Analytical Methods for High Dimensional Physiological Sensors

by

Gustavo Adolfo Lujan Moreno

A Dissertation Presented in Partial Fulfillment
of the Requirements for the Degree
Doctor of Philosophy

Approved April 2017 by the
Graduate Supervisory Committee:

George C. Runger, Co-Chair
Robert K. Atkinson, Co-Chair
Douglas Montgomery
Rene Villalobos

ARIZONA STATE UNIVERSITY

May 2017

ABSTRACT

This dissertation proposes a new set of analytical methods for high dimensional physiological sensors. The methodologies developed in this work were motivated by problems in learning science, but also apply to numerous disciplines where high dimensional signals are present. In the education field, more data is now available from traditional sources and there is an important need for analytical methods to translate this data into improved learning. *Affecting Computing* which is the study of new techniques that develop systems to recognize and model human emotions is integrating different physiological signals such as electroencephalogram (EEG) and electromyogram (EMG) to detect and model emotions which later can be used to improve these learning systems.

The first contribution proposes an event-crossover (ECO) methodology to analyze performance in learning environments. The methodology is relevant to studies where it is desired to evaluate the relationships between sentinel events in a learning environment and a physiological measurement which is provided in real time.

The second contribution introduces analytical methods to study relationships between multi-dimensional physiological signals and sentinel events in a learning environment. The methodology proposed learns physiological patterns in the form of node activations near time of events using different statistical techniques.

The third contribution addresses the challenge of performance prediction from physiological signals. Features from the sensors which could be computed early in the learning activity were developed for input to a machine learning model. The objective is to predict success or failure of the student in the learning environment early in the activity. EEG was used as the physiological signal to train a pattern recognition algorithm in order to derive meta affective states.

The last contribution introduced a methodology to predict a learner's performance using Bayes Belief Networks (BBNs). Posterior probabilities of latent nodes were used as inputs to a predictive model in real-time as evidence was accumulated in the BBN. The methodology was applied to data streams from a video game and from a Damage Control Simulator which were used to predict and quantify performance. The proposed methods provide cognitive scientists with new tools to analyze subjects in learning environments.

To my lovely wife, Liz and our wonderful son, Gustavo

ACKNOWLEDGMENTS

I would like to thank my advisor Dr. George Runger for all his support and guidance throughout all these years at ASU. He taught me everything I know about data science and believed in me and in this project. I would also like to thank my Co-Chair Dr. Robert Atkinson for providing me with the opportunity of working in his lab and providing all the necessary resources to make this project happen. I would also like to thank Dr. Rene Villalobos for offering me the opportunity to work with him in my first year and for providing wisdom and support. I am also thankful to Dr. Douglas Montgomery for his guidance and feedback provided in this work.

Special thanks to all my friends I made and my fellow students for all those days, nights and weekends of study, hard work, coffee and amazing moments. I found a lot of support from you guys even though we come from different countries and have a different culture, nevertheless we share the same dreams. Very special thanks to my friends Ghazal and Sangdi for helping me with the school paper work while I was away.

Finally, I would like to thank my father Carlos who taught me the value of education and for his unconditional support, my mother Carmen for her sweet conversations during the weekends that made me feel close to home. I would like to thank my sister Karla and brother Allan for taking care of my parents while I was away, I'm deeply indebted to you. Most importantly, I would like to thank my wonderful and smart son for learning a new language and facing this new world with an innocent bravery and finally thank you my dear wife Liz for all your patience and for believing in this crazy and amazing adventure.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

Figure Page

Figure Page

Figure Page

Chapter 1

INTRODUCTION

1.1 Motivation

Since the first publication of Picard's seminal paper on ”affective computing” in 1995 there has been a bloom of research related to this area (See Fig. 1.1). For example, in the very highly cited paper [1] the authors present a multimodal dataset in order to analyze human affective states. In this study participants watched 40 one-minute long videos with the objective of eliciting different emotions which were classified in terms of levels of familiarity, dominance, like/dislike and arousal. In this experiment electroencephalogram (EEG) was used for the classification of low/high levels of emotions using a Naive Bayes classifier. In another prominent paper [2], the authors provide a taxonomy for procedural content generation (PCG) algorithms in order to personalize user experience by analyzing the cognitive and affective state of users. In order to show the effectiveness of their approach they employ games which are a good source to study emotions in human computer interfaces (HCI) because they can elicit complex patterns of affective states. Moreover, in [3] authors use machine learning techniques such as Support Vector Machines and Neural Networks to identify affective states during context-specific scenarios using different sources such as: facial expressions, audio cues and shoulder gesture as physiological inputs.

1.2 Physiological Signals in Affective Computing

According to [4] the most basic emotions are normally classified as: joy, sadness, fear and anger. However, there is no complete agreement between theorists about

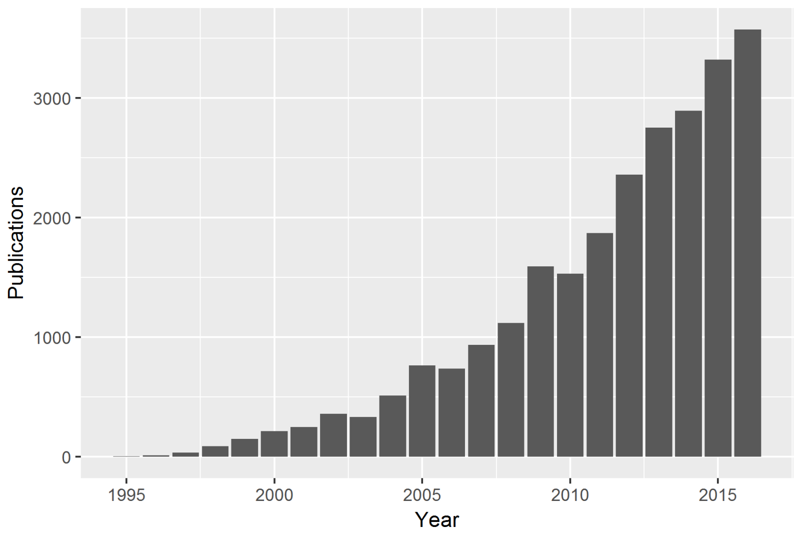


Figure 1.1: No. of Publications on Affective Computing According to Google Scholar.

what a basic emotions is, having researchers who list happiness and sadness as the only two basic emotions and other authors who list up to 20 different types. The two dimensional emotion model is another popular representation for affective states which is represented by a horizontal axis of negative/positive valence and a vertical axis of low/high arousal [5]. In this model valence ranges from pleasant (positive) to unpleasant (negative) emotions while arousal ranges from calm (low arousal) to excited (high arousal). Under this model, joy for example would be located in the upper right quadrant of positive valence and high arousal while sadness would be composed of low arousal and moderate negative valence.

Certain parts of the brain have been identified to play an important role in humans' affective states. For example, the left frontal region of the brain has been observed to be more involved in positive emotions such as happiness and joy while the right frontal area shows more activation for negative experiences such as sadness and fear [6]. Another study [7] observed that occipital high theta and low alpha asymmetry over the central nodes activity increased while participants played violent video games and confirmed that EEG was a reliable method when compared to approaches traditionally used in the field. Certain parts of the brain have also been associated with cognitive activity for example, elevated levels of activity in the pre-frontal cortex is believed to be associated with short-term memory in humans.

Sometimes researchers are also interested in analyzing not only affective states but also mental states such as awake/sleep and alert/drowsy [8] and these states have been linked to increase/decrease activity in certain parts of the brain. For example, in [9] researchers developed an EEG-based system to detect cognitive impairments in truck drivers in order to detect early signals of fatigue and drowsiness caused by sleep deprivation because these mental states have been linked to a decrease in cognitive performance. Working memory and mental workload are other constructs that are not emotions per se but are strongly related to cognitive performance in humans [10]. Working memory is responsible for the processing, manipulation and retrieval of information and it is critical for reasoning and learning while mental workload is the level of cognitive processes occurring in the brain and it establishes the relationship between cognitive tasks demands and the capacity of an individual's working memory [11].

The critical role of emotions and their influence on cognitive performance and learning have been studied recently and the body of literature has been steadily growing [12]. Moreover, it has been shown that people's emotions strongly affect productivity and the learning process and therefore it is important to be able to recognize and interpret the learner's different affective states in order to ensure an affective learning [13]. One area of application of Affective Computing is in the design of Intelligent Tutoring Systems (ITS) which are computer-based systems that try to adapt to humans in order to enhance learning. The human tutor has been seen as the gold standard in personalized learning and efforts are made to develop ITS which can provide the same level of instructional advantage. Humans are good at recognizing emotions and human tutors use this skill to engage students in order to generate good quality learning by encouraging creativity and facilitating a flexible environment for problem solving. The analysis and understanding of emotions and how they affect learning performance therefore becomes critical. As a consequence, a pre-requisite for a good ITS is to be able to monitor learner's affective states and take actions according to the information gathered from the subject regarding his individual and unique experience and also his emotions [12]. Hence, the first challenge of an ITS is how do we measure information about an individual's affective states and how do we build good predicting models to be included as part of a tutoring system. The next section talks about how physiological signals can help us model affective states.

1.3 Physiological Signals in Affective Computing

Different methods have been proposed to study and recognize emotions which include: neurophysiologic response, self-report, behavioral response and autonomic measurement among others [14]. Researchers have noticed that emotional changes in humans produce changes in different physiological response such as: blood pressure, heart rate, skin conductance and temperature [6]. These physiological signals are a good way of measuring users' affective states because they are non-intrusive and cause less distraction to participants unlike self-reported methods where participants need to interrupt the cognitive task in order to provide feedback about his current emotional state. Ideally we would like to use non-intrusive methods to measure affective states because subjects are not required to perform additional tasks that are not related to the primary goal. Non-intrusive measurements allow to quickly track affective state changes and don't rely on self-report methods which are disrupting and highly unreliable even within the same participant. Furthermore, physiological data is provided in real-time on a continuous stream while other type of overt behavior normally is incomplete and it is provided intermittent on discrete intervals [15]. Facial expressions, voice or speech are another type sources used to model affective states as well as body language and posture which have been used to study different emotions. However, most reliable and efficient physiological measurements come from four sources: brain, heart, skin and muscles. For brain we have electroencephalogram (EEG), functional magnetic resonance imaging (fMRI), electrocorticography (ECoG), functional near-infrared spectroscopy (fNIRS), magneto-encephalography (MEG), intracortical electrodes (ICE) and positron emission tomography (PET) [16]. For heart we have electrocardiograms (ECG o EKG), for muscles we have electromyogram (EMG) and in order to measure skin conductivity there are sensors which measure the electrodermal activity (EDA) [17]. All of these physiological signals have been used to model different affective states. For example, EMG has been shown to be a good predictor of motor preparation before a body movement [18] and there is a large body of literature involving the use of EEG to explore the relationship between affective states and cognitive performances [10], [19], [20], [21] and [22]. In the next section we will see how the EEG is a reliable and low cost physiological device.

1.4 EEG as a Reliable Physiological Input

The methods and techniques outlined in this work can be applied to different and even combined physiological signals which are not restricted to the ones we just mentioned. However, the current proposal is mainly focused on the use of EEG as the primary physiological signal. The first attempts to model cognitive activity can be traced back to Berger and his discovery of EEG and the alpha waves in 1929 and even though there are several devices used to study brain activity in terms of portability, reliability and costs EEG turns out to be a very practical tool to use in a research setting. EEG has been used to study different cognitive functions including but not limited to: emotions, language, memory, perception and social cognition. The signals captured by EEG can be associated with different brain processes and they are detected by the synchronization and desynchronization of neurons in specific parts of the brain [20]. However, the main goal is not to understand the physical properties of the brain but to be able to understand cognitive processes and behavior. Another reason why EEG is a good tool to analyze cognitive processes is that is able to directly analyze neural activity which is the voltage oscillations that can be measured on the scalp and those fluctuations are a direct result of cortex activity [23]. Perhaps the main reason why EEG is an excellent tool to analyze the brain is that it also provides a multidimensional signal which contains topographical information about neural processes. This is possible because each of the electrodes is located on a specific area of the skull and they capture voltage oscillations which contains both temporal and spatial information. As a matter of fact, EEG provides information in four domains: time, space, frequency and phase which enable researchers with different options to explore a wide array of psychological and physiological experiments.

In order to analyze brain waves different techniques have been proposed which include spectral analysis, synchronization and time-frequency transformations. EEG is widely used because it is able to capture brain dynamics at the time the cognition happens. The motor, emotional, linguistic, perceptual processes occur very quickly in matters of seconds and milliseconds and very few devices are capable of recording this information in real time in a reliable way at a low cost [23]. The activity produced by neurons is captured by the EEG and contains both spatial, temporal and spectral information. Spatial because the sensors located in different parts of the skull provide topographical information about the source and origin of the neural activity. Temporal because an increase/decrease of neural activity before, during and after and stimuli can provide some insights about the cognitive process and spectral because the signal can be decomposed into different frequencies which are known to be related to certain brain activity. The θ (theta) wave band is in the low side of the spectrum around 4-7 Hz and it is implicated in different cognitive functions such as cognitive control and memory. The δ (delta) is even slower at 1-3 Hz while the β (beta) bandwidth is in the 13-30 Hz frequency. Analyzing these bandwidths and some features generated from ratios researchers have found significant results for identifying different motivational traits, cognitive performance, emotions and even psychopathology [22].

The most popular feature extraction method for EEG signals is perhaps spectral analysis by using the Fourier transform. However, one of the shortcomings of this approach is that once we go from time to frequency domain all the temporal information is lost. As a consequence, other time-frequency approaches such as the discrete wavelet transform (DWT) and short time Fourier transform (STFT) have be proposed which consider both the temporal as well as the spectral information provided by the EEG. Hence, human cognition can be better understood by electrical changes in the participants' scalp which are associated with mental activity in a non-intrusive way.

In order to model the different cognitive and affective states different machine learning approaches have been proposed. In ideal situations the goal of any data mining or machine learning technique applied to the problem of modeling affective states should produce high accuracy while maintaining a low computational complexity [24]. For example in [10] the researchers fitted a discriminative model to classify engagement at different levels of intensity using a headset containing only 10 sensors. In another study [25] the authors used fuzzy k-means and fuzzy c-means as the clustering methods to classify emotions initially using 64 sensors and decreasing the number of electrodes down to 24 without sacrificing accuracy. The process followed by researchers to model different emotional states using EEG goes as follows: first the EEG raw signal is collected and a pre-processing phase is executed in order to reduce the noise of the signal as well as applying spatial and temporal filters to the EEG. After the signal has been pre-processed features are extracted applying different signal processing techniques such as even related potentials (ERPs), spectral power decomposition or phase synchronization. The last step is to apply either continuous (regression) or discrete (classification) machine learning models in order to estimate the various emotional states [14].

Another interesting construct that researcher have tried to model that is not directly related to affective states is cognitive workload which is defined as the total effort in the working memory. For example, in [26] the authors found that frontal theta activity in humans increases with the number of items retained in working memory while in [21] it was found for example that EGG alpha and beta bandwidths reflect cognitive and memory performance. For workload, features can be extracted using different techniques including: phase-based, spectral-based and time-based in order to characterize this construct like in [27] where researchers used a multi-class support vector machine (SVM) using the aforementioned features using EEG sensors as the raw signal in order to build a profile for working memory.

EEG headset providers rely on the techniques explained above in order to come up with robust models that perform well not only on a specific individual but that are able to generalize to larger populations. In order to achieve this objective, researchers extract feature that are good predictors across a wide range of different populations and perform baseline tests before the main experiment in order to accommodate individual differences and increase predictions accuracy. Topographical information about sensors is also considered during the modeling phase. For example, in [8] the experimenters found that topographical information derived from EEG provided useful information regarding the type of cognitive activity participants were performing on 14 different tasks. In the next section we talk about the contribution and organization of this dissertation considering the use of physiological signals to understand different affective states near the time of events.

1.5 Organization and Contributions of This Dissertation

The objective of this proposed research is to develop analytical methods for high dimensional physiological sensors. The methodologies developed in this work are applicable to numerous problems in learning science and also in industrial settings where high dimensional signals are present. In order to show the robustness of the methodologies proposed in this work we present different experiments in which we had a group of subjects participating on different learning environment systems. The characteristics of these dynamic learning environments is that they are complex and have multiple events embedded in time. In all the experiments a vector of physiological measurements y_t is continuously recorded where t represents time. Lessons learned in this work can be potentially used to analyze the relationship between events and physiological signals that later can be used to design robust ITS.

Chapter 2 proposes a robust methodology for performance assessment which can be applied to different event-driven learning environments using any type of physiological signal to monitor users' affective states. The methodology which we have called the event-crossover (ECO) has some analogies with the case-crossover which has been used to evaluate a measurement obtained continuously in real time near acute (abrupt) events. Unlike the case-crossover which usually considers longer time intervals such as days, the ECO works on time intervals of seconds and even milliseconds which adapts pretty well to the high frequency sampling rate of most physiological devices. The methodology is relevant to studies with the following characteristics: the experiment involves a learning environment with different types of events occurring a different points in time as a consequence of the user's decision making process and it is desired to evaluate the relation between these events which exhibit a random pattern and occur abruptly and the physiological measurement which is provided continuously in real time. The main advantage of the ECO is that each subject acts as its own control which avoids the traditional necessity of controlling for other confounded effects such as age, gender, health, skill level, gaming experience, etc. In this contribution we also introduce the concept of a moving hazard window in order to analyze the physiological signal before, during and after specific events. For illustration purposes the methodology is applied to analyze error rates in a popular video game using affective constructs provided by an EEG headset. Significant effects are detected following this methodology.

Chapter 3 introduces three approaches to analyze events that occur randomly during a learning environment and are embedded in time and where a multi-sensor physiological signal is provided in real-time. Instead of analyzing each of the signal components separately as in the previous contribution we propose a multivariate approach to simultaneously examine different patterns around events related to the participant's performance. The high dimensionality on this type of study is caused by the availability of different devices simultaneously recording physiological signals such as: electromyograms (EMG) to detect muscle activity, electrodermal activity (EDA) for skin conductivity, electrocardiogram (EKG, EOG) for heart rate, electrooculogram (EOG) for eyes movement and EEG to record brain activity. In the first approach we propose a multivariate version of the event-crossover where instead of analyzing the different physiological signals independently we use all the information of the input vector distribution near time of events using multivariate methods to draw conclusions. In the second approach we tried to address the question: Given a multidimensional physiological signal from a subject, what is the affective state of this subject at a given time or event and how can this state be represented? For this purpose we represent different physiological patterns in the form of weight combinations using self-organizing maps (SOM) which are a type of artificial neural network which maps high-dimensional data into a lower dimension representation without the need of any labels. The characteristics of interpretability and the preservation of topographical properties make this approach suitable for the analysis of physiological signals. Once the SOM was trained various statistical techniques were applied to analyze correlated proportions of node activations near time of events. In the last methodology proposed we compare for differences in the physiological signals between two groups at the time of specific events using univariate and multivariate methods. In order to show the effectiveness of these new methodologies a case study is presented analyzing events and the decision making process using a damage control simulator as the learning environment. The physiological signal used in the experiment is electroencephalogram (EEG). Significant results for all approaches are found. The methodologies proposed in this contribution can help better understand the decision making of participants around events in a complex learning environment. Lessons learned can be used by researchers and educators to improve the design of intelligent tutoring systems where physiological input signals are intended to be used to adapt the system, enhance user experience and improve learning.

Chapter 4 addresses the challenge of performance prediction using a physiological signal. In this contribution the methodology *Bag of States* is proposed with the goal of extracting features that later could be used on a machine learning model to make a two class (pass/fail) classification. The intention of the methodology presented is to provide cognitive scientists with the statistical and machine learning tools to be able to design a feedback system which considers different user's profiles in order to increase engagement, provide enjoyment, stimulate attention while preventing failure. We have employed the affective constructs provided by a research-grade EEG device as our main input but the methodology can be used to fit a wide variety of physiological signals. The methodology makes use of self-organizing maps (SOMs) to define different affective states but unlike the previous approach this time the SOM provides the time spent on a given affective state and this information is later used as input on a machine learning model. Once the SOM is trained with the physiological signal the number of activations is counted for each of the output nodes and fed into a logistic regression model. We have named the methodology *Bag of Affective States (BAS)* because it resembles a bag-of-words model which is one of the most popular techniques for object categorization [28]. The methodology was applied to a damage control simulator where participants required to perform several complex tasks with the objective of putting out a fire on a submarine. A high cross-validated accuracy was achieved even after reducing the number of affective constructs used to train the SOM and also reducing the number of nodes selected for the machine learning model. Findings suggest that participant who succeeded the mission were more likely to spent time in an affective state formed by a combination of low levels of engagement, distraction and drowsiness as well as low to moderate levels of workload in contrast with participants who failed who showed lower levels of engagement and higher levels of workload. The utility of this methodology relies on the fact that features were extracted in the first seconds of the simulation opening the door for a close-loop system to be able to recognize the user emotional state early on and adapt accordingly.

Chapter 5 introduces a novel way for predicting learners performance using Bayes belief networks (BBNs) in order to provide the individual with the appropriate guidance to maximize learning. The current methodology presents a way of using a BBN and its latent variables temporal information as inputs for a logistic regression model in order to make predictions about performance. The utility of this approach is that we consider time into the analysis and we focus on how early we can predict the final outcome. Traditionally, BBNs need a lot of evidence in order to be reliable and they are based partly on expert knowledge which sometimes could be biased. Furthermore, the latent nodes of BNNs are normally associated with specific skills and psychometricians generally analyze them in a unidimensional manner [29] not considering that students skills are most of the time highly correlated. In this experiment a BBN generated from a Damage Control Simulator was used to predict performance of 69 subjects. The posterior probabilities were updated in real-time as new evidence was presented. *Personal safety* which was one of the latent nodes turned out to be the most important predictor. The logistic regression model with 10-fold cross-validation achieved high accuracy where we were able to predict as early as 38 seconds into the simulation the final outcome of the session. Furthermore, by using logistic regression and the latent nodes scores as inputs we open the door to explore potential interactions between different latent skills challenging the ”unidimensional” approach that traditionally pychometricians embrace. Ignoring the fact that the majority of skills for a given objective are highly correlated can lead to miss important interactions [29]. The results show that Bayes Belief Network which are normally included in learning environments could be used in conjunction with machine learning algorithms to predict performance early on. Furthermore, we show that it is possible to identify the latent variables which have more discriminative power by means of variable importance. The utility of this approach will help cognitive scientists to use current available information embedded in tutoring systems to adapt and respond to different users’ needs.

Chapter 2

AFFECTIVE STATE ASSESSMENT IN A LEARNING ENVIRONMENT WITH PHYSIOLOGICAL SIGNALS AND EVENT-CROSSOVER ANALYTICS

2.1 Introduction

Recently, EEG has been used to understand a subject's affective states and predict reactions to stimuli. For example, [30] used EEG and other physiological signals to assess flow in games with the goal of anticipating users' intentions. This is an interesting area of research since emotions influence our actions, and therefore, dramatically affect our daily activities [31]. Some of the affective states that past studies have used or tried to model include happiness, surprise, anger, fear, disgust, sadness, fatigue, stress, drowsiness/alertness, task engagement and mental workload [9, 10, 31]. For example, in [20] the authors classified emotional responses when listening to different types of music where both EEG and self-reported ratings were used. This is a major improvement considering that the majority of the traditional affective assessments were only self-reported using different theoretical models such as the Keller's ARCS [32].

Brain Computer Interfaces (BCI) have been designed using EEG allowing humans to interact with their surroundings by using brain waves instead of muscles. The neuroscience community believes that BCI have a great potential to improve the quality of life of patients with disabilities [19]. Furthermore, BCI have moved from being mainly used for medical purposes to other applications such as video games [16]. For example, in [33] the researchers successfully classified experts from novices while playing a video game using logistic regression with a ridge parameter. BCI technology has started to have a commercial application with games such as Mattel's Mindflex, which uses brain waves to control the height of a floating ball. Kinect motion tracking sensor (Microsoft) has also been used in game environments using EEG functional brain mapping to isolate body movements of participants while playing a virtual ball game [34].

Intelligent Tutoring Systems (ITS) which try to combine cognitive sciences with artificial intelligence and computer software have also looked to understand the student's emotional states to customize the tutoring experience and enhance learning [35]. Unfortunately, the majority of the traditional techniques are invasive and disrupt the learning process by asking participants about their feelings or workload in the middle of a task. ITS in the form of Human Computer Interfaces (HCI) are designed in different shapes, from formal software indented for educational purposes to industry training, military training and simulators. In the case of education there has been an increasing interest in exploring different types of data which come from learning environment using data mining techniques [36].

Sometimes there is blurry line between a simulator and a video game to enhance learning. In fact, video games are an excellent tool to keep the subject engaged and motivated to perform specific tasks and they have been widely used to study the brain in human subjects [34]. For example, in [37] the authors used physiological information to study the social relationship between players on a first-person shooter video game. They focused on the participants' response to victory and defeat when playing against a friend or a stranger.

Performance assessment in many learning environments such as video games is more appropriately based on events that occur during the time interval of the game than an overall or average affective measure. Consequently, affective measures of players are similarly expected to be most sensitive to events. Furthermore, the measures near the times of events might be expected to be substantially different from measures during more routine periods of play without events. Thus, affective measures near events can provide more important summaries of a player's affective status than an overall summary. An important element of performance assessment is a statistically valid methodology to evaluate affective measures near events. The methodology should be robust to differences in subjects and environments, yet simple to evaluate and interpret. Moreover, the use of a control is a valuable addition to a methodology that facilitates a simple and interpretable approach.

Here we present a robust methodology for performance assessment that can be applied to a large number of event-driven EEG experiments such as the video game experiments that motivated this work. It has been found that in order to increase participant performance it is imperative to understand the learner in-game behavior [38]. A focus to measures near events has been considered in other domains, outside of affective monitoring, and approaches can be modified for the EEG domain. To this end, case-crossover studies have been used to evaluate a measurement (which is obtained continuously in real time) near acute (abrupt) events. For example, the approach has been successfully used in applications such as epidemiological studies to investigate the effect of transient effects on the risk of acute events [39]. Although there can be differences in the objectives of case-crossover studies from our objectives, a similar analytical framework can used. Also, case-crossovers usually consider longer time intervals (such as days) but we show that the methodology applies equally well to the higher-frequency measurements obtained in EEG. We provide a simple, robust approach that can be useful in a broad range of studies that involve events and continuous measurements recorded over time.

Although in our experiment an EEG headset is used as a physiological input to evaluate errors, other type of sensors such as eye tracking, facial expressions, electrocardiogram (ECG), electromyography (EMG), and galvanic skin response (GSR) can be used. Furthermore, the events don't need to be errors but rather any type of event defined by the experimenter according to the video game, tutoring system or learning environment. The layout of this chapter is as follows. The case-crossover methodology is fully explained in section 2.2. In section 2.3 the experimental protocol is explained in detail. Section 2.4 provides with the results and discussion. Finally, the conclusions are presented in section 2.5.

2.2 Event-Crossover Analysis

In this chapter we focus on events and the physiological measurements near the time of events. This methodology has analogies to studies used to assess the effect of exposures to outcomes in health care. A case-crossover study is used in health care to study the effect of transient exposures on rare, acute events [39]. In a case-crossover study, the case refers to the subject at the time of an event (hazard period), while the control (known as a referent) is the same person at another time (control period). The case-control pairs are used to study the effect of the exposure. It is similar to a matched, case-control study in that the analysis compares the distribution of exposures, not the proportion of events. The key benefit is that each case is matched with a corresponding control from the same subject to compensate for potential confounding from fixed subject attributes.

The analysis of a case-crossover follows an intuitive approach similar to a matched case control study. One compares the level of exposure in the hazard period to the level(s) of exposure in one or more control periods. A statistically significant difference (paired t-test) identifies the exposure as linked to the event. Control periods are selected to attenuate potential confounding effects. Usually some randomization is employed in the selection. Although frequently one control period is used for each event, more than one can be handled in a manner similar to that used in a matched case control study with several controls. The original objective of a case-crossover was to find the significant determinants of the event. We will use the case-crossover study as a framework, but we will modify the approach to more generally consider relationships between the events and physiological measurements.

Several methodology steps are used in our analysis approach. We share similarities with the case-crossover methodology but we find it useful to emphasize a more general applicability. Consequently, we refer to our approach for embedded event analytics as an event-crossover (ECO) study. The first step is to define an event case. It should be very clear and both easy to track and record. In a video game environment an event could be every error, or it could be every time a player shoots in a first person shooter game. It also could be when the player dies, makes a decision, goes into other level, clicks certain button, crashes, jumps, etc. Furthermore, analyses can focus on different event types. For example, one analysis can be applied to error event, while the same analysis can be applied only for decision events. Moreover, rather than a particular event instance, we find it useful to consider an interval of time that meets certain criteria as an event.

Rather than an exposure as a hazard as in a traditional case-crossover study, the exposure in our models refer to different physiological inputs for stimuli presentations, such as EEG, eye tracking, facial expressions, ECG, EMG and GSR among others. This is not restricted to only one input since the methodology can be applied to several inputs. It is also important to consider the sampling rate of each of the inputs. For example, it is not uncommon to see an EEG sampling rate of 256 Hz, which corresponds to recording 256 sample points every second. Moreover, according to [41] modern eye-trackers, which is another type of physiological signals, have a sampling rate from 25-2000 Hz. In conclusion, since each of the physiological inputs may have a different sampling rate it is good to define a unit time for the analysis where all the devices will be able to provide good resolution. A key difference from traditional case-crossovers is that we do not require the hazard period to occur before the event. We are also interested in the physiological measurements after an event occurs and we adapt the methodology to include these analyses.

It is convenient to partition the total time length of the game into equal length intervals of h seconds (with possible truncation) and let i denote the index of the associated time window, for $i = 1, 2, \ldots, n$. For a sufficiently small value of h, we can associate an event with a time window. But there is flexibility in how events are defined. In some of our analyses, we only consider an event to occur if there are more than 5 errors in a time window h. We define y_i as a particular physiological measurement in time interval i. For high-frequency measurements, y_i might be an average or median of measurements in interval i. In other cases, we might compare slopes of EEG measurements over intervals. The basic analysis will compare y_{ei} in the hazard period (where subscript e denotes that the interval i contains an event) to y_{ci} in a control period (where subscript c denotes an interval i without an event). The physiological measurement y_{ci} in our studies is selected randomly from the intervals without events for the same subject in the same game or session. Further restrictions can be placed on y_{ci} to remove additional confoundings. For example, in air quality studies a control period is selected on the same day of the week as the hazard period in order to control for difference in the intensity of traffic. In addition, if the playing environment changes dramatically, so, for example, the player dies and is reborn, one might restrict the control period to be selected within the same player lifetime as the hazard period. Environmental studies might select the control and hazard periods within the same season. Control periods might also be restricted to be selected either before or after hazard periods. Our environment was sufficient short in duration so that restrictions were not used.

An advantage of the ECO approach is that a simple, interpretable paired t-test is applied to the physiological measurements. Let n_e denote the number of intervals with events in the environment. A paired t-test is applied to the measurements from the hazard periods y_{ei}, $i = 1, 2, \ldots, n_e$ and the corresponding measurements from the control periods y_{ci}, $i = 1, 2, \ldots, n_e$, where one-to-one matching is assumed. More than one control period can be selected per hazard period and the analysis method can be modified in a manner similar to case-control studies with more than one control per case. See [40].

In addition, it is important to explore the relationship of the physiological input before the event happens or even after the event happens. This can be incorporated easily into the approach. We can define a hazard period as w intervals before or after the event interval. If y_{ei} is the measurement in event interval i, we can also consider $y_{e,i-w}$ and $y_{e,i+w}$ that denote the measurements w intervals before and after the event, respectively. Here intervals $i + w$ and $i - w$ define new hazard periods. Controls can again be randomly selected and a paired t-test can be applied to compare the measurements between these hazard periods and control periods. Consequently, the method can simply be adapted to compare physiological measurements before or after events. In our studies, we investigate several values for w before and after the event. See Fig. **??**.

We also look to study the effect of the random selection of control intervals. Consequently, we conduct analyses with R replicates where each replicate uses a different random selection of control intervals to compare to the hazard intervals.

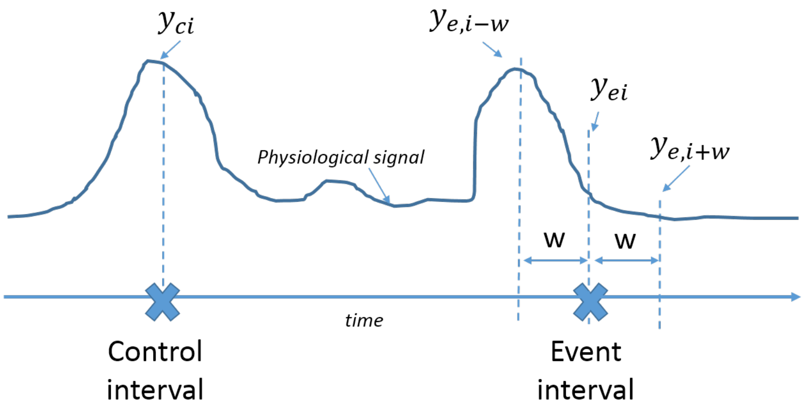


Figure 2.1: The Event-Crossover (ECO).

2.3 Experimental Protocol

2.3.1 Game Environment

As previously mentioned, the Guitar Hero video game was used as stimuli for the subjects. Guitar Hero is a game that involves holding a guitar interface while listening to music and watching a video screen. This type of video game is rich in graphics, multimedia, challenges and embodied activities triggers. This gives the best context in which to elicit changes in the affective states of the users. Guitar Hero provides a scenario where subjects are challenged in diverse ways that demand from them different skills related to a learning process such as concentration as well as visual, motor, and auditory skills. The user has five colored buttons to press on the guitar fingerboard. The objective is to use the left hand to press the correct button(s) while colored notes are streaming on the screen. The right hand is also used by depressing a switch that resembles a guitar strum or string picking.

2.3.2 Participants and Design

There were a total of 8 subjects recruited from Arizona State University of which 4 were men and 4 were women. Age ranged from 18 to 28 years. Participants were compensated and they had the option to leave the study at any time for any reason. Participants were asked to self-report their experience playing video games. Accordingly with the overall score and their self-report, four of them were classified as novices and four of them as experts. Data from two songs, an easy-mode and a hard-mode song, where collected. All selected subjects played the same songs in both modes. The easy song, "Story of my life", had length 5:40 (m:ss) and a total of 511 notes. The hard song, "One", had length 7:03 and 2189 total number of notes. Consequently, we had a total of 16 data sets, one for each player-difficulty possible combination.

2.3.3 EEG Recordings

Emotiv EEG is a high resolution, multi-channel, wireless portable EEG system. In [41] the findings suggest that Emotiv can provide a valid option to Laboratory EEG systems (Neuroscan) for recording auditory even related potentials. Another study assessed the quality of Emotiv to measure engagement, short and long term excitement and found consistency among these constructs, at least in learning environments [42]. However, authors in [43] do not recommend Emotiv for medical purposes such as rehabilitation or prosthesis control since the lack of reliability may cause negative consequences. Nevertheless, they agreed that Emotiv could be a good option for video game environments. Emotiv has 14 EEG channels with names based on the International 10-20 locations, these are: AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, AF4. The sensor data is referenced to left and right mastoid.

Emotiv internally samples at 2,048 Hz and applies band-pass filters in the range of 0.2-45 Hz. Finally, two digital notch filters at 50 and 60 Hz are applied. The output is downsampled to 128 Hz. Emotiv has a suite called Affective which uses the information derived from the channels to compute measures related to five affective states: short term excitement, long term excitement, engagement, meditation and frustration. The output that Emotiv generates of the raw signal from the 14 channels is provided at 128 samples per second (SPS) and the affective states at 2 SPS. The affective states defined by Emotiv Affective suite (2010) are the following: 1) short-term excitement, is a feeling of physiological arousal that is experienced with a positive value; 2) long-term excitement, it is similar to short-term excitement but it is measured over longer periods of time, typically minutes; 3) engagement, experienced as alertness and attention. The lack of engagement could be characterized as distraction; 4) meditation, experienced as relaxed and a clearness of the mind; 5) frustration, occurs when the difficulty level of a task is much higher than the skill level of the subject. This may cause a disconnection from expectation and reality.

2.3.4 Error Tracking

In addition to the EEG signal from the 14 channels and Emotiv affective states, the total number of errors for each of the seconds of the songs was also recorded. An error is made every time the player fails to hit the correct note. It is important to notice that for the hard song there is a guitar solo in the second half where there could be up to 15 notes per second. This means that the participant could make more than one error on a given second. Once the affective state features were generated, we plotted them along with errors against time to visualize the data. (See Fig. 2.2).

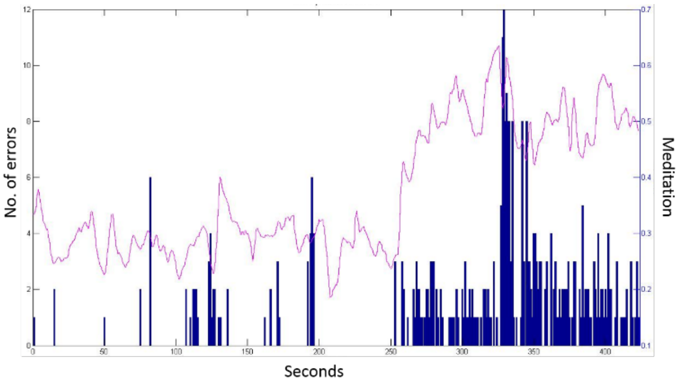


Figure 2.2: Number of Errors and the Affective State *Meditation*.

2.3.5 The Event-Crossover (ECO) Applied

In this study, the events are represented by the errors made by the participants when playing the video game. This is not a subjective measure because errors are reflected on the participant's score and there are also visual and audio cues that clarify when the errors happen. The reasons for errors include mistakes such as the right hand did not strum, the left hand pressed the incorrect button or buttons were not pressed at all, the note was played too soon or too late, the note was not held the required time, or any combination of these situations. The physiological input is defined by the measurements of the affective constructs provided by Emotiv EEG: short term excitement, long term excitement, frustration, meditation and engagement. We set $h = 1s$ as the interval length. Because Emotiv provides 2 samples per second of the affective states and our interval under study is 1 second length we averaged these two values to produce only one value per interval. From now on we will refer to this value as the affective state measurement instead of the affective state ”average”. Then we identified all of the intervals where at least one error happened among all participants, which are our events. Next, for each event we randomly selected a $h = 1$ second interval without error which defines the control. The controls were searched within the same participant playing the same song; as explained in the previous section. We performed this procedure using 4 different combinations for analysis: easy-novice, (novices playing the easy song), hard-novice, easy-expert and hard-expert. We applied a paired t-test to the results (a one t-test for the differences). To evaluate the effect of the random selection of control intervals, we replicated this process 15 times for each combination of participant skill and song difficulty.

In order to show that the methodology can be modified to serve different needs, we defined the events in three different ways: (1) a second in the song where at least one error occurred (> 0); (2) a second in the song where more than one error occurred (> 1); and (3) a second in the song where more than five errors occurred (> 5). This was done to study the relationship between the degree (or severity) of errors and the affective state. This also illustrates that the definition for the event can be modified for a different video game analysis. As mentioned, a player's affective state might not be related immediately to the players performance (the number of errors made). The affective state might predict errors, or result from errors. Consequently, we considered additional analyses. We also compared the participant's affective state value in intervals $w = 3$ seconds before and after, $w = 1$ second before and after. Including the interval that contains error(s), this provides 5 event intervals to be compared to control intervals. The intervals can be defined by the experimenter according to scientific or even empirical experience.

2.4 Main Results

For this analysis R statistical software version 3.0.3 was used. Table 2.1 shows the results for the analysis where the transient effect is immediate. This table shows the average of the p values and the average of the mean differences for the 15 replicates and the total number of events that satisfied each of the error conditions (> 0, > 1

Table 2.1: Results for the 4 Combinations in the Current Time (y_{ei}).

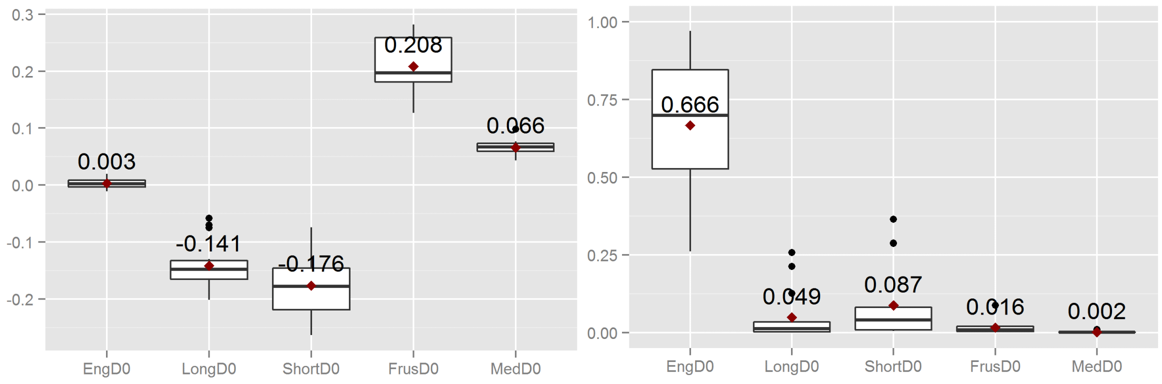
Combination →		Easy-Novice			Hard-Novice			Easy-Expert			Hard-Expert		
No. of errors →		> 0	> 1	> 5	> 0	> 1	> 5	> 0	> 1	> 5	> 0	> 1	> 5
	n →	153	46	NA	1513	1316	214	23	7	NA	658	410	75
Engagement	*p* value	0.512	0.558	NA	< 0.01	< 0.01	< 0.01	0.666	0.518	NA	< 0.01	0.003	0.396
	mean	-0.006	0.007	NA	0.027	0.029	0.045	0.003	-0.001	NA	-0.034	-0.028	-0.015
Long Term Excitement	*p* value	0.593	0.541	NA	< 0.01	< 0.01	0.585	0.049	0.146	NA	< 0.01	< 0.01	< 0.01
	mean	0.008	0.002	NA	-0.023	-0.025	0.004	-0.141	-0.218	NA	-0.137	-0.156	-0.215
Short Term Excitement	*p* value	0.52	0.468	NA	0.489	0.618	0.238	0.087	0.098	NA	< 0.01	< 0.01	< 0.01
	mean	0.003	-0.026	NA	0.002	-0.002	0.037	-0.176	-0.377	NA	-0.187	-0.21	-0.3
Frustration	*p* value	0.198	0.333	NA	< 0.01	< 0.01	0.133	0.016	0.056	NA	< 0.01	< 0.01	0.003
	mean	-0.025	0.023	NA	-0.036	-0.038	-0.033	0.208	0.248	NA	-0.164	-0.16	-0.172
Meditation	*p* value	0.611	0.53	NA	< 0.01	< 0.01	0.243	0.002	0.258	NA	< 0.01	0.001	0.196
	mean	-0.001	0.003	NA	-0.017	-0.016	0.009	0.066	0.053	NA	-0.041	-0.032	-0.026

and > 5). Because the range provided by Emotiv for each affective state is from 0 to 1 we can think of the mean difference as the percentage of variation between the affective state value during the hazard period and control period.

For the first column where an event was defined to be a second with at least one error (> 0) there are no significant results for the easy-novice combination. A couple of significant results for the hard-novice dataset were produced but we do not consider them to be relevant because the average of the mean difference is small and the large sample size yields the significance of the results ($n = 1513$). In this case, and for the rest of the discussion, we defined a mean difference to be large if the absolute value of the mean differences is greater than 0.1. In the easy-expert combination, long term excitement ($p = 0.049, \mu = -0.141, n = 23$) and frustration ($p = 0.016, \mu = 0.208, n = 23$) are significant. Meditation has a significant result, but according to our previous definition the mean difference is not large enough. The rest of the affective states for easy-expert do not show significant differences. Interestingly, for the hard-expert combination once again long term excitement ($p <$

$0.001, \mu = -0.137, n = 658$) and frustration ($p < 0.001, \mu = -0.164, n = 658$) had significant results and a large mean difference. This is intriguing because we see the same effect even though the games have different difficulty level and were played with 10 minute breaks between the first and the second song. This time also short term excitement ($p < 0.001, \mu = -0.187, n = 658$) is significant. Engagement and meditation were also significant but the mean difference was small.

For the > 1 case once again there were no significant results for the easy-novice combination. There are 4 out of 5 significant results for the hard-novice combination but the average of the mean differences is not large and the p values reflect the sample size ($n = 1316$). The easy-expert has no significant results ($n = 7$) and the hard-expert has the same pattern as in the > 0 case with the difference that this time $n = 410$. For the > 5 case we can see that there is a column with NA for the easy-novice because there are no more than 5 notes in any given second for the easy song. A similar result is shown for the easy-expert combination. The hard-novice combination has engagement as a significant result ($p < 0.001, \mu = 0.045, n = 214$), but we can see that the average mean difference is not very large. The rest of the results are not significant for that combination. Once again the hard-expert combination has long term excitement ($p < 0.001, \mu = -0.215, n = 75$), short term excitement ($p < 0.001, \mu = -0.3, n = 75$) and frustration ($p = 0.003, \mu = -0.172, n = 75$) with significant results. We can see that even with the sample size $n = 75$ we are still able to capture the effect as significant with the average of the mean differences even larger than the > 0 case ($n = 658$) and the > 1 case ($n = 410$). As an exploratory tool we can make use of the boxplots of the results over the 15 replicates as they allows us to not only see the central tendency but also the dispersion. Figure 2.3a shows boxplots of the mean differences for the case > 0 which refers to the interval at the time of the event (y_{ei}) and the easy-expert combination ($n = 23$). We observe that



(a) (b)

Figure 2.3: Box Plots for Mean Differences (a) and p Values (b), Easy-Expert.

the distribution of the difference in means has low variation across the replicates. In Fig. 2.3b we can see an example of the boxplot that shows the dispersion of the p values for the 15 replicates for the easy-expert combination. We observe, for example, that short term excitement has a larger mean difference (in absolute value) than long term excitement.

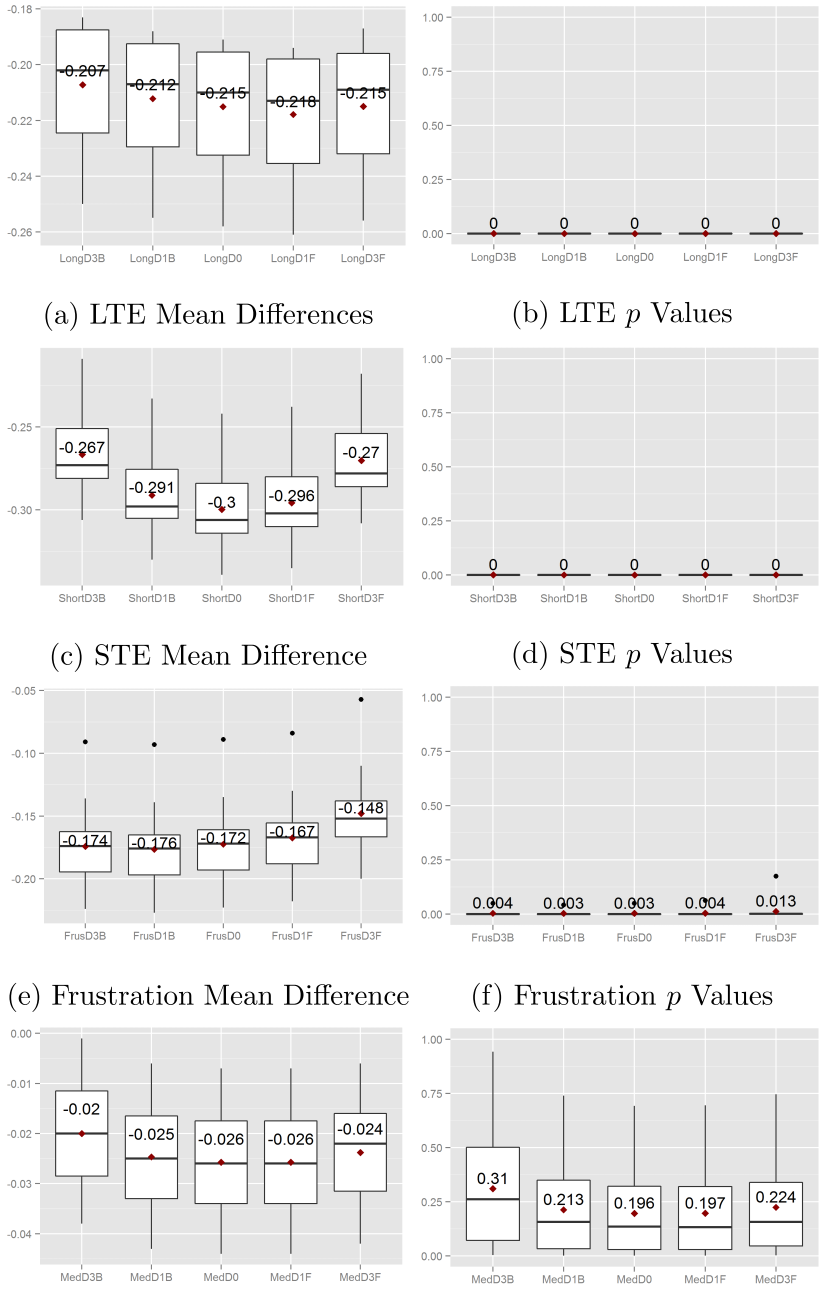
The 15 replicates procedure is recommended because we do not rely solely on one random selection that depends on the seed that was used to select the random control intervals. We also present the results to explore the relation of the affective construct values in time $i \pm w$ for $w = 3$. Figure 2.4 shows the results in the form of boxplots for the hard-expert combination and the five affective constructs. The average is marked as a dark red line in the box. In these plots we can clearly see a couple of outliers. For example the boxplot for the means in engagement (top left plot) we can see an outlier above the mean. This outlier represents one of the mean differences in one of the replicates. Also, from the p values boxplot for the same affective construct, we observe a couple of outliers for engagement 1 second before (EngD1B), 1 second after (EngD1F) and 3 seconds after (EngD3F). If once again we take the boxplot of means

Table 2.2: Results for the 4 Combinations Before the Event ($y_{e,i-3}$).

Combination →		Easy-Novice			Hard-Novice			Easy-Expert			Hard-Expert		
No. of errors →		> 0	> 1	> 5	> 0	> 1	> 5	> 0	> 1	> 5	> 0	> 1	> 5
	n	153	46	NA	1513	1316	214	23	7	NA	658	410	75
Engagement	p-value	0.546	0.475	NA	< 0.01	< 0.01	< 0.01	0.671	0.446	NA	< 0.01	0.014	0.696
	mean	-0.003	0.014	NA	0.027	0.029	0.044	-0.003	-0.02	NA	-0.032	-0.024	0.003
Long Term Excitement	p-value	0.565	0.552	NA	< 0.01	< 0.01	0.59	0.057	0.157	NA	< 0.01	< 0.01	< 0.01
	mean	0.009	0.007	NA	-0.023	-0.025	0.002	-0.136	-0.209	NA	-0.132	-0.15	-0.207
Short Term Excitement	p-value	0.346	0.557	NA	0.485	0.56	0.422	0.082	0.171	NA	< 0.01	< 0.01	< 0.01
	mean	0.03	-0.019	NA	0.004	0.002	0.023	-0.174	-0.308	NA	-0.177	-0.2	-0.267
Frustration	p-value	0.175	0.595	NA	< 0.01	< 0.01	0.047	0.015	0.018	NA	< 0.01	< 0.01	0.004
	mean	-0.026	0.001	NA	-0.037	-0.038	-0.043	0.212	0.297	NA	-0.16	-0.159	-0.174
Meditation	p-value	0.557	0.495	NA	< 0.01	< 0.01	0.223	0.001	0.059	NA	< 0.01	0.002	0.31
	mean	-0.005	0.011	NA	-0.017	-0.015	0.009	0.07	0.083	NA	-0.04	-0.03	-0.02

for engagement we will see that there is a steady increase (in absolute value) in the mean differences between the physiological measurement and the control point for different hazard periods w, starting from 3 seconds before (far left) to 3 seconds after (far right). There is also an interesting pattern in the short term excitement boxplots for the mean (first column). We observe a U shape from $w = -3$ to $w = 3$ seconds. This suggests that the mean difference in absolute value is greater for the interval at time of the event (ShortD0). In the case of long term excitement and frustration no pattern is obvious. For meditation, the mean differences are close to zero and there is no pattern. Additionally, the p values are far from significant and no pattern is obvious.

Table 2.2 presents the results for 3 seconds before the event ($i - 3$) and Table 2.3 contains the results for 3 seconds after the event ($i + 3$). We also did the same analysis for $w = \pm 1$ but the results are not shown here. The conclusions are very similar to those presented previously where we analyze events at the time they occur (y_{ei}).



(a) LTE Mean Differences

(b) LTE p Values

(c) STE Mean Difference

(d) STE p Values

(e) Frustration Mean Difference

(f) Frustration p Values

(g) Meditation Mean Difference

(h) Meditation p Values

Figure 2.4: Box Plots for Mean Differences (a) and p Values (b), Hard-Expert.

Table 2.3: Results for the 4 Combinations After the Event ($y_{e,i+3}$).

Combination →		Easy-Novice			Hard-Novice			Easy-Expert			Hard-Expert		
No. of errors →		> 0	> 1	> 5	> 0	> 1	> 5	> 0	> 1	> 5	> 0	> 1	> 5
	n	153	46	NA	1513	1316	214	23	7	NA	658	410	75
Engagement	p-value	0.38	0.579	NA	< 0.01	< 0.01	< 0.01	0.69	0.511	NA	< 0.01	< 0.01	0.084
	mean	-0.012	0.001	NA	0.027	0.029	0.048	0.005	0.012	NA	-0.039	-0.035	-0.041
Long Term Excitement	p-value	0.61	0.535	NA	< 0.01	< 0.01	0.622	0.049	0.143	NA	< 0.01	< 0.01	< 0.01
	mean	0.006	0	NA	-0.024	-0.026	0.002	-0.142	-0.223	NA	-0.14	-0.159	-0.215
Short Term Excitement	p-value	0.552	0.59	NA	0.524	0.427	0.279	0.162	0.127	NA	< 0.01	< 0.01	< 0.01
	mean	-0.005	-0.006	NA	-0.005	-0.01	0.033	-0.147	-0.338	NA	-0.188	-0.211	-0.27
Frustration	p-value	0.439	0.447	NA	< 0.01	< 0.01	0.186	0.02	0.051	NA	< 0.01	< 0.01	0.013
	mean	-0.015	0.016	NA	-0.036	-0.038	-0.03	0.201	0.251	NA	-0.161	-0.158	-0.148
Meditation	p-value	0.64	0.519	NA	< 0.01	< 0.01	0.554	0.002	0.277	NA	< 0.01	0.001	0.224
	mean	0.001	0.006	NA	-0.017	-0.016	0.004	0.063	0.047	NA	-0.041	-0.032	-0.024

2.5 Conclusion

The lessons learned in this study could be applied to develop more robust HCI or ITS. It is well known that students disengage overtime when using a HCI [44] and using noninvasive devices such as the EEG can help identify these signals to better design systems not only for entertainment but also for education. Furthermore, the ECO methodology can potentially be applied in real-time and provide feedback to the user. Making participants aware of their own affective states when making decisions can increase performance as shown by research focused on metacognitive processes [45]. The ECO methodology is most appropriate for studies where the main objective is to evaluate player performance where events are embedded over time with simultaneous sensors or physiological responses (skin conductivity, eye-tracking, EEG, heart rate monitors, etc.). The main advantage of our method is that each subject acts as its own control and no extra resources need to be spent to obtain controls. Furthermore, the methodology allows us to avoid the traditional necessity of controlling for other confounded effects such as age, gender, health, skill level, gaming experience, etc. The identification of the physiological changes around events can lead to a better design of HCI interfaces. Moreover, if these interfaces are optimized for both an effective and an efficient learning they can leverage the participant's skills and knowledge [35].

In the last 20 years there has been an explosion in generation of instructional material, especially online [46]. Thus, this methodology fits well for multimedia instructional material where it is desired to evaluate the relation between events which exhibit a random pattern and occur abruptly (acute) and a physiological measurement which is provided continuously in real time. In addition, the methodology considers replicating the results to find the true mean for the mean differences and the p values. The effectiveness of this methodology was illustrated using an example where participants played two Guitar Hero songs with different level of difficulty. The acute cases were defined in three different ways: intervals with at least one error (> 0), more than 1 error (> 1), and more than 5 errors (> 5). The physiological input was the values of 5 affective constructs provided by an EEG headset: engagement, frustration, meditation, long term excitement and short term excitement. The flexibility added to our methodology can easily be adapted to ITS in an inexpensive way both considering educational outcomes as well as machine learning issues. We are aware that this is only the first step and that further analyses need to be done which also include a cognitive task analysis and extensive evaluation. However, we believe that being able to understand the affective state of a subject at a given moment of time through the use of EEG is an important achievement. This methodology is even more relevant if we consider that the majority of the traditional ITS designs consider invasive ways of collecting information such as "think aloud" which can be very disruptive during a cognitive task [47]. Finally, the ECO approach identified the affective constructs of long term excitement, short term excitement and frustration as significant in the hard-expert combination for all type of cases. This implies that the affective state value of the participant for these emotional states is different when the player is making errors (events) than when the player is not making errors (control). A similar result was observed for the easy-expert combination in the > 0 case. Further analysis is needed to establish why this effect was observed in expert participants, but not in the novices. Nevertheless, once these differences are understood they can be used to perform individualized changes in HCI or ITS which respond to the user experience, something called the zone of proximal development, and have proven to be one of the most effective ways to enhance performance and learning [48].

Chapter 3

ANALYTICAL METHODS TO EVALUATE EVENTS AND PERFORMANCE IN AN EVENT-DRIVEN SIMULATOR

3.1 Introduction

Several studies have shown that emotions play an essential role in human cognition and perception [4] . Moreover, human performance depends not only on training and knowledge but also in the way a person is able to respond to different scenarios given his emotional baggage. In the field of affective computing the challenge for a human computer interface (HCI) is to recognize and respond to human emotions. Lately, several applications have been developed using the principles of affective computing in different areas such as: gaming, mental health and education [17]. In addition to the emotional component, cognitive load and memory are also related to performance and researchers have tried to study them through the use of physiological signals such as electroencephalogram (EEG). For example, in [21] evidence was found that the power spectral density (PSD) of EEG in the bandwidths alpha and theta were related to cognitive and memory performance. The relationship between emotions and workload are thus critical to better understand human performance and to design systems that adapt better to human emotional response.

The idealistic state would be to develop systems which employ real-time measurements of physiological inputs to predict the emotional state of a subject and adapt the system accordingly. Information provided by physiological measurements can help monitor and quantify the user's experience and proactively adapt in real-time [15]. The design of these dynamic difficulty adjustment (DDA) systems would not discard the information provided by traditional performance metrics, but it would complement it with affective feedback. This has already been done on simple games such as Anagram and Pong where a DDA mechanism was designed to adapt the level of the games by inferring probable anxiety levels in participants [49].

DDA systems are a growing area of research because the current state of intelligent tutoring systems (ITS) as well as video games for education or entertainment provide inaccurate challenge levels and they are normally based solely on performance but not in the affective states of user experience [50]. The aim of a DDA systems or an ITS should be to provide an experience tailored to users' specific characteristics. The most common method used to analyze affective states for a DDA consists of a real-time physiological signals that are constantly being monitored. For example, in [51] the authors developed a closed-loop real-time EEG-based drowsiness detection system. The system was designed to provide feedback to drivers just before sleep onset to try to prevent a car accident. In another study [30], a BCI was used to assess flow in games by adapting the system to different levels of difficulty and modeling the affective state of "flow" using machine learning methods such as support vector machines where the power spectral density characterization of the EEG was used as input. These studies while successful have been mainly focused on the continuous real-time physiological signal where a target is set and the DDA is expected to fluctuate around that target.

However, few studies have focused on more complex scenarios where different decisions are made and where events happen randomly sometimes as part of the system configuration and some other times as a direct result of the user's previous actions. One of these studies was conducted by [52] where instead of continuously analyzing an affective state throughout the session they focused around specific events for participants playing two scenarios: one in a virtual golf game and another one on a combat marksmanship simulator. Analyzing affective changes around events instead of continuously monitoring the physiological signal can provide more rich and complex information to try to understand the cognitive process and decision making of a participant under different scenarios.

It is in this line of event-based analysis that we propose this study. The current study describes three approaches to analyze events that occur randomly during a learning environment session where a multi-sensor physiological signal is provided in real-time. Instead of analyzing each of the signal components separately we propose a multivariate approach to simultaneously examine different patterns around events related to the participant's performance. Multivariate techniques can provide greater sensitivity to detect physiological responses and also help to dramatically decrease the number of statistical tests. On the other hand, univariate analysis normally ignores the relation or correlation between predictors while multivariate approaches take into account the joint distribution of the physiological signal.

The high dimensionality on this type of study is caused by the availability of different devices simultaneously recording physiological signals such as: electromyograms (EMG) to detect muscle activity, electrodermal activity (EDA) for skin conductivity, electrocardiogram (EKG, EOG) for heart rate, electrooculogram (EOG) for eyes movement and EEG to record brain activity. Furthermore, each of these devices could have multiple sensors to gather topographical information from the body and they could be combined at the same time. This multi-sensor combination makes the use of multivariate approaches more relevant when trying to analyze high dimensional input signals.

In our previous work to analyze affective effects [53], a video game with a simple graphical user interface was used and the task complexity was fairly simple. The techniques proposed in this study are applied to a much more complex learning environment where activities and decisions don't follow a predefined order and where the final outcome can greatly vary from participant to participant. It is under this scenario of high-dimensionality and user interface complexity that we present these approaches.

This chapter is organized as follows. Section 3.2 provides background. In section 3.3 different multivariate approaches to analyze events are presented. The experimental protocol explaining the learning environment, the physiological signal used and participants is presented in section 3.4. Results are shown in section 3.5. Finally, discussion and conclusions are developed in section 3.6.

3.2 Background

3.2.1 Event-Crossover Methodology (ECO)

An ECO methodology was proposed by [53] to explore the relation of physiological signals and events which alternate over time on a given learning environment which can be a simulator, a tutoring system, a video game or any type of HCI. The events are defined by the experimenter and it could be a person answering a question in a ITS, clicking some button, or in a video game environment a player turning right, making a mistake, jumping, dying during the simulation, etc. Events are related to decisions from the user and occur randomly while the physiological measurement is provided continuously in real time. The objective is to find if an event triggers a physiological response in the subject or, the other way around, if a sudden physiological change causes some type of event or decision. An advantage of this methodology is the detection of significant differences in the physiological response around events when compared to a control point. Moreover, control points are chosen within subjects so they act as their own control. Using controls from the same subject is a major advantage of the ECO methodology because it compensates for possible confounding effects such as age, gender, expertise, etc.

The ECO methodology focuses in the analysis of a physiological signal near the time an event happens [53]. A hazard period is the time at which an event occurs and, therefore, a change in the physiological response is expected. A control period is the time where an event does not occur. In this sense, the event-control pairs are used to study physiological changes within the same subject. A paired t-test is used to identify significant differences between event-control pairs. Another key attribute of the ECO methodology is that the hazard period could be set before, during or after an event occurs. This allows us to study the effect of a physiological signal on an event or the other way around, the effect of an event on a physiological response.

The first step in the univariate ECO [53] is to partition the time length into equal intervals of sized h (such as 1 second). For high-frequency physiological signals these intervals can contain several sample points. In this case, an average can be taken to associate a single measurement per interval. Denote n_e as the total number of events that we record in a session. A physiological measurement at event $i = 1, 2, \ldots, n_e$ is denoted as y_{ei}. For each of the events, a control interval without an event is randomly chosen within the same participant. The physiological measurement for this control interval is denoted as y_{ci}. The subscripts e and c are used to denote an event interval and a control interval, respectively. A paired t-test is applied to test for significant differences in the mean physiological values between the event and control intervals. A significant difference would imply that the mean affective state of the participant is lower/higher at the time of the event when compared to a control point. This information can help discover different behaviors and patterns which can be included in an ITS design.

In addition, to explore the relationship of the physiological signal before and after the event happens we can defined a hazard period w intervals before or after the event. If y_{ei} is the physiological measurement at event interval i then $y_{e,i-w}$ and $y_{e,i+w}$ represent the physiological measurement before and after an event, respectively. In this case intervals $i-w$ and $i+w$ represent new hazard periods. The control points are selected and denoted as before and a paired t-test is applied between the physiological measurements at these new hazard periods and the control intervals.

In order to analyze the stability of this procedure a series of R replicates is recommended. Depending on the number of events and the length of the session, the methodology can provide different results depending on the random seed that it is used to sample the control points. In each of the $r = 1, 2, \ldots, R$ replicates the mean difference $\sum_{i=1}^{n_e}(y_{ei} - y_{ci})/n_e$ of the physiological measurement between events and controls is recorded.

3.2.2 Self-Organizing Map (SOM)

Artificial neural networks (ANN) can be divided into two categories: supervised and unsupervised. In the supervised case there is a target which the ANN is trained to learn and guide the formation of the parameters. In the unsupervised case there is not a target and data is clustered using features inherent to the problem. SOMs belong to the unsupervised learning category and little knowledge is needed about the characteristics of the data [54]. The objective of the SOM is to map high-dimensional data into a lower dimension representation, usually a two-dimensional grid, therefore creating a discrete and spatially organized representations of input signals[55].

A SOM consists of an output layer normally arranged on a two-dimensional grid or lattice. The literature refers to the elements in the output layer in different ways: map units, cells, nodes, output nodes, neurons, etc. Each of these output nodes is fully connected to all of the input nodes (See Fig. 3.1) and each of these connections is associated with a weight w_{kq} from input node q for $q = 1, 2, \ldots, Q$ to output node

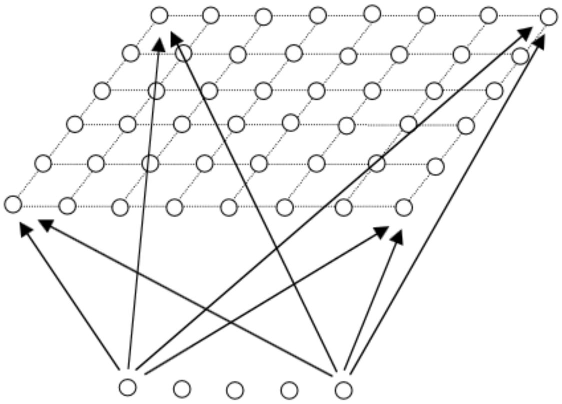


Figure 3.1: SOM Architecture.

k for $k = 1, 2, \ldots, K$. At the end of the training each output node has a weight vector $\mathbf{w_k}$ with Q elements and each weight vector represents a physiological pattern. It is critical for the formations of the SOM that weights are not updated independently, but in a manner that tries to preserve the topology [55]. A SOM provides a topology preserving map which means that if two instances are close to each other in terms of Euclidean distance in the original data space then they are expected to be mapped to nearby nodes in the output grid.

A SOM is trained as follows: first the topology for the output layer is defined in terms of the number of nodes and either a rectangular or hexagonal arrangement. In the first iteration weights are randomly initialized with small values. Next, an instance or feature vector y_i is selected and the Euclidean distance is computed between this instance and the weight vector $\mathbf{w_k}$ for each of the k output nodes. The output node k that provides the minimum distance is called the "best matching unit" (BMU) and it is denoted with the letter z. In order to preserve the topology a neighborhood of radius B_z around the BMU is defined to update the weights for the output nodes. Then, all the weights in the neighborhood of B_z are updated to more closely match the input vector y_i . The neighborhood could also be denoted as a function of time by $B_z(t)$ because at the beginning of the training the radius of the neighborhood is recommended to be wide to provide a coarse solution and as time goes by the neighborhood is decreased to provide a more refined update. This training process is repeated until a number of iterations defined by the user is reached.

Once the SOM is trained, all of the instances are presented to the SOM model and each is assigned to a single output node k that it matches most closely based on Euclidean distance, each of these mappings is called a node activation. This is the reason why SOMs are also seen as a dimensionality reduction technique. Once the SOM is trained each of the output nodes represent an affective state pattern expressed in the form of weights combinations. Moreover, the SOM has generalization properties because a new and never seen instance can be presented to the trained SOM and it will be assigned to one of the output nodes following the same minimum Euclidean distance criteria to select a BMU.

Finally, there is no a specific procedure to define the total number of output nodes. Each output node is a cluster itself and, therefore, large maps (maps with many output nodes) produce small but compact clusters while small maps have less output nodes and, therefore, instances assigned to each node will tend to be less compact. In unsupervised learning, compactness is a measure that is used to evaluate cohesion and as a consequence explain how closely related the instances inside a cluster are. Moreover, there are other evaluation measurements expressed in terms of cluster separation (isolation) to determine how well separated the clusters are from each other [56] .

3.3 Multivariate Analytical Methods.

3.3.1 Multivariate Event-Crossover

The objective of the next methodology is to see if the decision making around events can alter the physiological signal produced by the subject or the opposite, if the physiological signals interfere with the participant's decision making. The main disadvantage of the original approach is that correlations between features are not considered, hence, univariate tests can miss multivariate patterns. In the univariate ECO each of the attributes was analyzed separately and conclusions were made independently. Instead on analyzing the different input signals separately we are using all the multivariate physiological signal distribution to make inferences about the participant's cognitive state. If the differences are significant the conclusion would be that the multivariate distribution of the physiological input around a certain event is significantly different from that of a control point or interval. Knowing about the different affective states of the participant around events where an ITS can be aware of these behavioral cues is a prerequisite to design and implement adaptive systems with the ability to respond to user's needs [15]. Moreover, given the complexity of the task at hand focusing the analysis on specific events can provide a common starting point among participants and their decision making process.

For the proposed multivariate approach, once again we define n_e as the total number of events found during a session. The event intervals are indexed by $i = 1, 2, \ldots, n_e$. Furthermore, assume that a vector of Q physiological features are measured in each time interval. The physiological measurements for each of these intervals is given by y_{eiq} where $q = 1, 2, \ldots, Q$ is the index for the features. For each of these measurements a control interval without an event is randomly chosen within each participant to form the event-control pair. Similarly, the notation y_{ciq} is used to denote measurement q at the control interval that corresponds to event interval i. The subscripts e and c are used to denote an event and a control interval respectively. We define the difference as $d_{iq} = y_{eiq} - y_{ciq}$ and the vector of differences as $\mathbf{D_i} = [d_{i1}, d_{i2}, \ldots, d_{iQ}]^\intercal$. In addition, if we let $E[\mathbf{D_i}] = \boldsymbol{\delta}$ it is of interest to test $H_0 : \boldsymbol{\delta} = 0$ and $H_1 : \boldsymbol{\delta} \neq 0$. In other words, we are testing if the mean differences are zero. If mean differences are zero, the conclusion is that there is no difference between the physiological measurements at the event and control intervals. If the difference is significant, then we can further explore if the physiological measurement tends to be higher or lower at the event interval when compared to a random control point. We reject H_0 if:

$$T^2 = n_e \bar{\boldsymbol{D}}^\intercal \boldsymbol{S_d^{-1}} \bar{\boldsymbol{D}} > \frac{(n_e - 1)Q}{(n_e - Q)} F_{Q, n_e - Q}(\alpha) \tag{3.1}$$

where

$$\bar{\mathbf{D}} = \frac{1}{n_e} \sum_{i=1}^{n_e} \mathbf{D_i} \qquad \text{and} \qquad \mathbf{S_d} = \frac{1}{n_e - 1} \sum_{i=1}^{n_e} (\mathbf{D_i} - \bar{\mathbf{D}})(\mathbf{D_i} - \bar{\mathbf{D}})^\intercal \tag{3.2}$$

The relation between the physiological signal and the event can also be analyzed before and after the event happens. We can define a hazard period as w intervals before or after an event interval. That is, we denote $y_{e,i-w,q}$ and $y_{e,i+w,q}$ as the physiological measurements w intervals before and after events, respectively. Tests between events and controls can also be conducted for different w's.

In order to analyze the stability of this methodology, similar to the univariate ECO, for the multivariate ECO we run R replicates. At each replicate, a new set of random control intervals is selected and matched with the intervals containing events and the mean difference vector $\bar{\mathbf{D}}$ and the p values given by the T^2 statistic are recorded. Boxplots then can be used to explore the central tendency and the dispersion of each of the vector of mean differences $\bar{\mathbf{D}}_r$ and the p values denoted as α_r for each of the r replicates. In order to define the number of replicates we recommend observing the convergence of the average of the mean differences vector $\sum_{r=1}^{R} \bar{\mathbf{D}}_r/R$ as well as the convergence of the $\sum_{r=1}^{R} \alpha_r/R$ average. Empirically we have observed that with $R = 15$ both averages stabilize and we have enough samples to make use of boxplots to observe the central tendency and dispersion for the mean difference vectors $\bar{\mathbf{D}}_r$ and the α_r in each of the r replicates.

3.3.2 Self-Organizing Maps with Event-crossover

We are interested in discovering patterns around events that we can interpret. We are also interested in finding if these patterns around events are different from those without events. Therefore, the question is whether certain affective patterns are significantly different from other affective patterns at randomly chosen control points. SOMs are a good way of representing a continuous multidimensional physiological signal into discrete weight representations. As explained in the background section, each node activation represents a pattern at a certain point in time. The presence or absence of these activations can indicate certain cognitive patterns among participants given the weights observed in the corresponding node. In this section, we propose what could be considered as a discrete version of the ECO, where the multivariate input vector is decomposed into discrete time events using a SOM similar to that performed by [57].

In this approach, instead of using the original physiological signal, we use a SOM to find the different sources of variation or patterns by lowering the high dimensional feature vector into a two-dimensional space. The main motivation of using this approach is interpretability. The two-dimensional grid formed by the SOM enables us to see different patterns that can be more interpretable by the domain expert in the form of weights and their magnitude. In addition, because the topographical properties of the higher dimension are preserved in the two-dimensional space output nodes in the same neighborhood B_k would tend to have similar patterns in terms of weights. Besides interpretability another advantage of this method is that we can reduce a high-dimensional input signal to a few of patterns (nodes) learned empirically. The hexagonal grid is preferred over the square shape because it provides six adjacent neighbors enhancing isotropy quality and providing smoother maps [58]. In the discrete approach, a SOM is trained with the feature vector constructed with the physiological signals. This vector could contain not only the raw signal but also the features that the experimenter may want to include based on knowledge domain. Linear combinations of the original features could also be used to train the SOM. Once the SOM is trained we have the final weights w_{kq} according to the topography initially defined.

Modern EEG headsets such as Emotiv and ABM B-alert provide measurements that are associated with certain affective states such as: engagement, workload, drowsiness, meditation, excitement and frustration [10, 42, 9]. Having information about which states are activated around specific events and how they differ from intervals where there are no events open new opportunities for researchers and designers to explore different systems configurations and interfaces [59] . One way of doing this is to observe which nodes in the SOM output layer are activated (or not activated) before, during and after a specific event and which nodes are activated for their control intervals. Each output node represents an affective state pattern expressed as a weight combination.

Similar to the ECO methodology, we find all the intervals which contain an event and for each of these intervals we randomly choose within the same participant a control interval without any event. Hence, the total number of event intervals is equal to the total number of control intervals and it is denoted as n_e. Table 3.1 shows the four possible combinations where $n_{ab}, a = 1, 2, b = 1, 2$ represents the

Table 3.1: Contingency Table for a Specific Node k.

Node k activated during control interval Node k activated during event interval	True	False
True	n_{11}	n_{12}
False	n_{21}	n_{22}

counts (frequencies) of activations for an event and the corresponding control interval. The total number of activations is $n_e = n_{11} + n_{12} + n_{21} + n_{22}$. Once we have the contingency table, we compute the proportions $\pi_{ab} = n_{ab}/n_e$ of activations for each node k. The challenge is to know if the activations at event intervals are significantly different from the control intervals. In others words, we want to know if the affective state of the participant at the event is significantly different from intervals without events. One statistic technique that can be adapted to analyze this type of data is McNemar's test [60]. This test can be used for judging differences between correlated proportions. This test takes into account that the marginal proportion for events activations $(\pi_{11}+\pi_{12})$ and the marginal proportion for controls activations $(\pi_{11}+\pi_{21})$ are not independent because the event-control interval pairs are sampled within the same participant. The null hypothesis states that the two marginal probabilities for activations at events and activations at controls are the same: $\pi_{11} + \pi_{12} = \pi_{11} + \pi_{21}$. Because we have the same term π_{11} at both sides of the equality the null hypothesis can be stated as: $H_0 : \pi_{12} = \pi_{21}$ and $H_1 : \pi_{12} \neq \pi_{21}$. If the null hypothesis is true, this would imply that the affective state expressed as a discretized pattern is not different at an event when compared to a control interval. On the other hand, if the null hypothesis is rejected then we can conclude that the affective state at the event has a significantly different pattern when compared to a control point. Moreover, if the counts for events interval is higher than the controls $n_{12} > n_{21}$ we can say that pattern k is present during the event. On the contrary, if the counts of node activations is low at the event $n_{12} < n_{21}$ then this implies that pattern k is absent at the event or is very uncommon. Thus, the McNemar's test statistic is defined by

$$Z_0 = \frac{(n_{21} - n_{12})^2}{n_{21} + n_{12}} \sim \chi_1^2 \tag{3.3}$$

Here Z_0 is a random variable distributed as a chi-squared with 1 degree of freedom. The test is repeated for each of the k nodes to find significant differences in the proportion of activations.

The McNemar's test works only for 2×2 contingency tables and therefore can only consider the proportions of activations for a single output node k or in other words, the frequency a specific cognitive pattern is present or absent around an event. We could instead be interested in whether the distribution of affective states at events and the distribution of affective states for control points are in general the same across all nodes. The Generalized McNemar or Stuart Maxwell test [61, 62] can be used to test for differences in a distribution of K nodes. This test compares all the marginal probabilities for the event intervals and the control intervals. If the marginal probabilities are equal, then the proportions are said to have marginal homogeneity [61, 62] . This test is well suited for data that comes from repeated measures of subjects in which due to the matching of a control point the two samples are not statistically independent. Table 2 shows the information needed to compute the Generalized McNemar or Stuart-Maxwell test where each of the column/row correspond to a node k in the output layer of the SOM and the values at each cell are the counts of the activations for an event interval and its corresponding control interval expressed as $n_{ab}, a = 1, 2, \ldots, K, b = 1, 2, \ldots, K$. The marginal frequencies

Table 3.2: Counts of Activations for an Event-Control Interval.

Node k activated during control interval Node k activated during event interval	1	2	$\cdots$	K
1	n_{11}	n_{12}	$\cdots$	n_{1K}
2	n_{21}	n_{22}	$\cdots$	n_{2K}
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
K	n_{K1}	n_{K2}	$\cdots$	n_{KK}

are the row and column totals that are obtained in the following way: $\sum_{b=1}^{K} n_{ab} = n_{a+}$ and $\sum_{a=1}^{K} n_{ab} = n_{+b}$. In order to perform the Generalized McNemar's test we first define $\mathbf{g}$ as the $(K-1) \times 1$ vector of differences of the form $g_a = n_{a+} - n_{+a}$ for $a = 1, 2, \ldots, K-1$. The vector of differences is important because it contains the information to know if the distribution of affective states (expressed as weights in the output nodes) at events is different from the distribution found at random intervals without events. The Stuart Maxwell test (unlike the McNemar's test) evaluates these marginal frequencies simultaneously. If there is no significant difference between the affective state at an event interval and the control interval the vector of marginal differences is expected to be zero. On the other hand, if, for example, g_1 is positive then this implies that $n_{1+} > n_{+1}$ or that the proportion of node 1 activations is higher for events than for control points. Furthermore, define $\mathbf{S}$ as the $(K-1) \times (K-1)$ matrix of variances and covariances of the elements of $\mathbf{g}$. The hypothesis test is that $H_0 : (\pi_{a+} - \pi_{+a} = 0)$ and the alternative hypothesis is $H_1 : (\pi_{a+} - \pi_{+a} \neq 0)$ for $a = 1, 2, \ldots, K$. The hypothesis testing is important because it can detect patterns at events that can help monitor, analyze and respond to covert psychophysiological activity from the participant in real time. These patterns can be detected without any overt response from the participant and they can be used as a control signal for a ITS [15]. The Stuart Maxwell statistic is then defined as:

$$Z_1 = \mathbf{g}^{\intercal}\mathbf{S}^{\mathbf{-1}}\mathbf{g} \sim \chi^2_{K-1} \tag{3.4}$$

where Z_1 has a chi-squared distribution with $K-1$ degrees of freedom. We reject the null hypothesis if $Z_1 > \chi^2_{K-1,\alpha}$. When the number of categories or nodes is $K = 2$, this reduces to McNemar's test.

3.3.3 Comparing Performance Around Events

In the previous sections different approaches were proposed to analyze affective patterns around specific events considering all participants. However, we could also analyze the affective states around events dividing the participants between two groups. The objective in the following methodology is to test for differences at events for two different groups. Depending on the experiment under study these groups can be represented by: gender, level of expertise, type of training, type of treatment, etc. Analyzing differences between groups can shed some light about how to improve ITS and better understand the different group dynamics. For example, there are educational materials which aggravate differences between groups and put minority groups and women at a disadvantage [63].

As an example, consider two groups: participants who succeed in a task and participants who fail. Let y_{si} for $i = 1, 2, \ldots, n_s$ and y_{fg} for $g = 1, 2, \ldots, n_f$ denote the physiological measurements at the event intervals for success and failure groups respectively. Here n_s and n_f are the total number of events for success and failure groups respectively. Then we can apply a two sample t-test with $H_0 : \mu_s = \mu_f$ and $H_1 : \mu_s \neq \mu_f$. If we reject the null hypothesis, then we can conclude that the mean of the physiological signal for participants in the successful group (μ_s) around a specific event is significantly different from the mean of the physiological signal for the other group (μ_f) at the time of the same type of event. In the learning science context this could imply that participants from two groups have a different physiological response at the time of the event under analysis and this knowledge can potentially be integrated into ITS to customize and adapt a learning environment.

We can perform multiple comparisons in the univariate case. In fact, a test needs to be performed for each type of event defined in the learning environment as well as for each physiological signals. If the total number of physiological signals is Q and the total number of different types of events is E then the total number of tests is QE. However, care should be taken when performing many hypothesis tests because the likelihood of incorrectly rejecting a null hypothesis (Type I error) increases with the number of hypotheses tested. One strategy to compensate this multiple comparison problem is to use the Bonferroni correction [64]. This correction maintains the overall error rate at a desired level of statistical significance. The statistical tests which are being performed may be dependent or independent because no assumption about dependence is made between p values [65]. Another strategy would be to use a multivariate approach as explained below.

This process can also be extended to the multivariate approach where we follow a similar procedure but instead of analyzing the physiological signals separately, we perform a multivariate two-sample Hotelling's T^2 test for differences between groups (success and failure). In this case, instead of a single value for a physiological signal q we take all the available Q physiological signals. In the same way we define the $Q \times 1$ vectors $\mathbf{y_{si}} = [y_{si1}, y_{si2}, \ldots, y_{siQ}]^\intercal$ and $\mathbf{y_{fg}} = [y_{sg1}, y_{sg2}, \ldots, y_{sgQ}]^\intercal$ for success and fail participants, respectively. If we let $\boldsymbol{\mu_s}$ be the unknown vector of means for the successful group and $\boldsymbol{\mu_f}$ the mean vector for the other group then we can test the hypothesis $H_0 : \boldsymbol{\mu_s} = \boldsymbol{\mu_f}$ and $H_1 : \boldsymbol{\mu_s} \neq \boldsymbol{\mu_f}$.

In the univariate case, multiple comparisons need to be performed for each type of event and physiological signal combination. In the multivariate case, the comparison is only done by type of event because all the physiological signals are being considered on a single input vector. The multivariate approach dramatically decreases the number of tests and provides further advantages already explained in the multivariate ECO such as: ability to detect interactions, reveal multivariate differences and can help to control for type I error.

3.4 Experimental Protocol

In the following section we illustrate an example of the methodologies previously explained. In our experiment we are using the affective states provided by ABM B-alert series EEG headset and using a damage control simulator as our HCI or learning environment.

3.4.1 Learning Environment

The Damage Control Simulator (DCS) was created by researchers at UCLA's National Center for Research on Evaluation, Standards, and Student Testing (CRESST). The main task in this simulator is firefighting aboard a ship and the participant's specific objective is reporting, putting out and preventing re-occurrences of fires. In this simulator the participant has to make several critical decisions in order to successfully complete the missions. For this purpose, the participant has available different team members in the ship repair locker but only a few are essential in firefighting. One of the main members is the scene leader (SL) whose principal activities are: report the fire, request and set isolation, manage and request other teams, request testing, debrief and final report. Among the team members the SL is the only capable of re-

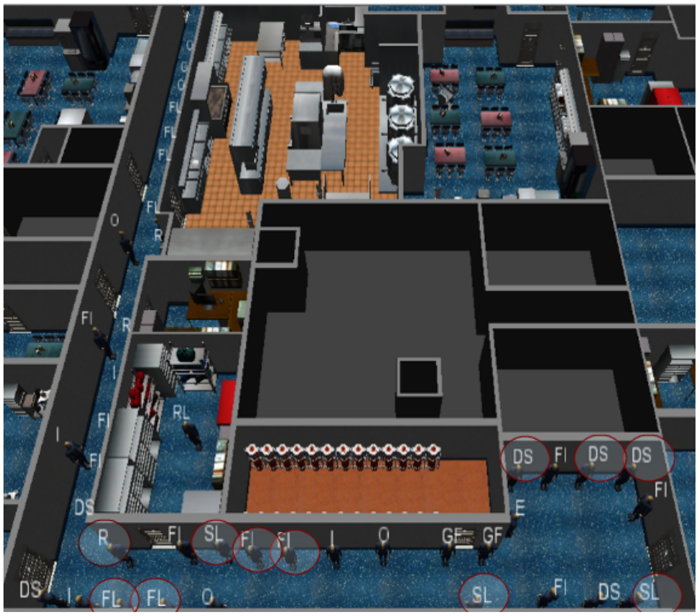


Figure 3.2: Ship Repair Locker of the Damage Control Simulator (DCS).

questing the fire fighter team (FI), desmoke team (DS) and reflash (R). The FI team is normally comprised of three persons and they are mainly responsible of putting out the fire. The DS is a two person team and they are responsible for clearing the area of smoke during and after the fire is put out. The reflash is normally one person who watches over extinguished fire to insure it stays out.

The participants are trained to follow 5 steps when fighting a fire: 1) reporting, 2) choosing the right equipment, 3) putting out a fire, 4) removing smoke from the area and 5) debrief/final report. The participant is also introduced to the 3 types of agents he has available for fighting the fire: carbon dioxide (CO2), aqueous fil forming foam (AFFF) and potassium bicarbonate (PKP). Furthermore, the participant is instructed to select any of these agents according to different types of fires: alpha fire (white or gray smoke, ash producing material on fire), bravo fire (black smoke, flammable liquid on fire) and charlie fire (blue smoke, electrical fire).

The DCS has a catalog of equipment where to choose from. The options vary

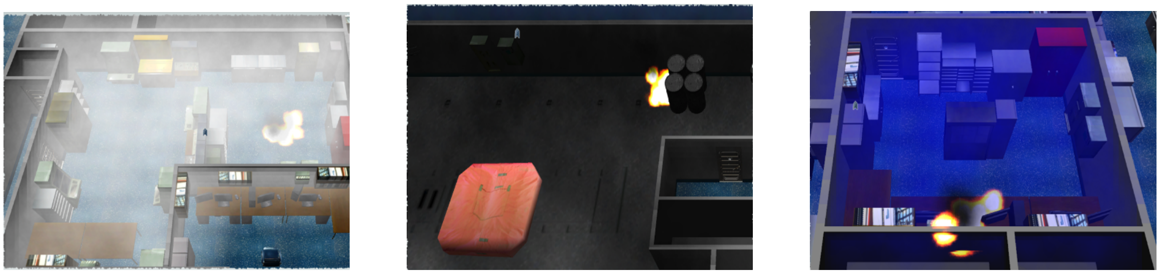
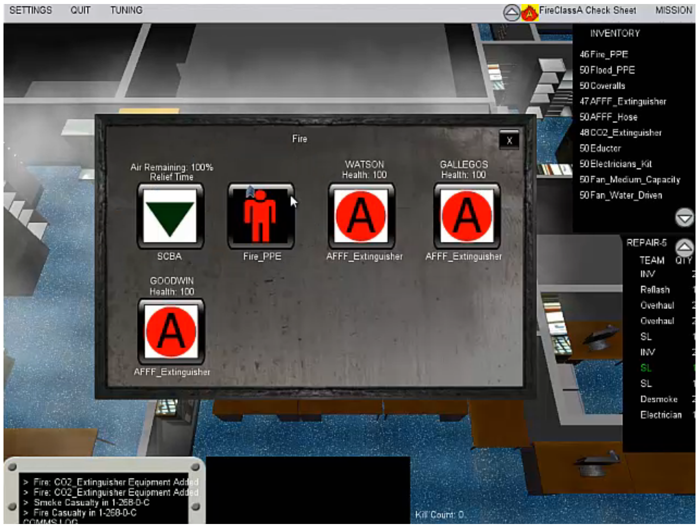


Figure 3.3: Types of Fire (Alpha, Bravo and Charlie).

from type of extinguisher, hose and fans to clothing, headgear, etc. The simulator also provides real time feedback such as the fire health (0 meaning the fire was put out and 100 meaning the mission was failed) and smoke intensity. The participant is supposed to perform several activities before sending a team to attack the fire such as: testing the air is on, testing the type of agent, checking for electrical isolation among other activities. Monitoring the health of each member of the team is also critical and the participant should make the decision of taking a member out of the fire area if the person's health is in danger. The mission is successfully completed when the team is able to put out the fire. At this point the SL debriefs and sends the final report. As we can see the simulation environment is complex because it tries to resemble the situation a firefighter would normally find in a real scenario. Moreover, the time and the tasks are not fixed and they don't have to follow a specific order. Therefore, the traditional methods like time-locked ERP for analyzing events using an EEG fail and more flexible techniques that are able to consider short and long term affective states need to be considered.

The simulator keeps a log regarding the activities and times they were executed. Among the main activities that are tracked are: report, request fire team, set zebra, investigate, request set boundaries, test agent, turn air on, report casualty out, request desmoking team, request reflash watch, return to station and check equipment. The



INVENTORY
46 Fire_PPE
50 Flood_PPE
50 Coveralls
47 AFFF_Extinguisher
50 AFFF_Hose
48 CO2_Extinguisher
50 Eductor
50 Electricians_Kit
50 Fan_Medium_Capacity
50 Fan_Water_Driven

REPAIR-5
TEAM QTY
INV
Reflash
Overhaul
Overhaul
SL
INV
SL
SL
Desmoke
Electrician

```
> Fire: CO2_Extinguisher Equipment Added
> Fire: CO2_Extinguisher Equipment Added
> Smoke Casualty in 1-268-0-C
> Fire Casualty in 1-268-0-C
COMMS LOG
```

Figure 3.4: DCS Equipment Selection Window.

simulator provides an output file with the timestamp during the simulation at which each of these activities were performed. Consequently, it is possible to analyze events off-line with high precision.

3.4.2 EEG and Constructs

Although other methods to analyze the brain are normally used such as near-infrared spectrography (fNIRS), functional magnetoresonance imaging (fMRI) and magnetoencephalography (MEG), EEG headsets have proved to be practical, noninvasive, safe, portable and low cost devices [66] . Traditionally, EEG has been used to analyze very short term events under carefully controlled laboratory conditions in the form of event related potential (ERP). In contrast, video game and simulator sessions last more than a couple of milliseconds, sometimes minutes or even hours and as a consequence different techniques are needed to take into account a more long term approach. For this purpose, several portables headsets exist that provide with mid and long term affective constructs which are derived by building classification or discriminative models using a reasonable sample size of participants. For example, [57] tried to model neurodynamics in submarine navigation teams using a measurement of engagement provided by a low cost but medical grade EEG headset.

For this experiment the ABM B-Alert X10 series were used. This device is 9-channel EEG headset comprising the midline and lateral EEG sites with an optional channel for ECG, EMG or EOG data. The sampling rate is 256 Hz and allows a wireless signal transmission (up to 10 meters) via Bluetooth. The headset requires the application of electro conductive gel to the sensors which are positioned in: Fz, F3, F4, Cz, C3, C4, POz, P3 and P4 locations. B-alert provides 4 classifications or affective constructs: high engagement, low engagement, distraction and drowsiness as well as three measures for workload: workload average, workload FDS (forward digit span) and workload BDS (backwards digit span). Average workload is computed by ABM as the average between the workload FDS and workload BDS. Depending on the task the participant is performing engagement and workload have been shown to work either concordantly or independently [10] . ABM requires all participants to complete three neurocognitive tests: three-choice vigilance test (3CVT), eyes-open (EO) and eyes-closed(EC). This process creates definition files that are needed to compute the B-alert affective states and workload in real time and offline. The classification models developed by ABM use general features from the tested population but also uses subset of additional features from the subjects' baseline to accommodate for individual differences, [10, 9]. The affective states as well as the workload measures are given as a probability (range from 0 to 1).

3.4.3 Participants and Protocol

A total of 60 participants were recruited from the Arizona State University from which 31 were female and 29 males. Ages ranged from 18 to 31 years old with mode of

19 and average of 20.7 years old. Participants were compensated and had the option to leave at any time during the study. Participants were required to participate in two different sessions with a maximum of two weeks between sessions. The purpose of session 1 was to train the participant in the concepts of firefighting, to get familiar with the simulator user interface and to collect demographic information. First, the participant undertook a tutorial for the damage control simulator developed by our team. This tutorial contained information about the different types of fires, team names and their responsibilities (scene leader, fire team, desmoking team, etc.) as well as the type of equipment. Once the participant was aware of the terminology he was asked to play a tutorial embedded in the simulator which was more oriented to get familiar with the keyboard and mouse commands to perform the desired activities in the simulator. The participant then had to play a session where the settings were pre-adjusted to make it easier (slow fire grow, little smoke, high efficiency in the extinguisher, etc). Subsequently, the participant was asked to successfully complete at least one game which was normally done in 2 or 3 attempts. The reason for doing this was that we wanted the participant to be familiar with the controls during session 2 so her affective states reflected the decisions made during the game instead of the struggle to remember how to use the keyboard and the mouse appropriately. Session 1 finished with a pre-test to know how much knowledge the participant acquired and a demographic survey. Session 1 lasted between 45 and 60 minutes and the participant didn't wear any type of physiological equipment.

In session 2 the participant wore the ABM headset. An impedance check was performed at the beginning of the session to make sure the values were below 40 $k\Omega$. Next, participants took a baseline test needed to compute the B-alert classifications (affective states and workload). Participants were required to play 3 scenarios in the damage control simulator representing three levels of difficulty: easy, moderate and difficult. The difficulty was modified changing different parameters such as: the intensity of the fire, the grow rate of the fire, the intensity of the smoke and the efficiency of the equipment among others. The difficult level was set up expecting to have 50% of the participants failing. The purpose of doing this was to challenge participants to make fast decisions and to induce pressure.

3.4.4 Features

Out of the seven constructs provided by ABM B-alert three were selected: high engagement, distraction and workload. Low engagement is negatively correlated by high engagement. Preliminary analysis also showed that participants who failed tended to have higher levels of distraction, as a result it was included. The feature vector is built similarly to the approach followed by [52] with the difference that in our case heart rate was not included and at the same time they didn't include distraction.

Time was partitioned in intervals of $h = 1s$. ABM provides two affective states measurements per second, therefore we took an average of these two values to provide a single measurement by interval. A feature vector was constructed from these three affective states by taking the measurement 3 intervals before and 3 intervals after current time. The objective was to reduce autocorrelation from consecutive intervals. We empirically observed that 3 intervals was a good tradeoff between reducing autocorrelation and still being close enough in time to an event during study. In mathematical notation the feature vector is defined as:

$$\mathbf{y_i} = [y_{i-w,dis}, y_{i,dis}, y_{i+w,dis}, y_{i-w,wl}, y_{i,wl}, y_{i+w,wl}, y_{i-w,he}, y_{i,he}, y_{i+w,he}]^\intercal \tag{3.5}$$

Where $i - w$ and $i + w$ represent the physiological signal w intervals before and after current time respectively. For this specific study $w = 3$, dis represents distraction, wl is workload and he is high engagement. In order to avoid the cluttering notation and to express everything in terms of the current interval i for the rest of the chapter the feature vector is denoted as:

$$\mathbf{y_i} = [y_{i1}, y_{i2}, \ldots, y_{i9}]^\intercal \tag{3.6}$$

each element represents: distraction 3 second before the event (DIm3), distraction at the current time (DI0), distraction 3 seconds after the event (DIp3), workload 3 seconds before the event (WLm3), workload at the current time (WL0), workload 3 seconds after the event (WLp3), high engagement 3 seconds before the event (HEm3), high engagement at the current time (HE0) and high engagement 3 seconds after the event (HEp3).

The length of the sessions varies among participants and it also depends on either the participant completing the mission successfully or failing. Participants who succeeded were, in general, more inclined to last longer, but this was not always true. Sometimes a given participant would finish the mission successfully quickly and other times the participant would take longer time to play only to fail at the end. The different length implies that some participants would tend to have more events during their session than others. If the number of events by participant is unbalanced, the experiment would tend to capture the effect in the physiological response for the participants with higher frequency of events and this may not represent the behavior of all the population. As long as the number of participants is large enough and the number of events is not substantially disproportional between participants then an adjustment might not be necessary. We did not adjust for differences in our experiments. Also, the total number of events have a direct influence in the results because large sample sizes tend to yield significant results.

Finally, the feature vector could be z-scored or mean-centered depending on the domain knowledge. For this study we didn't perform any type of scaling because the measurements are already restricted to the range 0-1 and also normalized by the EEG headset software to account for individual differences using a baseline correction.

3.4.5 Events

The DCS provides a log file which contains a time stamp and type of events as they happened during the simulation. Therefore, we cannot only know with high precision the time an event happens but also the sequence in which they happen. Among the most common events there were: report, attack, request fire team, air on, test agent and debrief/report. The synchronization between the events provided by the log file of the DCS and the affective states output file provided by ABM B-alert series was done with R statistical software using the time stamp of both sources as the reference to merge both files. The time stamp provided in the log file of the DCS allows knowing the exact time up to milliseconds precision when an event happens. Once we know the exact time an event happens when can precisely locate the physiological signal in the ABM B-alert output file searching in the time stamp the closest time indicated by the DCS log file.

In this study, we selected only those events that had high representation, this is, were performed by the majority of the participants. The different frequency of events happens because the tasks or actions can sometimes be performed in a different order or a participant omits an event at a given time. For example, *active desmoking* is a task that very few participants requested because is not completely necessary, but it could in general help to have a better vision of the scene. On the other hand, an activity such as *report* or *request fire team* could be very difficult, if not impossible, to omit.

Table 3.3: Univariate ECO for Participants Who Passed.

		DIm3	DI0	DIp3	WLm3	WL0	WLp3	HEm3	HE0	HEp3
Report (n=42)	*p* value	0.45	0.58	0.061	0.399	0.609	0.591	0.525	0.32	0.042
	mean diff	0.027	0.016	0.092	0.026	0	0.012	-0.036	-0.092	-0.162
Attack (n=179)	*p* value	0.45	0.374	0.623	0.598	0.264	0.123	0.031	0.066	0.564
	mean diff	0.017	0.016	0.007	0	0.02	0.027	-0.093	-0.081	0.015
Request Fire Team (n=79)	*p* value	0.427	0.499	0.341	0.399	0.677	0.298	0.399	0.357	0.138
	mean diff	0.02	0.001	0.032	0.022	-0.004	-0.024	-0.05	-0.054	-0.101
Air on (n=163)	*p* value	0.462	0.572	0.545	0.45	0.587	0.501	0.443	0.41	0.488
	mean diff	0.014	0.008	0.006	0.013	0.003	-0.003	-0.029	-0.03	-0.024
Test Agent (n=94)	*p* value	0.529	0.639	0.617	0.534	0.352	0.407	0.469	0.349	0.506
	mean diff	0.011	0.002	0.009	-0.01	0.02	-0.016	-0.034	-0.042	-0.037
Debrief and Report (n=34)	*p* value	0.44	0.779	0.426	0.483	0.148	0.492	0.5	0.588	0.684
	mean diff	0.031	0.006	0.038	-0.028	-0.059	0.021	0.046	-0.031	-0.013

3.5 Results

3.5.1 Event-Crossover

The univariate ECO was applied to the data set of the *difficult* scenario as explained in the methodology section using $R = 15$ replicates. Table 3.3 shows the results for those participants who passed. We can see that the only significant result for the event *report* ($n = 42$) is high engagement 3 seconds after (HEp3), $p = 0.042$. The table also provides the average of mean differences across those 15 replicates, in this case -0.162. This means that the high engagement 3 seconds after the event *report* happens is in general lower when compared to a random control point. For *attack* ($n = 179$) there is also one significant result in high engagement 3 seconds before (HEm3), $p = 0.031$. The average of the mean difference for the 15 replicates is -0.093 which implies that on average this affective construct is lower 3 seconds before the event *attack* happens when compared to a control point. Fig. 3.5 shows a visual

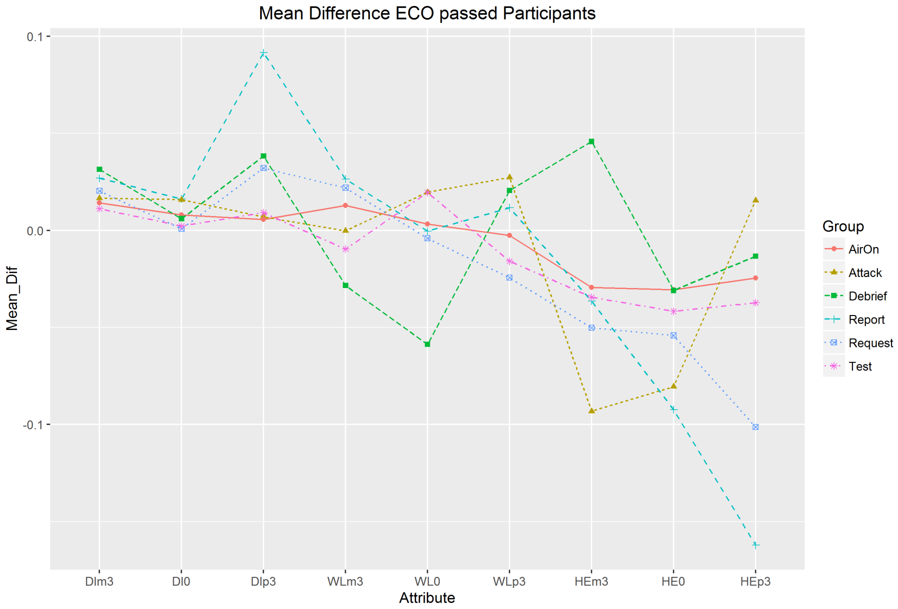
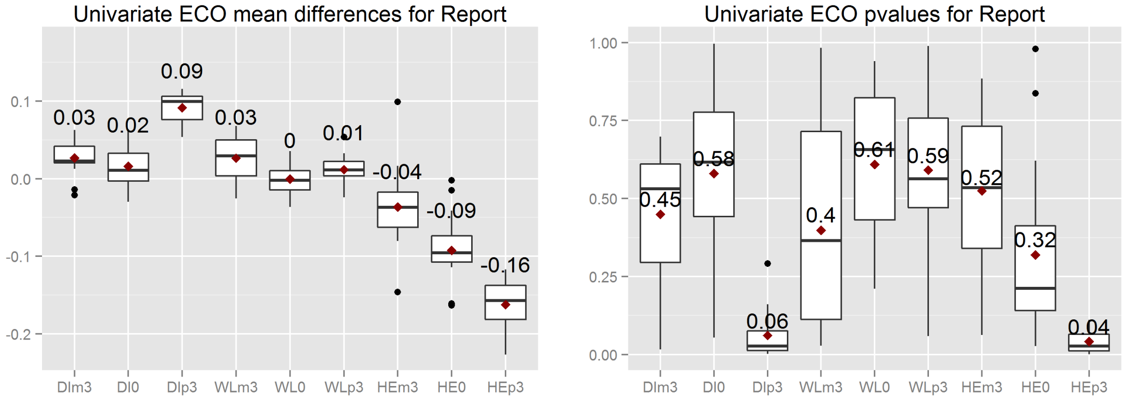


Figure 3.5: Univariate ECO of Mean Differences Plot, Participants Who Passed.

representation of the results presented in the previous table. In this plot the average of the mean differences for those participants who passed are shown for each of the event. At the bottom right of the plot we can see the large difference for the event *report* in high engagement 3 seconds after (HEp3). The event *report* ($n = 42$) for DIp3 also shows a large difference and on average it was close to significance with $p = 0.061$. Another interesting pattern is that the majority of the mean differences tend to be near zero, which implies that no large differences around events where found when compared to a random control point. However, if we look at the high engagement pattern we see that in general this affective construct tends to be lower around all the types of events when compared to a control point. In order to analyze the stability of our methodology, $R = 15$ replicates were run, each time using a different random seed. The rationale behind this procedure is that because each time we are randomly sampling control points there is variability in the results. For each of the r replicates we kept track of the mean differences vector and also the p values

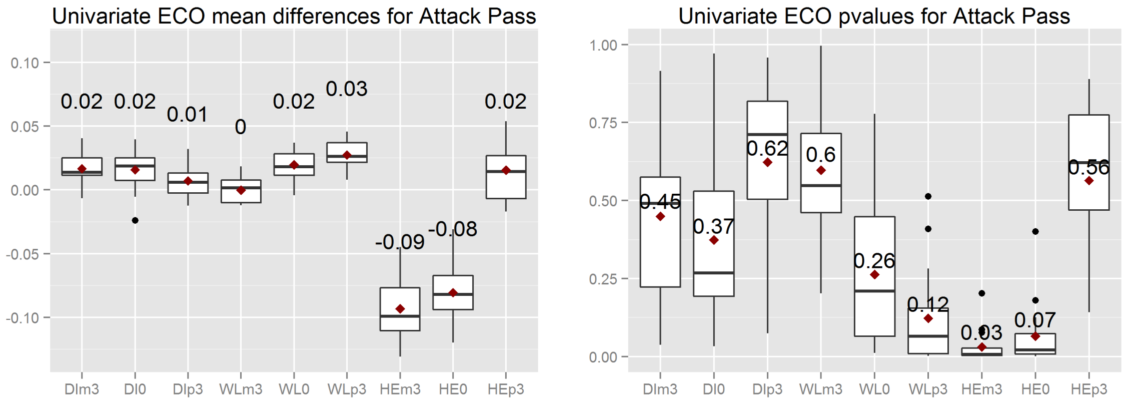


(a)

(b)

Figure 3.6: Box Plots for *Report*, a) Mean Differences and b) p Values, Pass.

for the hypothesis tests. Boxplots of these mean differences and p values for the 15 replicates are shown in Fig. 3.6. In Fig. 3.6a the mean differences for report and participants who passed are shown. Again the majority of the differences are around zero and we can see a trend in the high engagement group. The box plots show not only the central tendency, but also the spread of the data. Looking at the length of the whiskers, we can argue that the results for mean differences across the 15 replicates seem to be very consistent. On the other hand, in Fig. 3.6b the p values for the same mean differences are shown. All features present relative high variation except for HEp3 which is significant and DIp3 which was close to significance. Fig. 3.7 shows the same box plots but this time for the event *attack* ($n = 179$), and again only for those participants who passed. Most of the mean differences shown in Fig. 3.7a are around zero except for HEm3 and HE0 which show a large absolute mean difference, -0.09 and -0.08 respectively. The p values are shown in Fig. 3.7b and we observe that for HEm3 the dispersion of the p values across replicates is very tight with only 1 or 2 outliers outside the whiskers. For HE0 the spread of the p values also looks tight with two outliers possibly pulling the average up to 0.07. Table 3.4 shows results for



(a) (b)

Figure 3.7: Box Plots for *Attack*, a) Mean Differences and b) p Values, Pass.

those participants who failed. This time we see that only one significant differences was detected in the average of the 15 replicates. The event attack ($n = 83$) has a significant average results ($p = 0.041$) for workload current time (WL0). The average mean difference for this event-feature combination is 0.046 which is a moderate magnitude. Fig. 3.8 shows the average of the mean differences for those participants who failed. We see that for the event report ($n = 22$) HEp3 has a large absolute mean difference (-0.125) but the average of the p values across the 15 replicates didn't turn out to be significant ($p = 0.309$). The other event-feature combination that shows large mean difference is *report* in WLm3 but the average p value was 0.144, which although not significant was one of the lowest in the table.

Fig. 3.9a shows the boxplots for the mean difference and the event *attack* ($n = 83$) for those participants who failed. Workload current time (WL0) as well as high engagement 3 seconds after (HEp3) show large absolute differences but only WL0 has a stable and low p value average ($p = 0.04$) as shown in Fig. 3.9b. The rest of the p values are on average large and have a greater variation as seen in the same figure. Perhaps only WLp3 has a low p values but it is still far away from 0.05. Similar to

Table 3.4: Univariate ECO for Participants Who Failed.

		DIm3	DI0	DIp3	WLm3	WL0	WLp3	HEm3	HE0	HEp3
Report (n=22)	*p* value	0.465	0.568	0.368	0.144	0.664	0.665	0.421	0.747	0.309
	mean diff	0.061	-0.021	0.063	-0.07	0.02	0.008	0.021	-0.024	-0.125
Attack (n=83)	*p* value	0.561	0.627	0.484	0.554	0.041	0.169	0.584	0.482	0.553
	mean diff	0.018	0.004	-0.023	0.009	0.046	0.031	-0.022	0.033	-0.036
Request Fire Team (n=71)	*p* value	0.571	0.478	0.545	0.658	0.626	0.68	0.658	0.608	0.568
	mean diff	0.018	0.029	-0.006	-0.008	0.01	0.003	-0.003	-0.018	-0.022
Air on (n=77)	*p* value	0.497	0.191	0.402	0.44	0.226	0.615	0.308	0.505	0.38
	mean diff	0.024	0.06	0.025	0.017	0.033	-0.005	-0.06	-0.031	0.047
Test Agent (n=52)	*p* value	0.484	0.611	0.671	0.467	0.37	0.475	0.596	0.527	0.453
	mean diff	-0.035	-0.012	-0.012	0.023	0.022	0.013	0.025	-0.038	-0.037
Debrief and Report (n=0)	*p* value	NA	NA	NA	NA	NA	NA	NA	NA	NA
	mean diff	NA	NA	NA	NA	NA	NA	NA	NA	NA

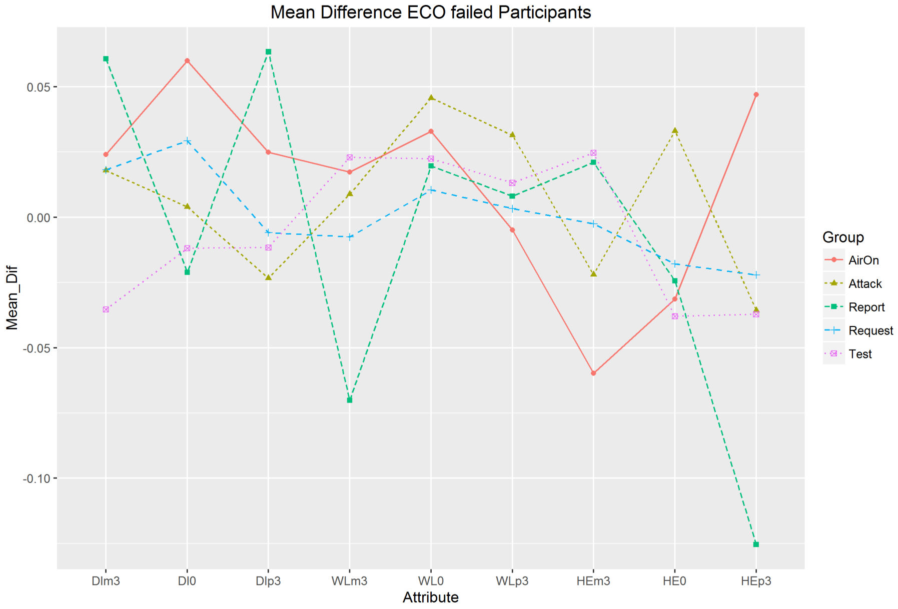
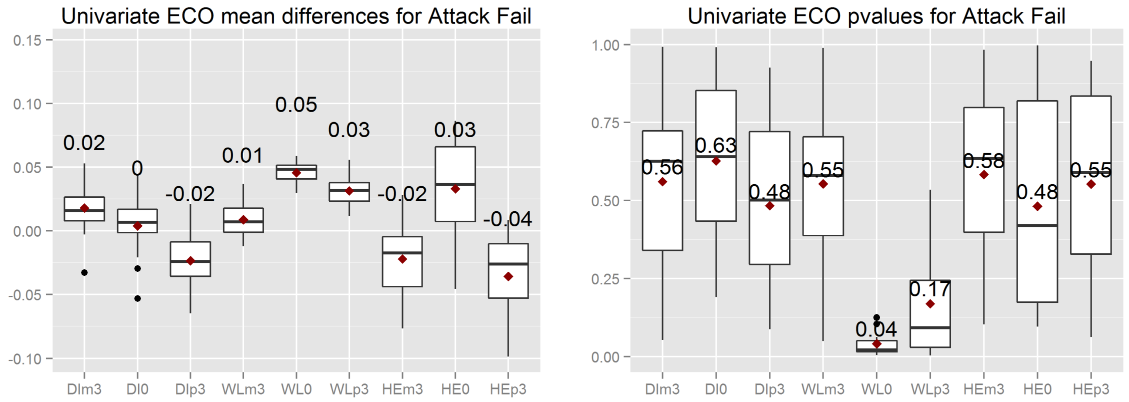


Figure 3.8: Univariate ECO of Mean Differences Plot, Participants Who Failed.



(a)

(b)

Figure 3.9: Box Plots for *Attack*, a) the Mean Differences and b) p Values, Fail.

the univariate ECO, we present the results for the multivariate event-crossover (See Table 3.5). In this case, only *attack* ($n = 179$) was on average significantly different from the control point ($p = 0.02$) for those participants who passed. This result is somehow consistent with the univariate ECO because we also found that *attack* had an average significant difference. The multivariate ECO didn't show significant results for those participants who failed with *attack* and *air on* as the events with the lowest average p value but none of them below 0.05.

3.5.2 SOM-ECO

We trained the self-organizing map using R version 3.2.2 and the package "Kohonen" version 2.0.19 using a hexagonal grid with 3 columns and 3 rows. The feature vector was defined in equation 3.5 and the short notation in equation 3.6. We will use the last notation to express physiological signal interval in terms of i. Because the feature vector was constructed with rows representing seconds, the number of activations we expect to see from each participant is the total number of seconds that they played in the simulator. Moreover, the time of each session is not fixed but

Table 3.5: Multivariate ECO for Participants Who Passed and Failed.

Event	Pass				Fail			
	T^2	df1	df2	p value	T^2	df1	df2	p value
Report	1.71	9	33	0.22	0.79	9	13	0.64
Attack	2.56	9	170	0.02	1.65	9	74	0.21
Request Fire Team	1.77	9	70	0.21	0.53	9	62	0.82
Air on	0.8	9	154	0.62	1.63	9	68	0.2
Test Agent	0.87	9	85	0.56	0.74	9	43	0.68
Debrief and Report	1.42	9	25	0.38	NA	NA	NA	NA

depends on the participant's performance. For this reason the number of activation by participant varies.

The grid is traditionally built using a $\sqrt{K} \times \sqrt{K}$ configuration where K is the total number of desired nodes (patterns) to explore. However, other asymmetric grids can be built. Empirically we observed that a 3×3 arrangement provided good results in terms of compactness in our experiments. We also used a 4×4 SOM, but it provided clusters that were not as well separated. Separation is a measure of cluster evaluation for unsupervised learning which determine how well distant the clusters are from each other [56]. A 2×2 was another setting. In this case the weights for each of the four nodes (clusters) were well separated but the instances assigned to each cluster were not very compact (low cluster cohesion), this is, instances assigned to a cluster were not closely related within the cluster. All the measures in the feature vector range from 0 to 1 so no additional scaling was needed. The node and weights are shown in the following Fig. 3.10. Although the grid looks rectangular the nodes are really on a hexagonal grid. As explained in the background section nodes close to each other

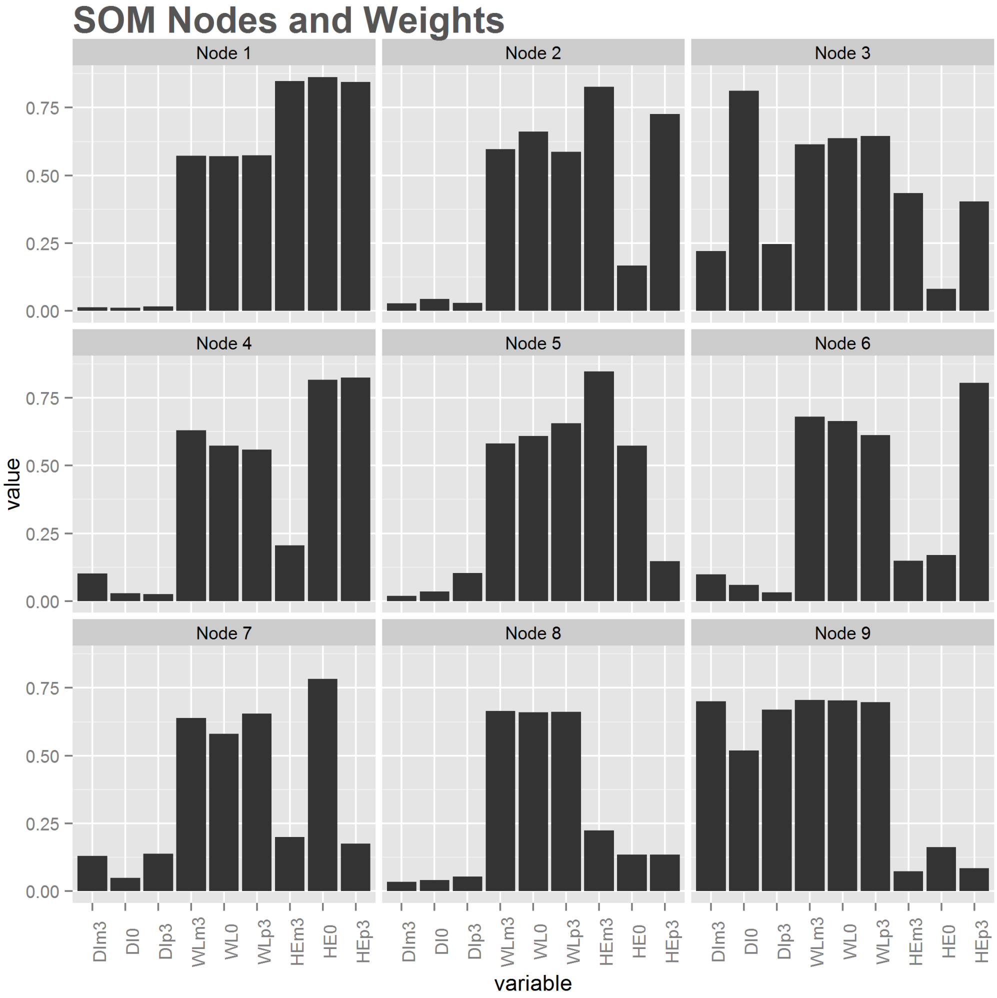


Figure 3.10: Weights for Each of the Nodes in the SOM.

in the two-dimensional grid will tend to have similar weights. We can see that node 1 is a combination of low distraction and moderate workload and high engagement. If we go to the opposite side of the map to node 9, we see, as expected, the opposite configuration with high distraction and high workload, but very low high engagement. An advantage of the SOM is this easy visualization of the nodes. Another interesting fact is that workload never appears low in any of the weight combinations. The total activation counts are shown in the following Fig. 3.11. Node number 1 has the

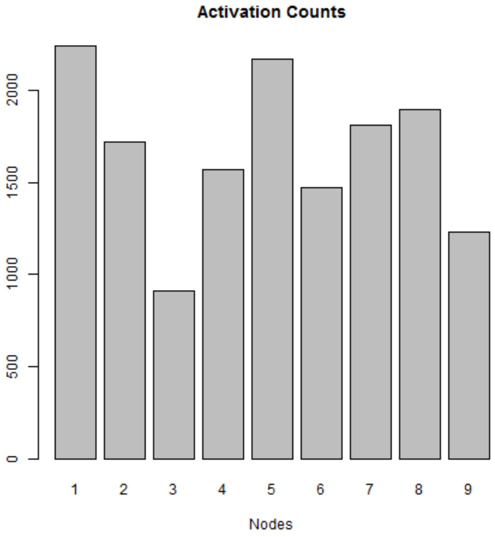
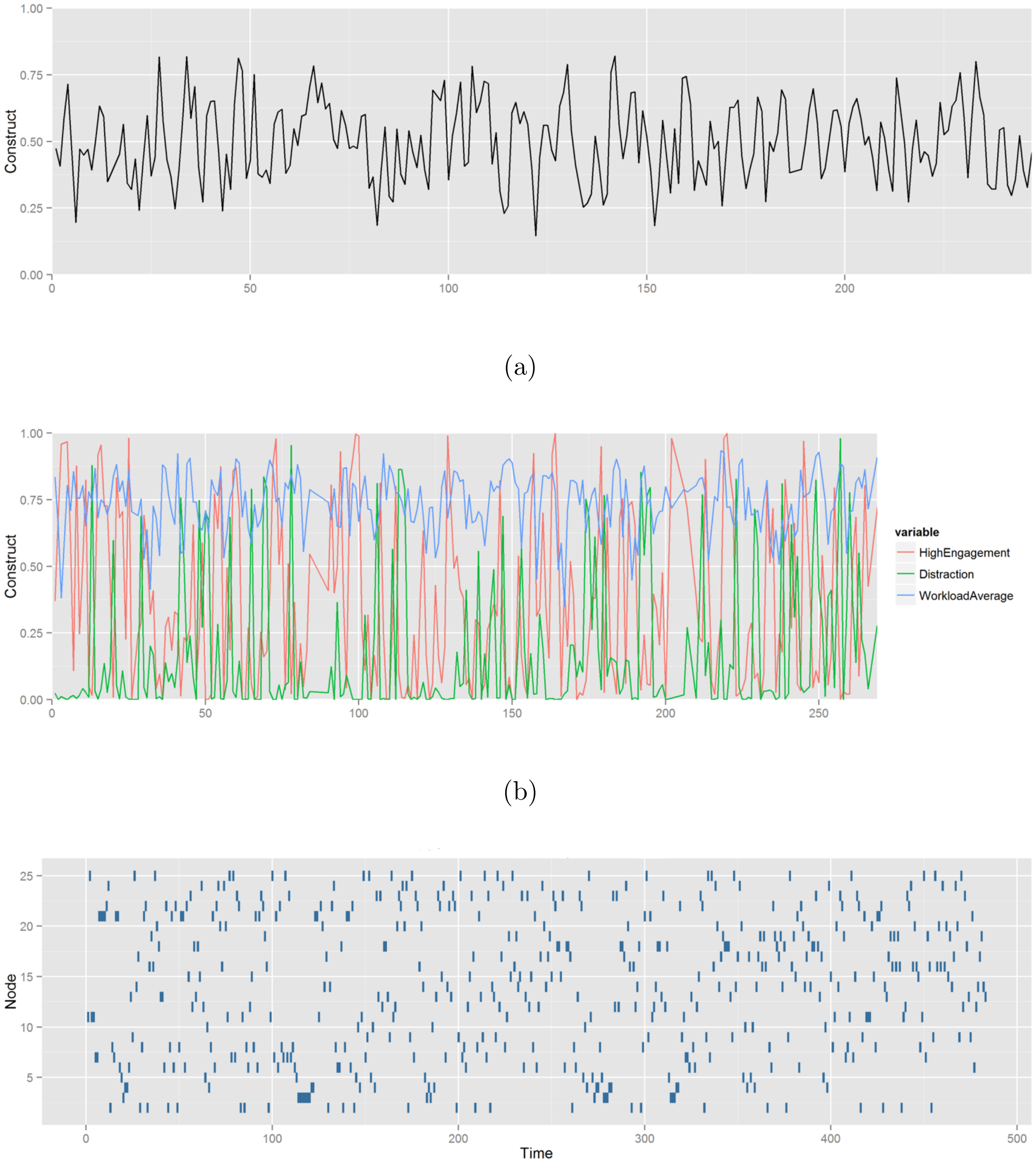


Figure 3.11: Activation Counts.

most activations followed by node number 5. Node number 3 has the least number of activations. Figure 3.12 shows how the node activation can help create a discrete version of the time series. In Fig. 3.12a we see a physiological time series with only one feature (univariate). Fig. 3.12b shows a multivariate time series with three affective constructs: high engagement, distraction and workload. Finally, Fig. 3.12c shows the activations of node k at interval i. Note that there can be only one activation per interval because each time an instance is presented to the SOM it chooses the BMU by computing the minimum Euclidean distance. Moreover, it is easy to see that as the number of physiological signals increase, the more difficult it is for the human eye to detect patterns in the traditional multivariate time series representation as shown in Fig. 3.12b. Table 3.6 contains the p values for each of the event-node combinations. The number of node activations at each of the events is also shown (n). The events appear in the first column as well as the number of events that were used for each computation. This table only presents results for those participants who passed. Node 1 and event *report* ($n = 42$) have a significant results with $p = 0.04$. Fig.



(a)

(b)

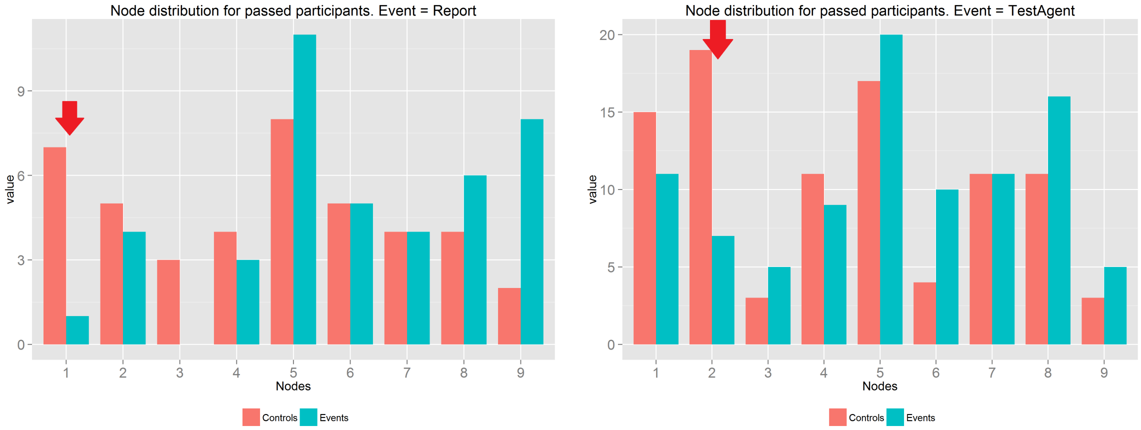
(c)

Figure 3.12: A Time Series of Physiological Input (a), a Multivariate Signal (b) and Node Activations of a 5 × 5 SOM (c).

Table 3.6: McNemar's Test Results for Participants Who Passed.

	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6	Node 7	Node 8	Node 9
Report (n=42)	0.04	1.00	0.25	1.00	0.58	1.00	1.00	0.72	0.08
Attack (n=179)	0.17	0.56	0.82	0.86	0.20	0.74	0.73	0.06	1.00
Request Fire Team (n=79)	0.19	0.26	1.00	0.23	0.69	0.77	0.63	0.06	0.50
Air on (n=163)	0.76	0.39	0.42	0.74	0.43	1.00	0.61	0.20	1.00
Test Agent (n=94)	0.52	0.02	0.72	0.80	0.72	0.18	1.00	0.42	0.68
Debrief and Report (n=34)	0.23	0.34	1.00	1.00	0.23	1.00	0.50	0.37	1.00

3.10 can be used as a road map for the interpretation of the weights. For example, node 1 has almost zero distraction in any of the three distraction features, it also has moderate workload and moderate high engagement. This implies that participants who succeeded the mission normally lack this pattern when they were first reporting the situation when compared to randomly chosen control point. The node 2 at event *test agent* ($n = 94$) has also a significant result ($p = 0.02$). Node 2 is very similar to the previously described node 1 because they are next to each other in the two-dimensional grid. Fig. 3.13a shows the node distribution for the event *report* for those participants who passed. The red arrow shows the large difference between the number of activations for control and events, therefore confirming the results observed in Table 3.6. Fig. 3.13b shows the same type of plot only this time for *test agent*. Again node 2 shows a large difference between controls and events. Table 3.7 presents the results for the McNemar's test for those participants who failed. There are two significant results. The first one is the node 5 at event *report* ($n = 22$) with $p = 0.04$. In Fig. 3.14a we can see there is a large difference in node 5 activations around the event *report* when compared to a control interval. If we observe node 5 in Fig. 3.10 we notice that distraction is very low, there is moderate to high workload and there is a decreasing level of high engagement, going from high 3 second before the event to



(a)

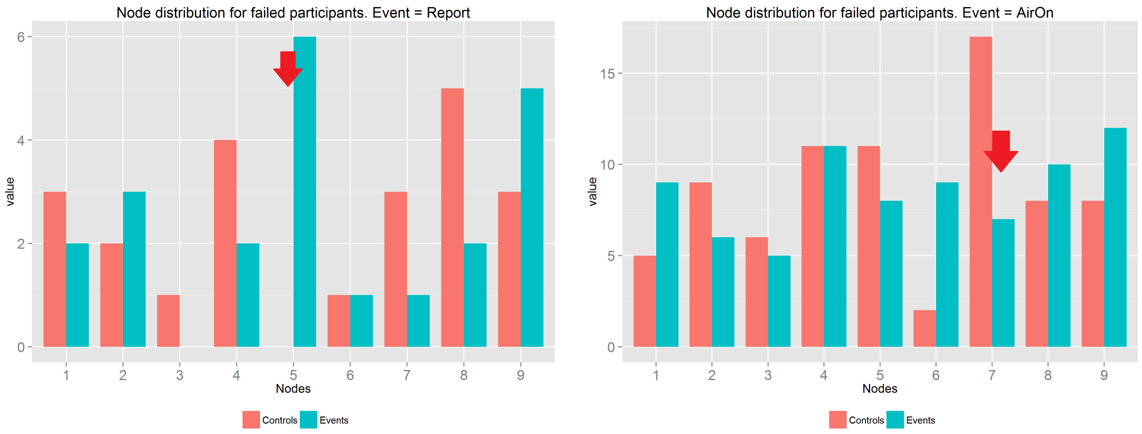
(b)

Figure 3.13: Node Distribution for Controls and Events, Pass.

Table 3.7: McNemar's Test Results for Participants Who Failed.

	Node 1	Node 2	Node 3	Node 4	Node 5	Node 6	Node 7	Node 8	Node 9
Report (n=22)	1.00	1.00	1.00	0.68	0.04	1.00	0.62	0.37	0.62
Attack (n=83)	0.55	1.00	1.00	0.42	0.40	0.27	0.31	1.00	0.34
Request Fire Team (n=71)	0.34	1.00	1.00	0.11	1.00	1.00	0.77	0.79	0.22
Air on (n=77)	0.34	0.58	1.00	1.00	0.61	0.07	0.04	0.79	0.22
Test Agent (n=52)	0.45	0.34	0.22	0.55	0.10	0.68	0.15	0.34	0.37
Debrief and Report (n=0)	NA	NA	NA	NA	NA	NA	NA	NA	NA

low 3 seconds after the event. The other significant result is in node 7 at event *air on* ($n = 77$) with $p = 0.04$. In Fig. 3.14b we can see that the node 7 activations around the event *air on* are unusually low when compared to a random interval. Table 3.8 provides with information about the Stuart-Maxwell test or Generalized McNemar's. The way of interpreting the results of this table is that the distribution of nodes activations around the events under study is significantly different (or not) from the proportion of nodes chosen randomly, where the random selection is done within each participant working as its own control. If we put all participants together we have



(a)

(b)

Figure 3.14: Node Distribution for Controls and Events, Fail.

Table 3.8: Generalized McNemar's Test.

	All	Pass	Fail
Report	0.041	0.065	0.287
Attack	0.403	0.299	0.457
Request Fire Team	0.093	0.075	0.474
Air on	0.335	0.522	0.117
Test Agent	0.029	0.023	0.074
Debrief and Report	0.339	0.269	NA

significant results for *report* $(n = 64)$ and for *test agent* $(n = 146)$ with $p = 0.041$ and $p = 0.029$ respectively. If we only consider those participants who passed then *test agent* $(n = 94)$ is significant $p = 0.023$. There were other results that were close to significance for this group, namely *report* $(n = 42)$ with $p = 0.065$ and *request* $(n = 79)$ with $p = 0.075$. No significant results were found in the fail group, with perhaps *test agent* $(n = 52)$ close to significance $p = 0.074$. In order to analyze the

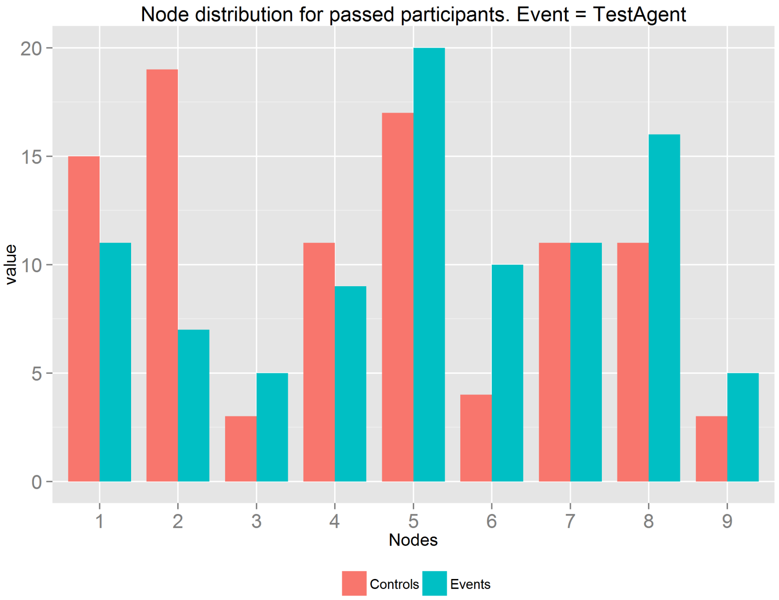


Figure 3.15: Node Distribution for Participants Who Passed at Event *Test Agent*.

finding of Table 3.8, we look to Fig. 3.15 which displays the activations distribution across all 9 nodes for those participants who passed and for the event *test agent*. We can clearly see for example that nodes 1 and 2 have low counts when compared to the controls and that nodes 5, 6 and 8 have relatively high counts around events when compared to controls.

3.5.3 Contrasting Performance

In the univariate pass vs fail results we found that there are significant differences between the affective states of the participants who passed from those who failed around certain events as seen in Table 3.9. For the event *attack* ($n = 262$) there were significant results for all three workload features. There is also a significant result for HEp3 ($p = 0.036$) with mean difference -0.102, where participants who failed had lower values for this feature than those who passed. *Request* ($n = 140$)

Table 3.9: Results of the One-Sample t-Test for Differences in Pass vs Fail.

		DIm3	DI0	DIp3	WLm3	WL0	WLp3	HEm3	HE0	HEp3
Report (n=64)	p values	0.437	0.903	0.866	0.13	0.362	0.705	0.568	0.609	0.817
	mean diff	0.068	0.007	0.015	-0.067	0.042	0.017	0.059	0.052	0.022
Attack (n=262)	p values	0.093	0.155	0.407	0.023	0.002	0.024	0.878	0.239	0.036
	mean diff	0.067	0.054	0.031	0.05	0.061	0.046	0.007	0.055	-0.102
Request Fire Team (n=150)	p values	0.167	0.026	0.393	0.938	0.201	0.181	0.932	0.883	0.5
	mean diff	0.064	0.102	0.037	-0.002	0.033	0.038	-0.005	-0.009	0.039
Air on (n=240)	p values	0.033	0.018	0.042	0.023	0.001	0.074	0.099	0.371	0.559
	mean diff	0.089	0.106	0.087	0.045	0.071	0.039	-0.079	-0.046	0.03
Test Agent (n=146)	p values	0.394	0.15	0.278	0.005	0.074	0.01	0.86	0.71	0.679
	mean diff	0.037	0.068	0.052	0.083	0.046	0.071	0.011	-0.024	-0.026
Debrief and Report (n=0)	p values	NA	NA	NA	NA	NA	NA	NA	NA	NA
	mean diff	NA	NA	NA	NA	NA	NA	NA	NA	NA

has one significant result, *air on* ($n = 240$) has 5 significant results and *test agent* ($n = 146$) has 2 significant results. The event *debrief and report* could not be done (marked as NA's) because those who failed in theory are not supposed to debrief and report because this is the last action that needs to be taken after putting out the fire. In general there were no significant differences between pass/fail participants for the event *report* ($n = 64$). Fig. 3.16 shows the visual summary for mean differences for all events. HEp3 for the event *attack* has a large absolute value and it clearly departs from zero and from the other mean differences. As we saw in Table 3.9 this value turned out to be significant. Distraction in current time (DI0) also has two large mean differences for the event *request* and *air on* both with significant $p$values ($p < 0.05$). Fig. 3.17 shows an additional visual aid. This time the event *air on* is further analyzed by a more detailed data visualization for the simple reason that it has the most significant results from Table 3.9. We observe that distraction is consistently higher for those who failed when compared to those who passed. This could imply

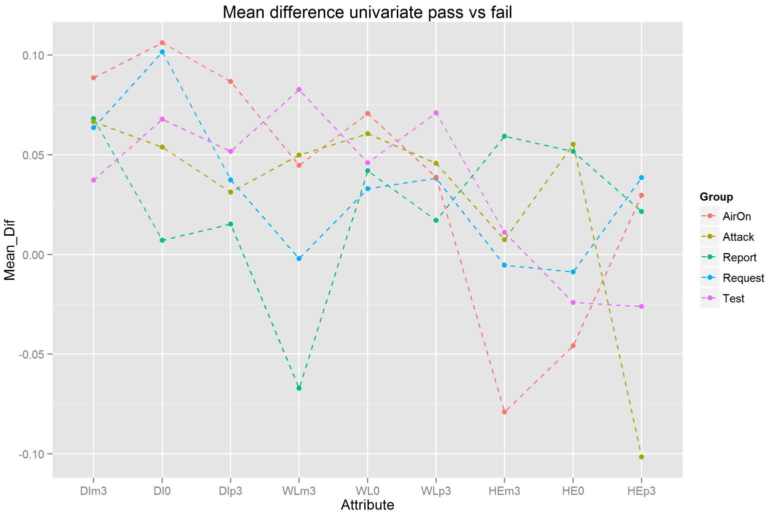


Figure 3.16: Plot for Mean Differences for Participants Who Passed vs Failed.

that more successful players tend to express lower levels of distraction. Another interesting fact is that the distraction feature group presents a greater number of outliers when compared to the other two groups (workload and high engagement). Workload is also higher for those participants who failed and the significance of the results was confirmed in the previous table for WLm3 and WL0. High engagement on average shows similar differences than workload in terms of magnitude but the spread of the data is larger and therefore no significant results were obtained. Each of the other events can be analyzed using this type of plots to increase the detail of analysis. Finally, Table 3.10 presents the results for the multivariate approach for pass vs fail. The event *attack* ($n = 262$) and *air on* ($n = 240$) have significant results with $p = 0.002$ and $p = 0.004$ respectively. No significant results were found in the rest of the events.

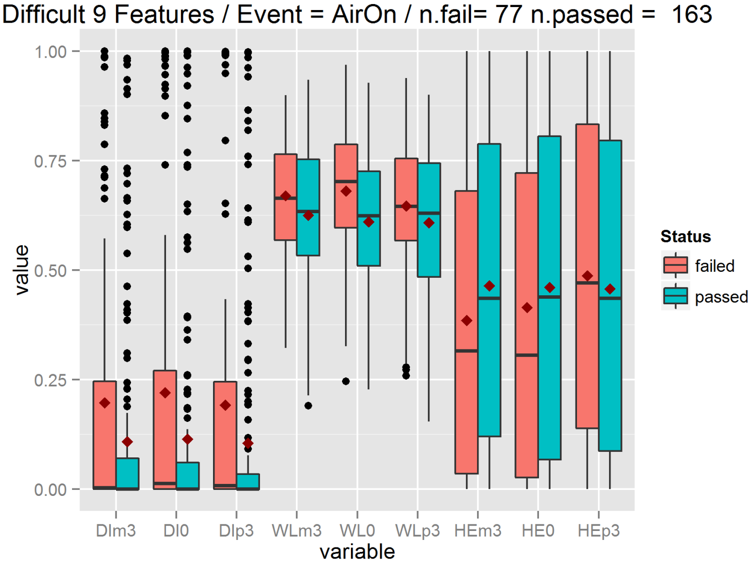


Figure 3.17: Box Plots for Mean Differences Between Pass vs Fail.

Table 3.10: Results for Multivariate 2-Sample Hotelling's T^2.

	Hotelling-Lawley	num Df	den Df	*p*value
Report	0.11	9	54	0.73
Attack	0.11	9	252	0.002
Request Fire Team	0.07	9	140	0.331
Air on	0.11	9	230	0.004
Test Agent	0.11	9	136	0.123
Debrief and Report	NA	NA	NA	NA

3.6 Discussion and Conclusions

The aim of our work was to introduce a set of analytical tools to analyze events in a learning environment and where a physiological input is recorded in real-time. The methodologies proposed here are suitable for situations where events occur randomly and are embedded in time. In this study the methodology was explained in detail and an applied example was illustrated using a damage control simulator. The experiment used data from 60 participants and its objective was to identify differences around events.

The first proposed methodology was the multivariate ECO. First, the univariate ECO was applied where *report-HEp3* and *attack-HEm3* had significant results for those participants who passed. This implies that high engagement 3 seconds after the event *report* happens is lower when compared to a control point. Something similar happens with high engagement 3 seconds before attack, it is in general lower than the control points. The methodology suggests using box plots to further analyze patterns and trends. For example, we learned that in general high engagement for those who passed tends to be lower at events than at control intervals and there are few outliers across replicates. The p value boxplots for *report* shows that DIp3 and HEp3 have low variation when compared to the rest. For participants who failed the mission the event *attack* at WL0 was the only significant ($p = 0.041$) , this implies that this group has higher workload at the exact moment the event happens. This is consistent with other studies where they have found that subjects who are struggling with a learning task tend to have higher cognitive load [67] . On the other hand, in the in the multivariate ECO only the event *attack* ($n = 179$) shows a significant result ($p = 0.02$) for those participants who passed. This is consistent with the univariate approach which also detected this event as significant but only for a single feature.

In the second approach we tried to address the question: Given a multidimensional physiological signal from a subject, what is the affective state of this subject at a given time or event and how can this state be represented? We proposed a discrete version of the ECO methodology where the high dimensional feature space was lowered to a two-dimensional space using a SOM. The characteristics of interpretability and the preservation of topographical properties make this approach suitable for the analysis of physiological signals. Once the SOM was trained the feature vector was presented to the model and the node activations were tracked. The McNemar's test for participants who passed showed that node 1 was significantly different from a random chosen interval for the event *report*. Node 1 is characterized by very low distraction, moderate workload and high high-engagement. In this regard, the number activations at the time of this event are unusually low. The combination node2 and *test agent* also had a significant results. Node 2 is similar to Node 1 with the difference being that it has very low high engagement in the current time. For the participants who failed, node 5 and *report* had a significant result where the proportion of activation for this event is large when compared to the controls. Node 5 is characterized by low distraction, a moderate workload that increases over time, and high engagement that starts high 3 seconds before the event, goes to moderate during the event and falls abruptly after 3 seconds. This is an interesting pattern that requires further analysis because it was observed only in failed participants but not in those who succeeded. Lastly, node 7 at event *air on* also had a significant result.

The Generalized McNemar's or Stuart-Maxwell test showed two significant results for *report* and *test agent* when all participants were considered as a whole group. When groups were split between pass and fail only the first ones showed a significant results in the event *test agent*. This event is normally done as a control check before sending the fire team to attack. The significant results implies that the node distribution is very different at the time of this event in comparison with the rest of the session.

Finally, when directly comparing both groups (pass/fail) we found that workload is consistently different from these two groups where the poor performers group tend to have high workload. As a consequence, a good strategy to maximize the probability of success could be to minimize the workload in this type of participants. Similar patterns of workload are observed at events *air on* and *test agent* where once again, good performers have lower workload levels. Another interesting result seen in *request fire team* and *air on* events is that poor performers have high levels of distraction when compared to good performers. This is true at current time (DI0) were the largest mean difference is observed, but also before and after the event *air on* happens. The result makes sense as one would expect poor learner to be more distracted compared to successful participants. This finding suggests that in order to successfully finish the mission in the simulator a participant needs to have low levels of distraction and on the other hand, high levels of distraction could be an early warning for failures and an intervention could be triggered. The multivariate comparison between pass and fail further confirm the hypothesis for differences between these two groups especially around the events *attack* and *air on*.

In summary, the methodologies proposed in this work can be used to better understand the decision making process around events in a complex learning environment. The conclusions drawn by applying these tools can enable researchers and educators to improve the design of HCI and ITS by enhancing the user experience and improving learning.

Chapter 4

BAG OF AFFECTIVE STATES TO PREDICT PERFORMANCE EARLY IN AN EVENT-DRIVEN SIMULATOR

4.1 Introduction

Several efforts have been made in the last couple of years to develop mechanisms which allow computers to adapt and automatically tailor learning experience to a participant profile [50]. It is thought that students emotions need to be considered in order to motivate them and improve learning [13]. In spite of the challenge of identifying or even defining human emotions, the consensus is that if we are to create an effective Intelligent Tutoring System (ITS) it should be able to recognize feelings and mood in order to maximize learning experience. Unfortunately the current state of communication between machine and participants is asymmetric, this is, humans are able to obtain a lot of information from a computer such as operating system, memory, processor speed, etc. but the machine has very little information about the user[15]. The asymmetry is even greater when the participant is a person with motor impairments where conventional computer interfaces such as mouse, keyboard or game controllers cannot be used. Given this challenge several biometric sensors have been proposed to interact with these environments such as: heart rate, skin conductivity, electromyography (EMG), respiration, electroencephalogram (EGG) and electrocardiograms (ECG). Brain computer interfaces (BCIs) have the capacity to enable humans to control and interact with different learning environments from video games to military simulators. Valuable information regarding the cognitive state of the subject can be derived from BCIs and this information can be sent back to the computer with the goal of narrowing the communication gap between the user and the computer. The basic believe in the area of Affective Computing is that if a computer automatically recognizes and adapts to the user's emotions and affective states the quality of the interaction is enhanced and as a consequence the learning environment becomes more enjoyable and thus more effective [17]. However, detecting emotions is a hard task and many individual differences across participants is very challenging.

Performance prediction is a new area under research in ITS and EEG has been used to provide information about the ability of a user to successfully accomplish a task. In addition, EEG provides technological benefits to model different affective states that are task-independent and non-intrusive [9] which allows monitoring a subject's emotional state in real time and in quantitative way [30]. BCIs can also be used to accommodate individual differences in skills and emotional traits that are largely ignored in learning environments which have static difficulty levels [50]. In order to accomplish this goal performance has to be predicted ahead of time to avoid participants getting low engaged due to an easy difficulty level or overwhelmed and frustrated when the level doesn't match their skills. Furthermore, several studies have shown that traditional performance metrics are not enough to adapt learning environments but affective information should also be considered. ITS that detect and respond to user's affective states should be able to more accurate predict performance. For example, in [52] EEG was used to assess the impact of fatigue on a cognitive test and they were able to predict with high accuracy future performance. In another study, EEG was used to identify potentially impaired drivers in an unpredictable and dynamic driving simulator where researcher were able to group poor and good performers [68]. Performance can not only be predicted for individuals, in [57] complexity theory principles were used to derive physiologic models for teams on a submarine simulation using EEG. This is possible because is hypothesized that an individual's affective state is affected by other individuals during interactions. The number of applications in this new research can be applied to any field, from health-care, education and public safety. For example, in [51] a close-loop system was developed using EGG which is capable of detecting drowsiness and responds by sending alarms in a driving simulator. Advances in technology like wearables integrate hardware and software solutions that are able to respond and adapt to users in real time and the applications go from fashionable smartwatches to sensory eyewear. Physiological signals, in fact, can provide very valuable information about the user to understand his affective state.

In this chapter we address the challenge of predicting performance on a learning environment using a bag of states model where the final outcome is a binary classification problem (e.g. success/ fail). The intention of the methodology presented is to provide cognitive scientists with the statistical and machine learning tools to be able to design a feedback system which considers different user's profiles in order to increase engagement, provide enjoyment, stimulate attention while preventing failure. The methodology makes use of self-organizing maps (SOMs) to define different affective states. Once the SOM is trained with the physiological signal the number of activations is counted for each of the output nodes and fed into a machine learning model. We have named the methodology *Bag of Affective States (BAS)* because it resembles a bag-of-words model which is one of the most popular techniques for object categorization [28] and it has been widely adopted and successfully used in language processing as well as computer vision [69], [70], [71]. The novelty of our approach is that we are able to make predictions early on the learning session using only the available information at time t unlike other studies where prediction is a posteriori using all available information from the session and where participants are grouped given certain physiological information. The methodology is explained on an applied experiment where affective constructs provided by a research-grade EEG device is used as our main input on a damage control simulator.

The rest of the chapter is organized as follows. Section 4.2 explains the background of SOMs. Section 4.3 explains the methodology of *Bag of Affective States* while section 4.4 describes the simulation environment, participants and EEG recordings. Main results are detailed in section 4.5 and final conclusions are drawn in section 4.6.

4.2 Background

Self-organizing maps (SOM) are a type of artificial neural network that map a high dimensional input vector into a 2-dimensional grid. The grid is normally arranged on a $\sqrt{K} \times \sqrt{K}$ hexagonal layout but other arrangements could be used. In this case K is the total number of desired nodes and it is defined by the user. SOM belong to the unsupervised category of algorithms in machine learning because no labels are required for training. Once the SOM is trained each of the samples used for training are presented to the model and they are mapped to only one of the K output nodes by computing the Euclidean distance between the instance to be mapped and each of the output nodes. The node which provides the minimum distance with respect to the instance is chosen and we say that a node activation has occurred. The total number of activations should be equal to the number of instances (samples) in our training dataset. The SOM model created after training can be fully described by the node topology as well as the weights associated with each of the output nodes. Therefore, new or never seen instances can be presented to the SOM and they will be mapped to a single output node. This property can be very useful when designing ITS because we only train the model once. In a SOM all the input nodes are fully connected to the output nodes also called output layer. Each of these connections are denoted as w_{kq} where index q denotes an input node $q = 1, 2, \ldots, Q$ and k the output node $k = 1, 2, \ldots, K$. During the training process the weights of output nodes close to each other in the 2-dimensional grid are updated together. This process allows the SOM to preserve the topology of the original high dimensional space. In other words, instances close to each other in the high-dimensional space will be close in the 2-dimensional grid.

4.3 Bag of States Methodology

The first step in the methodology is to train a SOM using a physiological signal as the input. The high dimensional physiological signal mapped to a specific output node in the SOM can be seen as cluster where the centroid is the average of all the instances associated to that node. The centroid is described by the weights of a given node and can be represented as the vector $\boldsymbol{w_k} = [w_{k1}, w_{k2}, ..., w_{kQ}]^\intercal$. Therefore, each node can be interpreted as a meta affective state, this is, the combination of different affective measurements provided by physiological device. For example, a combination of affective states provided by an EEG headset such as engagement and workload can be seen as an affective state itself. For instance, a high level of engagement and high level of workload can be related to a subject that is learning a new task. On the other hand, low engagement, low distraction and low workload can be related to a person performing a well-known or automated task, such as driving a car (experienced driver).

Once the SOM is trained and we have all the nodes K and their respective weights $\boldsymbol{w_k}$ for $k = 1, 2, \ldots, K$, we can present the instances to the model and track their node activations. The physiological signal that is used as input is a time series where the time unit will depend on the sampling rate of the device or it can be defined by the experimenter. For now let's assume that our time unit is one second. Denote the number of activations for node k as y_k. Hence $\sum_{k=1}^{K} y_k = T$ where T

is the total number of seconds in the session. This process is done for each of the participants. Ideally, we would like to perform this procedure early into the session so we can predict final performance or in the case of an ITS to be able to adapt accordingly. Finally, another way to see the number of activations y_k is the amount of time spent by a participant on a node k or in other words, in an affective state $\boldsymbol{w_k}$. Finally, the input vector to a machine learning algorithm can now be constructed if we define $\mathbf{y} = [y_1, y_2, ..., y_K]$. In other words, we are using the distribution of node activations for each of the participants to make predictions about his performance where the class label can be defined as a binary classification problem (e.g. pass vs fail). The methodology can be extended to a multiclass classification problem as long as the machine learning algorithm used supports it. In the next section an applied experiment is presented where the *Bag of States* methodology is illustrated.

4.4 Experimental Protocol

4.4.1 Simulation Environment

The Damage Control Simulator (DCS) was developed by the CREEST lab from UCLA with the goal of improving Navy warfighter skills while participants respond to an on-board ship emergencies. The DCS was specifically designed for naval operations with the intention of developing a low-cost computer-based solution. The simulator is single player and is played on a third person view where each of the team members available has a specific skill such as: firefighter, electrician, technicians and a scene leader, all of them controlled by the player. The game adds realism by introducing random and unexpected situations while the player tries to put out a fire making the situation even more challenging. Participants have to make many diverse and complex decisions in order to successfully accomplish the mission like selecting the right equipment, deciding the best strategy, monitor personnel health and prioritizing tasks. The skills that are targeted to improve are: resource management, personnel safety, communications, tactical planning and adherence to Navy protocol. The action required during the game include: reporting, requesting teams, setting boundaries, investigating spaces, requesting mechanical/electrical isolation, checking equipment and attacking emergencies. The session is divided in three main parts: the first one consists in reporting the fire and assessing the initial situation, the second phase is about fighting the fire and the last part is monitoring the scene and making sure the fire is under control.

4.4.2 Participants

Sixty participants were recruited from the Arizona State University where 31 were female and 29 male. They were paid and had the option to leave the experiment at any time. Subjects were asked to participate in two sessions. The first one lasted 90 minutes and participants were required to fill out a demographic survey, they later took a training session followed by a pre-test. Participant were asked to review the tutorial embedded in the DCS and after that they played a damage control scenario in easy-mode. The goal of the first session was for the participants to get used to the user interface and also get familiar with the tasks and decisions in order to successfully finish the mission. Session 2 was on a different day with no more than 2 weeks between sessions. This time the participant wore the EEG headset and played the DCS in three levels with different difficulty: easy, moderate and hard. At the end of the session the participant was required to take a post-test.

4.4.3 EEG Recordings

For this experiment B-alert X series from Advanced Brain Monitoring (ABM) was used. The head set is a portable, easy-to-use EEG device which provides high quality recordings. It consists of 9 channels which are located in the mid-line and lateral sides of the skull (Fz, F3, F4, Cz, C3, C4, POz, P3, P4) plus an additional channel for ECG, EMG or EOG. The sampling rate is @256 Hz and the setup is done with gel which takes an average of 15 minutes. In order to check for signal quality the device performs an automated wireless impedance check and it allows for the signal to be transmitted up to 10 meters. The devices comes with a suite which allows providing the most common frequencies for EEG studies: delta, theta, alpha, beta, gamma and high gamma as well as 4 affective constructs: *high engagement, low engagement, distraction* and *drowsiness.* The suite also provides three measurements of memory workload: *workload average, workload forward digit span (WFDS)* and *workload backward digits span (WBDS).* The derivation as well as the validity of these measurements are explained in [9] and [10] while an application using this device in close-loop system can be found in [51].

4.5 Main Results

4.5.1 Results Using the Full Model

The input vector used to train our SOM consisted in the fours affective constructs provided by ABM B-Alert: *high engagement, low engagement, drowsiness* and *distraction.* These measurements are provided as the probability of a specific affective state being present at a given time in the subject and they sum up to one. We also used the three measurements of workload: *workload FDS, workload BDS* and *workload average* all of which were explained in subsection 4.4.3. Several topologies

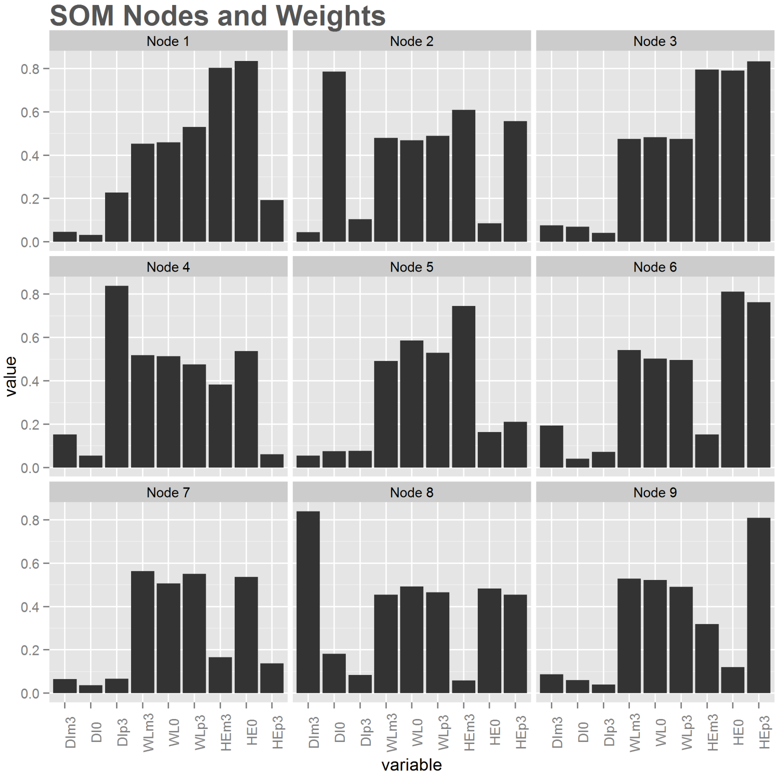


Figure 4.1: SOM Nodes and Weights After Training.

using different number of nodes and grids were studied where the SOM using a 3×3 hexagonal grid provided the best results. The nodes and weights after training can be seen in Fig. 4.1. Each node can be interpreted as an affective state itself. For example, node 7 is a combination of *high engagement* and low *workload* while node 3 is represented by *low engagement* and high *workload*.

ABM B-alert provides with the measurements per second, therefore a node activation occurs every second. From the 60 participants in total the first participant failed at second 59. Hence, we decided to use the node activations up to second 58. This way we ensure that all participants have the same number of activations and most importantly we are interested in using the node activations from the beginning of the session to predict the final outcome. In this sense, our features are generated

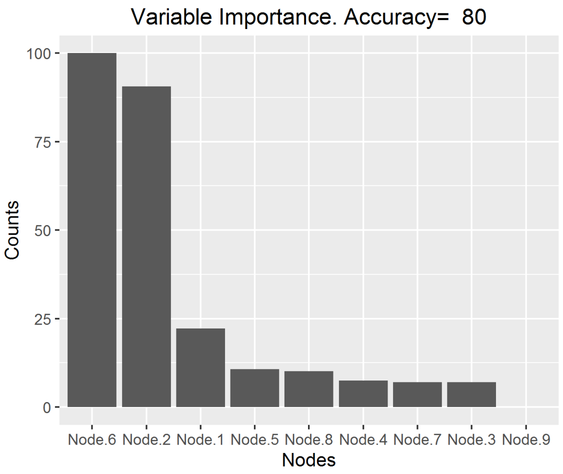


Figure 4.2: Variable Importance for Full Model.

Table 4.1: Confusion Matrix for Pass/Fail, Full Model

Predicted Actual	Pass	Fail
Pass	32	6
Fail	6	16

by counting the number of activations by node with a total of 58 per participant.

We used logistic regression to predict performance (pass/fail) were the input vector was constructed as explained in the methodology section 4.3 with a total of 9 nodes and where the total counts for each of these nodes can be seen as the total time spent on a specific affective state. The variable importance is shown in Fig. 4.2 where we see node 6 and 2 as the most important features. The 10-fold cross-validated confusion matrix is shown in Table 4.1 where 6 *pass* subjects were misclassified and 6 *fail* participants were also misclassified. Accuracy was 0.8, true negative rate 0.727 and true positive rate 0.842 where the positive class was *pass*.

Table 4.2 shows the performance of the *Bag of States* approach at different times.

Table 4.2: Bag of States Performance for Different Seconds.

Seconds	Accuracy	TPR	TNR
5	0.550	0.763	0.182
10	0.650	0.921	0.182
20	0.550	0.763	0.182
30	0.650	0.816	0.364
40	0.700	0.763	0.591
50	0.817	0.842	0.773
58	0.800	0.842	0.727
100	0.850	0.895	0.773

At $t = 5$ we observe that there is not discriminative power and the prediction has low accuracy. As time goes by the accuracy starts to increase and around $t = 50$ we have achieved an accuracy above 0.80 which remains constant up to $t = 58$ where we start losing participants due to mission failure. We went further and made predictions at $t = 100$ and the accuracy improved up to 0.85 but we also have to consider the fact that we have lost more participants and the prediction is unbalanced towards the successful class.

Fig. 4.3 shows a heatmap where the x axis represents the 9 nodes and the vertical axis shows all participant in two blocks. The upper part of the axis shows participants who passed and the lower part participants who failed. The horizontal line is a visual aid to separate them both. The color represents the counts of node activation for a given node k and a participant p. Furthermore, within each pass/fail block node 6 is sorted in descending order. Visually we see that the block of participants who passed tend to have fewer number of activations in node 6. On the other hand, the block of

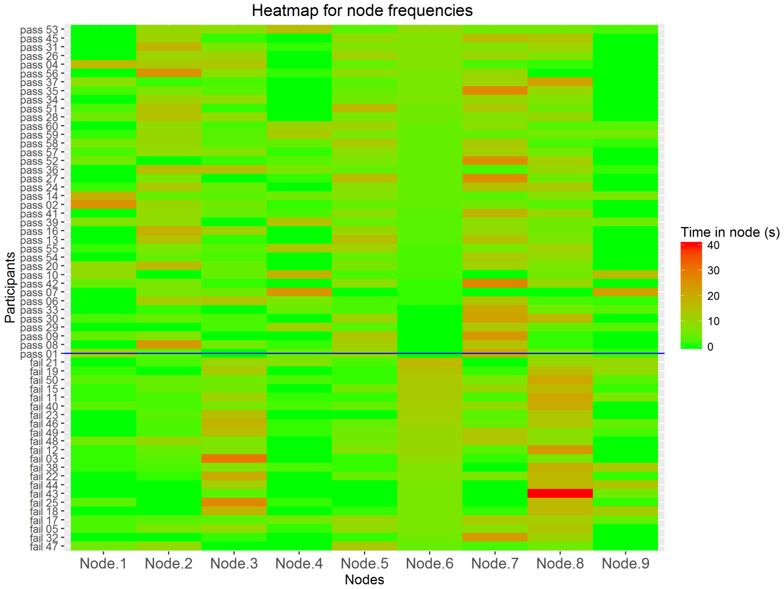


Figure 4.3: Heatmap for Node 6 Activations Counts.

participants who failed have a larger number of activations. Although, node 2 (the second most important feature) is not sorted we can also see a trend. In case of node 2 participants who passed have more activations than those who failed.

In order to further analyze the instances associated with node 6 and 2 boxplots were constructed. In Fig. 4.4 we see the central tendency as well as the dispersion for each of the affective states in node 6. *High engagement* and *low engagement* are close to 0.5 and there are no outliers (instances beyond the whiskers). *Distraction* is low with several outliers going from 0.15 to 0.37 approximately. A similar trend is observed in *drowsiness* although the mean is very close to zero. The 3 measurements of *workload* on average are close to 0.6 where the main box (from percentile 25% yo 75%) ranges from 0.5 to 0.75.

Fig. 4.5 shows the boxplots for node 2. In this case we observe instances for *high engagement* with values falling between 0 and 0.25. The opposite is true for *low engagement* where instances are on average around 0.85. *Distraction* and *drowsiness*

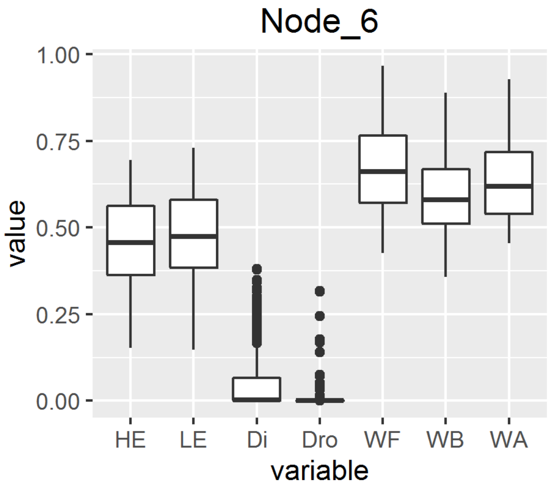


Figure 4.4: Box Plots for Instances Mapped to Node 6.

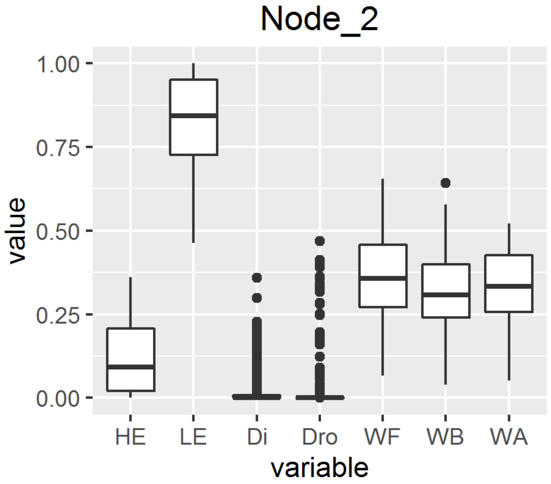


Figure 4.5: Box Plots for Instances Mapped to Node 2.

are again low and values for *workload* are on average below 0.50.

Plotting the number of activations for node 2 and 6 for each of the participants provides a good idea of why these nodes have good discriminative power. In Fig. 4.6 we see a scatter plot where the x axis represents counts for node 2 and the y axis counts for node 6 for each participant. Furthermore, participants are identified by pass/fail. An imaginary line with intercept in the y axis a slightly below 5 and with

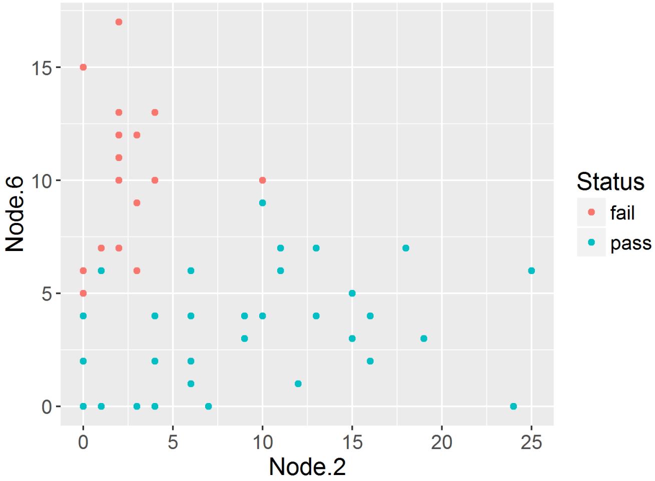


Figure 4.6: Scatter Plot of Activation Counts for Node 2 and 6 with Pass/Fail Color.

a slope around 1 could provide a good separation between these two groups.

Since node 6 appear to be the most important we decided to plot node 6 activations in time for all participants. Fig. 4.7 shows in green activations for node 6. The x axis is time in seconds and the y axis is arranged in four groups: the top one is represented by participants who passed but were misclassified. The next group are participants who passed and were correctly classified followed by participants who failed and were misclassified and finally the participants who failed that were correctly classified. Three horizontal lines separate these groups. Similarly to Fig. 4.3 we observe that participants that failed have more activations in node 6 when compared to those who passed. Further patterns are not completely clear. For example, participants who failed and were misclassified appear to have fewer activations at the beginning of the session. On the other hand, participants who passed and where correctly classified appear to have more activations at the beginning of the session. The same type of plot could be used to further analyze node 2.

We also compare node 6 against the rest of the nodes. In Fig. 4.8 we have 7 subplots, one for each affective construct provided by ABM B-alert. The x axis

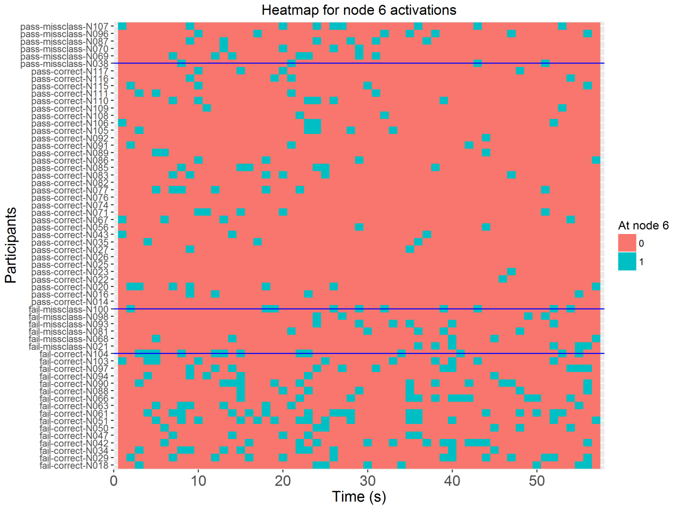


Figure 4.7: Node 6 Activations in Time by Participants.

represents the 9 nodes and the vertical axis represents the weights. In red we see node 6 and how it compares to the rest. For example, *high engagement* is very close to node 5 (around 0.5) but the rest of the nodes are either too low (1, 2, 3, 4 and 9) or too high (7, 8). Regarding workload, node 6 is similar to node 3, 8 and 9 where they show mid to high values.

The model was further reduced to include only the 3 most important features: counts for node 6, 2 and 1. Fig. 4.9 shows the variable importance where node 6 once again was the most important closely followed by node 2. Node 1 importance was zero. Consequently, we further reduced the model to only two variables: node 6 and 2. The 10-fold cross-validation accuracy was 0.833, the true positive rate was 0.868 and the true negative rate 0.773. The confusion matrix can be seen in Table 4.3 where each of the classes had 5 instances (participants) incorrectly classified.

In previous analyses we observed that *distraction* didn't show up among the most important variables and as a consequence it was probably not contributing too much to the performance prediction. Moreover, the three *workload* measurements seem to be highly correlated and they contain similar information. Therefore, we proceeded

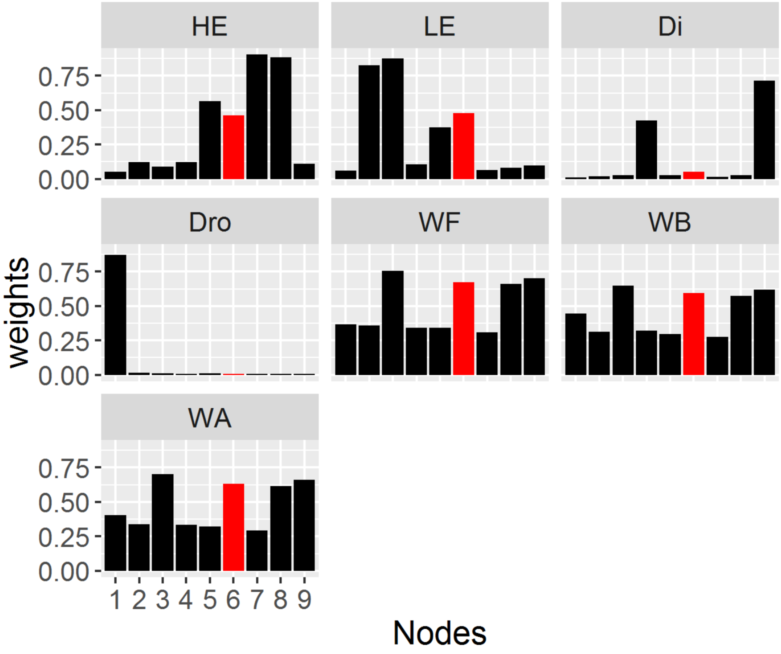


Figure 4.8: Node 6 (in Red) Compared to the Rest of the Nodes.

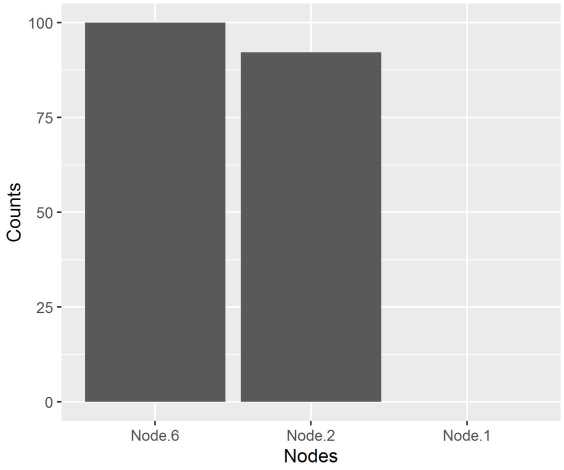


Figure 4.9: Variable Importance with the Model Reduced to 3 Predictors.

Table 4.3: Confusion Matrix for Pass/Fail, Reduced Model with 3 Features

Predicted Actual	Pass	Fail
Pass	33	5
Fail	5	17

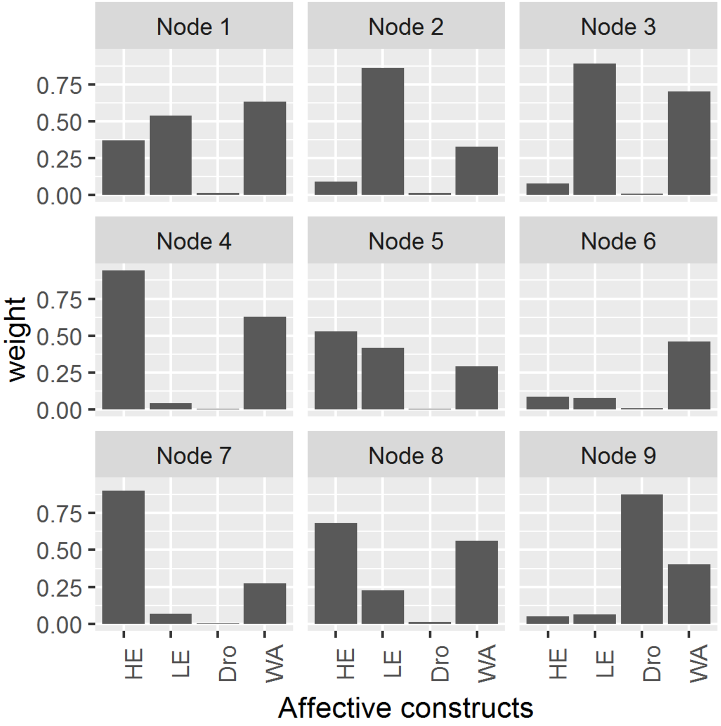


Figure 4.10: SOM Nodes and Weights after Dropping *Distraction, Workload FDS* and *Workload BDS*

to train the SOM this time dropping *distraction, workload FDS* and *workload BDS*. The new nodes and their weights can be seen in Fig. 4.10.

The 10-fold cross-validated results using logistic regression yield an accuracy of 0.85, a true positive rate of 0.895 and a true negative rate of 0.773. Once again, we computed the variable importance and the results are shown in 4.11. Node 2 was the most important closely followed by node 7 and 6. Node 2 is presented by having low values for *high engagement* and high values for *low engagement*. It also shows very low levels of *drowsiness* and moderate levels of *workload*. The model was further reduced using only the 3 most important nodes as inputs (*Node.2, Node.7* and *Node.6*) yielding the same performance.

4.5.2 Results Using a Subset of Physiological signals

Further efforts were made to reduce the number of affective states used to train the SOM. Being left with only these four affective constructs we proceeded to try the four possible combinations: *(HE, LE, Dro), (HE, LE, WA), (LE, Dro WA)* and *(HE, Dro, WA)* using a 3×3 and 2×2 topology. Nodes and weights for the 3×3 topology are shown in Fig. 4.12 where the four combinations are shown. Unfortunately, the performance was poor and the highest cross-validated accuracy for any given combination was no better than 63.3%. Nodes and weights for the 2×2 topology are shown in Fig. 4.13. Cross-validated accuracy for this topology was no better than 53% for any of the four combinations. Therefore, we conclude that given the affective contructs provided by ABM and for this specific learning environment *high engagement, low engagement, drowsiness* and *workload average* provide good information for early on performance prediction and this is the most parsimonious model we can get without sacrificing accuracy.

4.6 Conclusion

Emotions or affective states play an important role in human performance. The capacity of detecting and recognizing these affective states is an important aspect of human interaction[5]. Recent neurological studies show that in order for machines to be able to efficiently assist humans they should be able to recognize emotions [4].

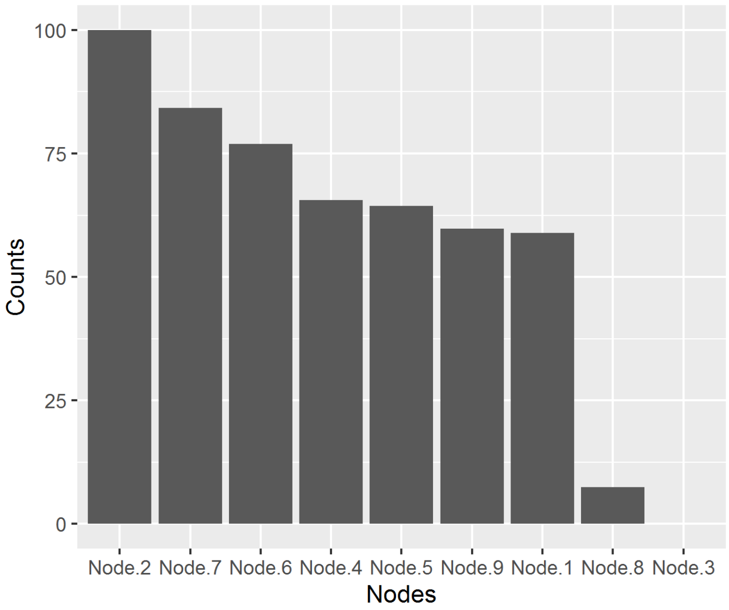
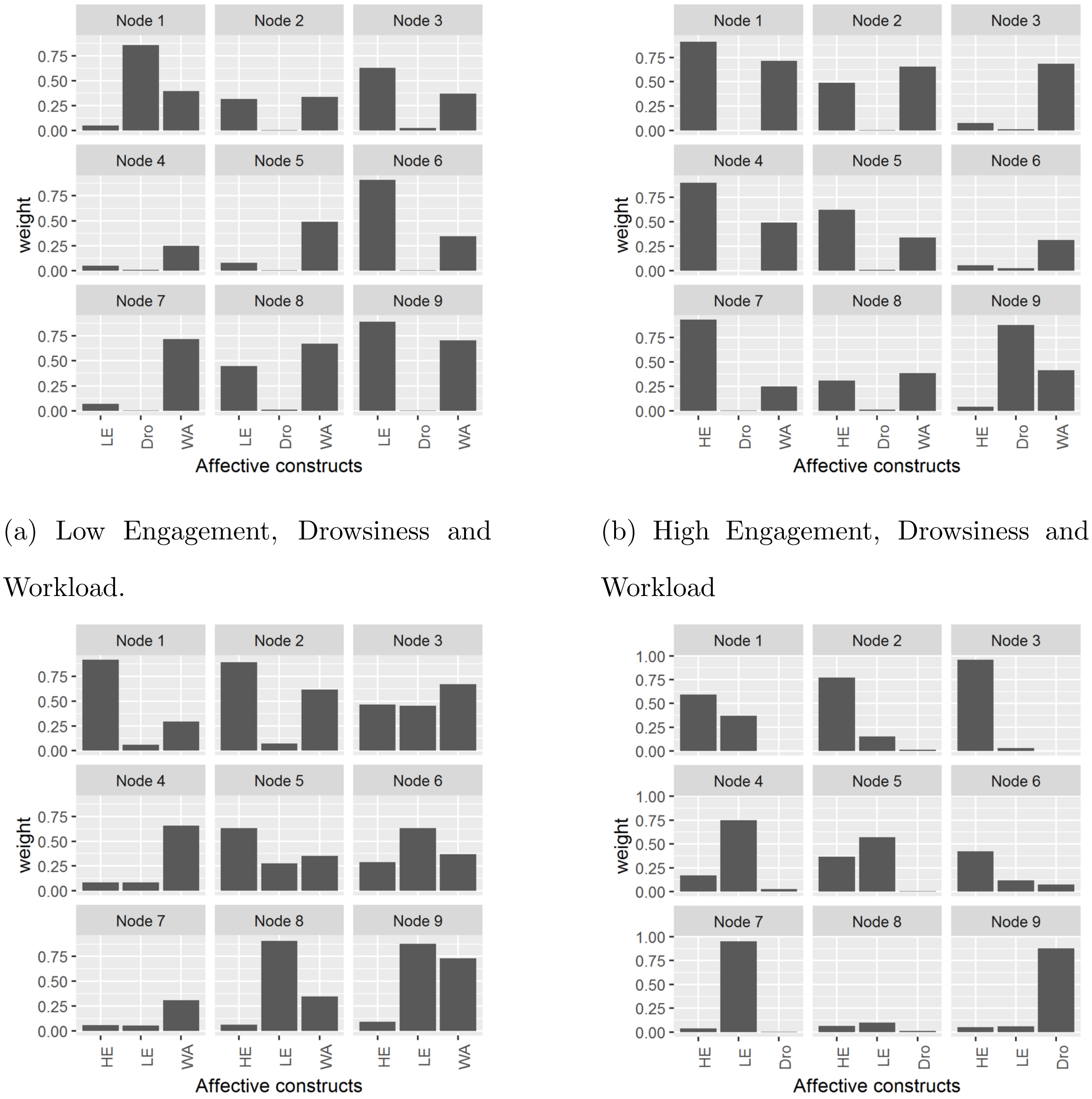


Figure 4.11: Variable Importance for the Reduced Model after Dropping *Distraction, Workload FDS* and *Workload BDS*

Moreover, physiological signals have been used to provide objective assessment in cognitive tasks prior to engaging in more complex learnings scenarios [72].

The aim of our work was to develop a methodology to take a physiological input and predict performance early on a learning environment. Performance prediction embedded on a ITS could be used to train participants to maximize learning while at the same time providing them an enjoyable experience [32]. The methodology considers training a SOM and monitor the number of activations for each of the output nodes at the beginning of the session. The counts of each of the activations can be seen as time spent on a certain affective state. The methodology was called *Bag of Affective States* because it resembles a bag-of-words approach widely used in machine learning. The meta affective states generated can be seen as a combination of different constructs provided by physiological devices which in the case of an EEG headset could inclide : *high engagement, low engagement, distraction* and *drowsiness* as well



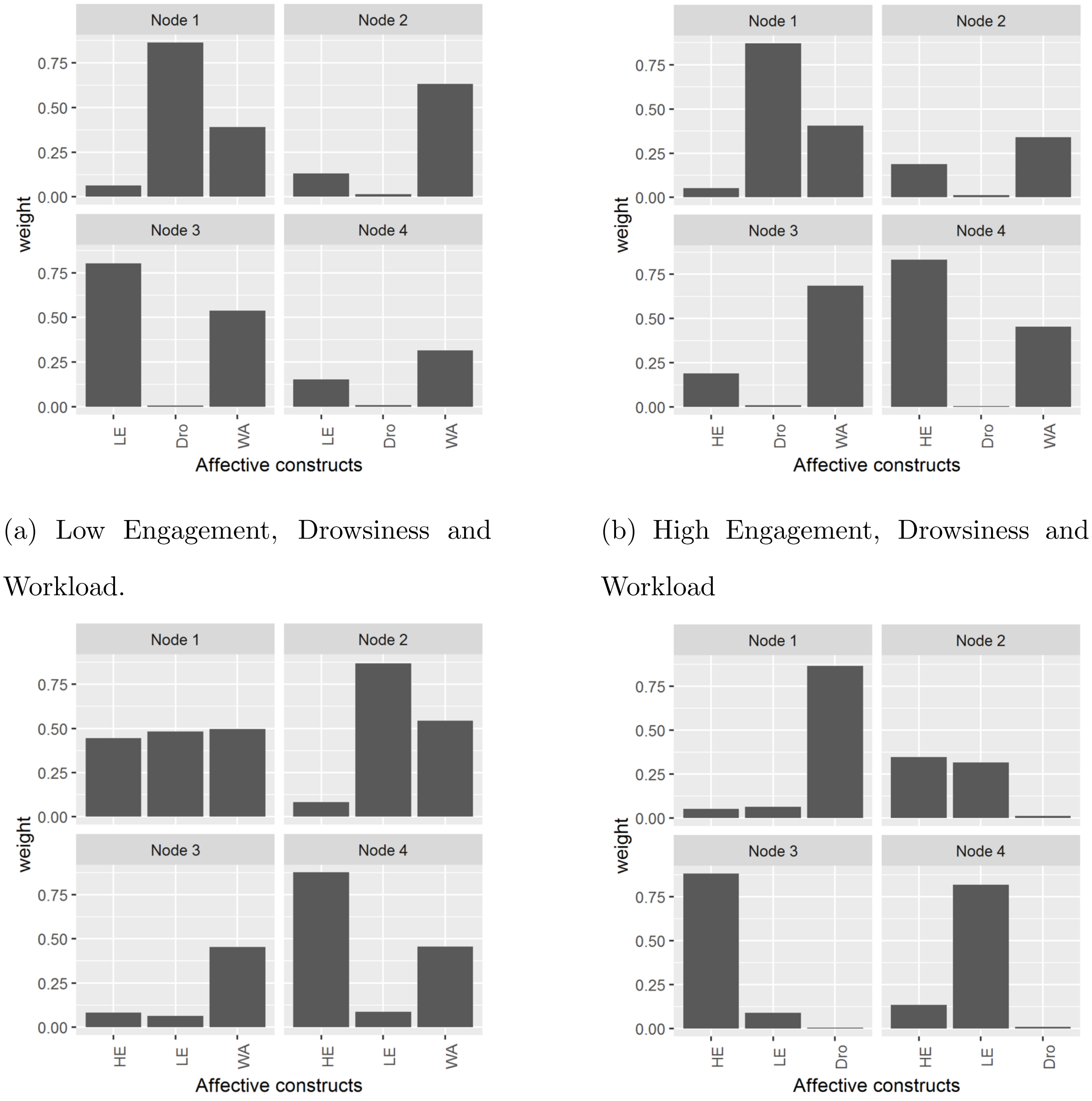
(a) Low Engagement, Drowsiness and Workload.

(b) High Engagement, Drowsiness and Workload

(c) High Engagement, Low Engagement and Workload

(d) High Engagement, Low Engagement and Drowsiness

Figure 4.12: Trained SOM Using Different Affective Constructs Combinations on a 3×3 Grid



(a) Low Engagement, Drowsiness and Workload.

(b) High Engagement, Drowsiness and Workload

(c) High Engagement, Low Engagement and Workload

(d) High Engagement, Low Engagement and Drowsiness

Figure 4.13: Trained SOM Using Different Affective Constructs Combinations on a 2×2 Grid

as measurements for memory workload. The novelty and utility of this methodology relies on the fact that performance prediction is done early on the simulation using only the available information at time $t << T$ where T is the duration of the session unlike other approaches were the information about the session is used in retrospective.

The methodology was applied on a Damage Control Simulator using affective constructs provided by a EEG headset and were cross-validated results showed an overall accuracy of 80% for a two class pass/fail prediction and where the true negative and true positive rates presented good balance. Node 6 and 2 were the most important features where the first one is characterized by moderate levels of *high engagement* and *low engagement*, low levels of *distraction* and *drowsiness* and moderate to high levels of *workload*. Participants who failed spent more time in this affective state when compared to participants who succeeded. On the other hand, participants who successfully completed the mission spent more time in node 2 which is described by low values of *high engagement* and high values of *low engagement*. This node has even lower values of *distraction* and *drowsiness* when compared to node 6. Node 2 also shows lower levels of *workload*.

Different time lengths at the beginning of the sessions were tried in order to identify how early we could predict performance. The first seconds of the session don't provide enough discriminative information and the machine learning algorithms performed poorly. However, as we consider more time in the analysis performance improves reaching its peak around second 50. The model was further reduced to considered only two nodes (6 and 2) and the performance was not degraded.

Looking for a more compact model the number of affective constructs used as inputs to train the SOM was reduced while at the same time preserving similar performance. We found that with only four affective constructs (*high engagement,* *low engagement, distraction* and *workload average*) similar cross-validated accuracy was achieved. Node 2 in the reduced model looks very similar to node 2 in the original model with low values for *high engagement* and high values for *low engagement* as well as very low levels of *drowsiness* and moderate levels of *workload*. This pattern was present more frequently in participants who passed. *High engagement* is generally associated with visual scanning, sustained attention and information-gathering while *workload* is associated with memory load and problem-solving activities [10]. In this manner, participants who passed showed lower *workload* and *engagement* which can reflect knowledge or expertise gained from session one where they learned how to use the simulator and where the mission protocol and goals were introduced. On the other hand, participants who failed spent more time in affective states with higher levels of *engagement* which may imply that they spent more time gathering information and scanning visually the screen which can be an indicator of the struggle to make sense of the data given certain scenario. *Engagement* and *workload* have shown to increase concordantly when the difficulty of a task is also increased [10] something that is even more prevalent in participants that find a task very challenging.

BCIs used in conjunction with ITS in simulation environments can maximize learning and detect areas of improvement as experiments can be conducted before sending users to dangerous tests [66]. Moreover, we showed that EEG cannot only be used to discriminate human cognitive activity like in [8] but it can also predict performance. Future research could include the study of individual differences in emotion-related cognitive tasks in other types of environments and see if a screening and categorization of participants is achievable. Different configurations could be implemented in ITS to accommodate specific needs based on a profile of a subgroup.

Another challenge to address for future research is the lack of temporal information of a bag of states approach. It is assumed that the sequence of events on learning environment contain important information about the user's cognitive state but that information is not considered in the bag of words. Sequential pattern mining techniques which consider temporal information on streams of data that are delivered in a sequence could be used in order to find relevant patterns with good prediction power. EEG raw signals can also be used instead of the affective constructs provided by a commercial, research or medical-grade headset. A large body of literature discusses the different frequencies, bandwidths and ratios of the sensors raw signal and they have been shown to be associated with response-inhibition, affective traits and attentional control [22]. The work presented here sets the foundation to continuously monitor affective states and use this information on a close-loop system which is able to adapt to users' emotions with the goal of improving performance in a variety of environments from education to military.

Chapter 5

BAYESNET FOR PERFORMANCE PREDICTION ON AN EVENT-DRIVEN SIMULATOR

5.1 Introduction

In order to provide learners with adequate assistance a precise estimation of student's future performance is a prerequisite [73]. A diverse number of factors can influence present and future student performance such as: educational background, family environment, teaching strategies as well as personal factors. In the field of education student performance has become increasingly important and even more when high-stakes tests are each year more and more critical for academic success [48]. For this reason a promising area for the application of affective computing is performance prediction which is not only restricted to the educational settings but also to more diverse fields from health care to military. For example, being able to predict when a soldier will succeed or fail on a given mission can provide some insights about the cognitive process of the most successful and least skillful participants and in the process this could probably save lives. A popular approach to model student's performance is Bayesian belief networks (BBNs) which use historical information as well as new evidence to model students' behavior and has been shown to help the decision-making process of educators regarding the strategies to enhance learning experience. As a matter of fact BBN modeling has become increasingly popular in distance education courses where a tutor keeps track of student progress and guides students according to their needs and abilities [74]. Educators use BBNs to model behavior because unlike other machine learning algorithms such as random trees and neural networks they provide a structure that is easy to interpret. BBNs are normally built using a combination of historical information and expert knowledge which also poses the challenge of continual adaptation of the model as new variations are introduced in the learning environment. Experts of a research field can easily manipulate the BBNs layout to include new information and constructs and in this way create more robust and accurate predictive models [75]. Recently, Intelligent Tutoring Systems (ITS) have incorporated embedded BBNs which provide real-time analysis of student performance in the form of posterior-probabilities. However, these ITS normally operate under high uncertainty regarding students' information and this leads to the BBN model to be incomplete and unable to capture all the student's interaction on a given time [76]. Another disadvantage of traditional BBN is that they are generally modeled to do long-term assessment and prediction of student's action so they are less reliable at the beginning of the tutoring session where there is little evidence available [77]. A mayor challenge is to be able to predict the final outcome of the beginning or middle part of the session since it is useless from the practical point of view to predict performance once we know the final outcome. Moreover, if the final objective is to build a close-loop system this should be able to adapt before is too late or the participant has already failed. Fig. 5.1 shows a time series of a learning environment session and we have divided the time in three parts: beginning, middle game and end game. Ideally we would like to extract features from the first and second part because the last part is already too late and we probably already know the final outcome.

The methodology in this chapter presents a way of using a BBN and its latent variables temporal information as inputs for a logistic regression model in order to make predictions about performance. The utility of this approach is that we consider time into the analysis and we focus on how early we can predict the final outcome. Traditionally, BBNs need a lot of evidence in order to be reliable and they are based

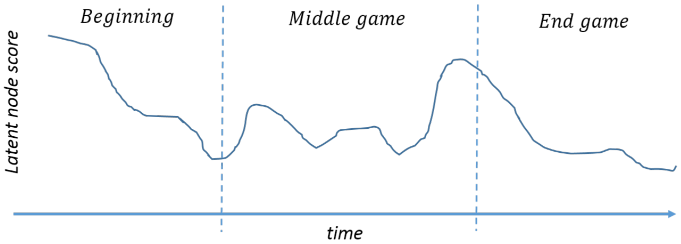


Figure 5.1: Time Series Representation of a Latent Node Score in a BBN.

partly on expert knowledge which sometimes could be biased. Furthermore, the latent nodes of BNNs are normally associated with specific skills and psychometricians generally analyze them in a unidimensional manner [29] not considering that students skills are most of the time highly correlated. In the present investigation we proposed the use of temporal information provided by a BNNs to predict performance early on a learning environment. The chapter is divided as follows: in section 5.2 the theory about BBNs is reviewed and in section 5.3 the methodology used in this contribution is explained. In section 5.4 the experimental protocol including the simulation environment, participants and the BayesNet is introduced. The applied example is illustrated in section 5.5 and the results are described in section 5.6. Finally conclusions are drawn in section 5.7.

5.2 Bayes Belief Network Background

A very popular machine learning technique used in intelligent tutors is the Bayesian belief networks [35]. BBNs are used to model student knowledge in a learning environment and allow making predictions considering past information. The BBN is a graphical representation of probabilistic relationships between different predictors. The network is composed of two main elements: 1) an acyclic graph showing the relation between the predictors (nodes) and 2) the probabilities associated with each of these nodes [56].

BBNs are very useful when we only have statistical dependencies among different variables. These casual dependencies can be defined by a domain expert which combined with prior belief can weigh new observable data. Computers have made easier to compute prior and conditional probabilities to make inferences in real time. BBNs are normally built using historical information and expert knowledge. One of the most popular application is in ITS where historical data and information from experts can help better assess knowledge or skill acquisition of students based on a set of observable tests or tasks [35]. In a BBN we are trying to infer a posterior probability after observing some data or collecting evidence. This posterior probability is derived from the Bayes' theorem which is given in the formula 5.1.

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)} \tag{5.1}$$

In Figure 5.2 we can see a simple BBN in which we have 5 nodes: A, B, C, D and E. Each of the nodes has a probability associated to it. These probabilities are obtained from historical information, running participants, expert opinion or a mixture of some of these approaches and they are stored in so called probability tables. In the same figure we see for example that nodes A and B don't have precedent nodes and therefore only prior probabilities are stored in the probability tables. On the other hand, for nodes C, D and E we have conditional probabilities because these nodes depend on the state of their respective parent nodes.

In a ITS setting node A could denote the acquisition of skill A while node C could be the score of a test which we are able to observe. Once we observe C then we can make inferences about what is the probability of the participant having skill A given than we observed C, expressed in mathematical notation as $P(A|C)$. Each of these nodes are represented by discrete states but with continuous-valued associated probabilities [78]. For example the node A can have two states: the participant has

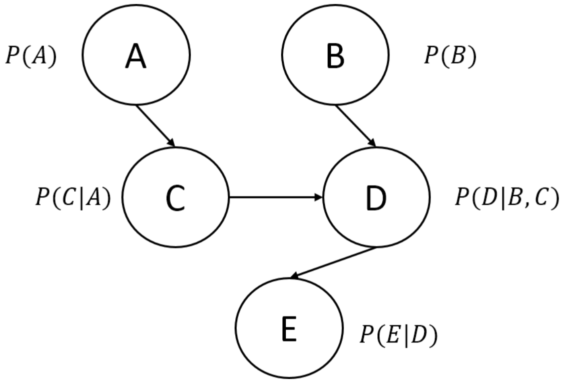


Figure 5.2: A Simple Bayes Belief Network with 5 Nodes.

the skill A or the participant does not have the skill A. Sometimes to reduce clutter notation these states can be simply represented as *yes* and *no*. The observable nodes also have a discrete and finite representation of states. In our current example we mentioned that C represented a score of a test which would normally be continuous. However, we can discretized a continuous score by defining ranges such as: 1-3 to represent low, 4-7 could be moderate and 8-10 representing a high score, assuming that the score scale goes from 1 to 10. Therefore, it is always recommended to explicitly define the states when computing the posterior probabilities. For example, the probability of participant having skill A given that he scored a 6 in test C should be defined as $P(A = yes|C = moderate)$.

5.3 BayesNet for Performance Prediction Methodology

The methodology presented here allows to use posterior probabilities of latent variables considering time information to make performance predictions. In order to accomplish this, let's define $y_{k,t}$ as the posterior probability of latent node k at time t where $k = 1, 2, ..., K$ and $t = 1, 2, ..., T$. This posterior probability can statistically be defined as $y_{k,t} = P($ Skill k = good | all evidence available at time $t)$. In other words, we will try to estimate the probability of a skill or ability k to be acquired (it's good) at time t using all the information provided by the observable nodes which contribute to the computation of this probability. In this case, all the evidence is accumulative which means that all events which happen before time t are considered. Furthermore, we are assuming that the posterior probabilities are updated every time new evidence is presented. For example, if no new evidence is presented between time t and $t{+}1$ then $y_{k,t} = y_{k,t+1}$. Normally in BBNs used in learning environments the observable nodes are presented as scores and these have to be discretized as explained in the background section. The next step is to define the feature vector which can be represented as $\mathbf{y_t} = [y_{k,t}, y_{k,t+1}, ..., y_{K,t}]$. This feature vector is the input which will be used in a machine learning algorithm to make predictions at time t. Finally, depending on the machine learning algorithm used different variable importance metrics could be used. For example, in random forest the most common approaches are Gini importance which considers the mean Gini gain produced by the variable $y_{k,t}$ over all trees and the permutation importance which tracks the decrease in classification accuracy after permuting $y_{k,t}$ over all trees. In the case of logistic regression variable importance is normally computed using the absolute value of the t-statistic for each latent variables $y_{k,t}$. The next section will illustrate this methodology with an applied experiment.

5.4 Experimental Protocol

5.4.1 Simulation Environment

The Damage Control simulator (DCS) was developed in UCLA by the National Center for Research on Evaluation, Standards and Student Training (CREEST). The simulator provides with realistic shipboard emergencies in which multiple fires can occur simultaneously. The player is required to respond to different threats by putting together a team of people with different skills. The primary goal is to put out a fire where the player has to make a series of complex decisions based on previous training. The different characters to choose from the simulator include: technicians, fire fighters, electricians, investigators, scene leaders, etc. The players can employ diverse tactics and they respond to a scenario that is not deterministic but introduces random and unexpected changes like equipment malfunctions which forces the player to change his strategy and adapt. Instead of focusing on very specific skills such as how to select the appropriate equipment based on the type of fire or how to fix a mechanical failure the DCS tries to improve higher order skills which are critical for a marine. The system has a Bayesian network which provides real time assessment of the situation which is updated by gathering information from observable actions such as: reporting, selecting equipment, checking agents, performing mechanical or electrical isolation and deployment of specific crews. The latent skills that the system is designed to target are: adherence to Navy protocol, communication, tactical planning, personnel safety, resource management and situational awareness. The BayesNet embedded in the simulator was designed using input from participants as well as from expert Navy instructors. This ensures that the expected skills to be acquired match the expectation of a human instructor.

5.4.2 Participants

Sixty-nine participants (37 female and 32 male) were recruited from Arizona State University. The students were paid for their participation and they have the option to leave at any time during the study. Students participated in two sessions with no more than 2 weeks in between. In the first session participants were trained to put out a fire according to the Navy protocol developed for the simulator. Participants also took a tutorial which is embedded in the simulator in order to get familiar with the user interface. In the second session participants were asked to play three different levels of difficulty: easy, moderate and hard. The difficulty was modified by changing different parameters such as: fire speed, smoke growth, fire damage, agent modifier, smoke emissions, etc. For this contribution we selected the outcome of hard session because it provided a good balance between the number of successful and unsuccessful results (42 passed, 27 failed). For the easy and moderate sessions the number of participants who passed was very high making the dataset highly unbalanced and unfit for this experiment.

5.5 BayesNet Applied Experiment

The UCLA Damage Control Simulator has an embedded BBN which is constantly being updated as new evidence is presented to the model. Figure 5.3 shows the architecture of the BBN which was designed based on expert knowledge. At the top we can see the node called *Damage Control Management* which summarizes the casualty proficiency of two different areas: fire and flood casualty management. In a lower level we have 6 latent nodes described as: *communications*, *compartment integrity*, *personal safety*, *casualty management*, *situation awareness/decision making* and *checksheet adherence*. At the very bottom of the BBN we see the observable nodes: report, set boundaries, set zebra, mechanical isolation, test agent, request reliefs, among many others. All the latent and observable nodes have two states: good performance and bad performance. In order to define the current state on an observable node a score or rubric is generated according to some evidence collected in real time during the simulation. These scores are later discretized into two states: good or bad. Once the observable nodes are updated the effect is propagated to all the BBN when the posterior probabilities are computed. An output file is provided by the DCS which computes these probabilities in the following mathematical notation: $P(Communication = Good|Report = Good)$ in this case the probability of the skill *communication* being good given that the observable action *report* was done correctly. Every time new evidence is presented to the BBN new and updated posterior probabilities are computed considering previous information. Following our example $P(Communication = Good|Report = Good, Mechanical - Isolation = Bad)$, in other words what is the probability of the skill *communication* being good given that the observable action *report* was good and the *mechanical isolation* was bad. Each participant has an output file where each row represents a posterior probability which is generated every time new evidence is presented and the columns are the latent variables. Therefore, the first row contains only one evidence: *report fire* the first task a participant is supposed to perform and it is the same for all participants, while the last row contains the posterior probabilities considering all evidence presented throughout the simulation. The question we are trying to address is: How early can we predict success or failure given the information provided in real-time by the BBN? In order to answer this question we followed a simple approach: at each point in time (seconds) we will use the available information provided by the DCS BBN in the form of posterior probabilities and only using the 5 latent nodes: *communications* (comm), *compartment integrity* (comp-int), *personal safety* (per-safety), *casualty management* (casualty) and *situation awareness/decision making* (sit-awa).

5.6 Main Results

We performed a logistic regression using the inputs described in the previous section for a two classification model (pass/fail) at each second starting from the beginning of the session up to when the first participant failed (second 58). Figure 5.4 shows the progression of the error. The model was built using the *caret* package in R using the ”glm” method. The error was computed using 10-fold cross-validation. We observed that there is a trend starting at time 34 and the model achieves its lowest

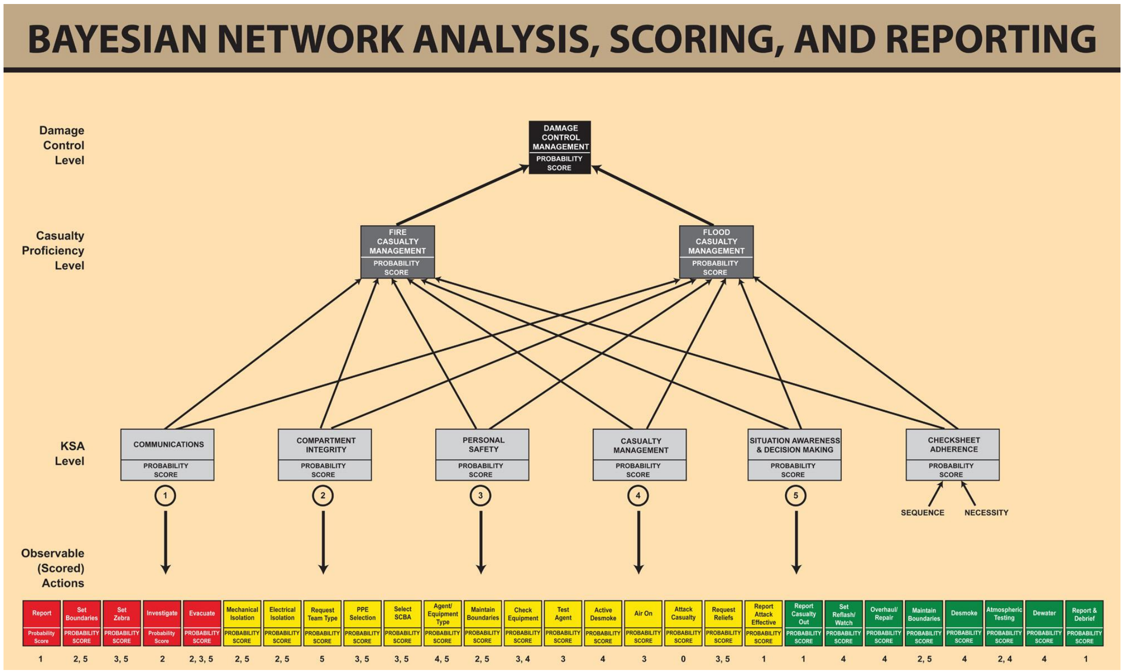

BAYESIAN NETWORK ANALYSIS, SCORING, AND REPORTING
Damage Control Level
Casualty Proficiency Level
KSA Level
Observable (Scored) Actions
DAMAGE CONTROL MANAGEMENT PROBABILITY SCORE
FIRE CASUALTY MANAGEMENT PROBABILITY SCORE
FLOOD CASUALTY MANAGEMENT PROBABILITY SCORE
COMMUNICATIONS PROBABILITY SCORE
COMPARTMENT INTEGRITY PROBABILITY SCORE
PERSONAL SAFETY PROBABILITY SCORE
CASUALTY MANAGEMENT PROBABILITY SCORE
SITUATION AWARENESS & DECISION MAKING PROBABILITY SCORE
CHECKSHEET ADHERENCE PROBABILITY SCORE
1
2
3
4
5
SEQUENCE
NECESSITY
Report Probability Score 1
Set Boundaries PROBABILITY SCORE 2, 5
Set Zebra PROBABILITY SCORE 3, 5
Investigate Probability Score 2
Evacuate PROBABILITY SCORE 2, 3, 5
Mechanical Isolation PROBABILITY SCORE 2, 5
Electrical Isolation PROBABILITY SCORE 2, 5
Request Team Type PROBABILITY SCORE 5
PPE Selection PROBABILITY SCORE 3, 5
Select SCBA PROBABILITY SCORE 3, 5
Agent/ Equipment Type PROBABILITY SCORE 4, 5
Maintain Boundaries PROBABILITY SCORE 2, 5
Check Equipment PROBABILITY SCORE 3, 4
Test Agent PROBABILITY SCORE 3
Active Desmoke PROBABILITY SCORE 4
Air On PROBABILITY SCORE 3
Attack Casualty PROBABILITY SCORE 0
Request Reliefs PROBABILITY SCORE 3, 5
Report Attack Effective PROBABILITY SCORE 1
Report Casualty Out PROBABILITY SCORE 1
Set Reflash/ Watch PROBABILITY SCORE 4
Overhaul/ Repair PROBABILITY SCORE 4
Maintain Boundaries PROBABILITY SCORE 2, 5
Desmoke PROBABILITY SCORE 4
Atmospheric Testing PROBABILITY SCORE 2, 4
Dewater PROBABILITY SCORE 4
Report & Debrief PROBABILITY SCORE 1

Figure 5.3: Bayes Belief Network Provided by the Damage Control Simulator Developed by UCLA-CRESST.

error rate around second 38. The results for the model at time 38 are shown in Table 5.1. The 10-fold cross-validated accuracy was 0.826, the true positive rate 0.833 and the true negative rate 0.815 with *pass* being the positive class.

We performed variable importance where *personal safety* turned out to be the most important latent node followed by *situational awareness* and *communication*. The latent nodes *casualty management* as well as *compartment integrity* seem to have very low importance. Results are shown in Fig. 5.5

In Fig. 5.4 we observe that the lowest error rate is achieved around second 38. However, the error rate starts to increase once again. In order to explain this behavior

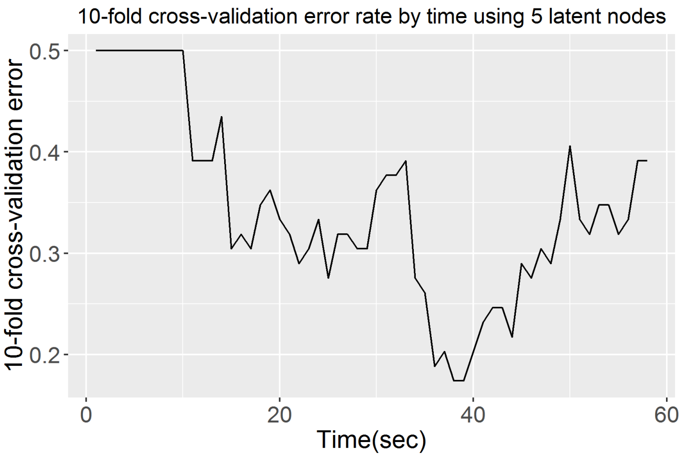


Figure 5.4: 10-fold Cross-Validated Error Using Logistic Regression.

Table 5.1: Confusion Matrix for Pass/Fail at Second 38.

Predicted Actual	Pass	Fail
Pass	35	7
Fail	7	20

and given that *personal safety* was the most important variable we created a heatmap shown in Fig. 5.6. In this map the x axis is time in seconds starting at second 1 up to 58 (first participant failes at second 59). The vertical axis represents the participants and they are further divided into two blocks: the upper block are the participants who passed and the lower block those who failed. We added a horizontal line to separate these two groups and a vertical line to point the time where we get the lowest error rate. The color of the graph represents the score provided as a probability. The lighter the color the closer to 1 and the darkest the closer to zero. We see that at the beginning of the session all participants start with the same probability which is close to zero. As the participant progresses in the simulator by performing several tasks new evidence is presented to the observable nodes and the probabilities are back propagated to

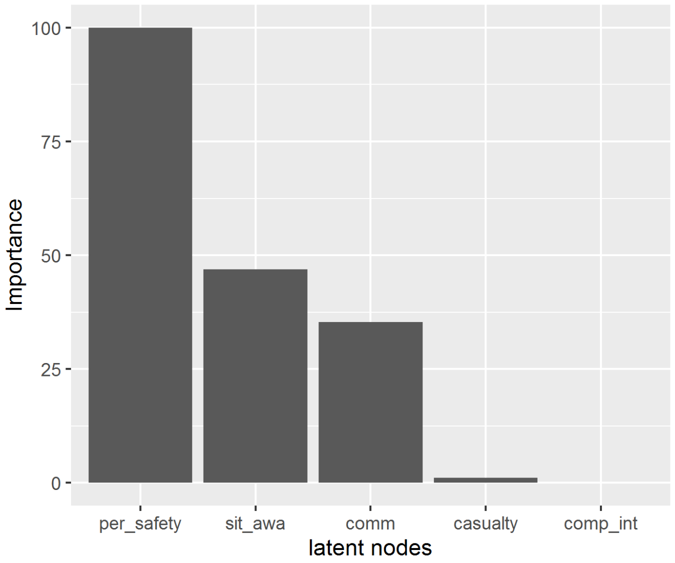


Figure 5.5: Logistic Regression Variable Importance at Second 38.

the latent nodes. Participants in the heatmap were further sorted within each group (pass/fail) by the average score on *personal safety*. The top participants in the *pass* group seem to increase the score very early while the bottom of the plot (participants who failed) do it very late in the session. The vertical line at second 38 provides a good visual aid to demonstrate why we are able to discriminate between *pass* and *fail* around that time. After second 38 those participants who failed catch up again and they are able to increase the score by performing different tasks and once again we lose the discriminative power after second 50. This is the reason of the behavior of the error rate shown in Fig. 5.4. The main take away here is that: we are able to discriminate between *pass* and *fail* based on the latent score of *personal safety* where successful participants are able to increase the score very quickly (before 38). On the other hand, participants who failed tend to be slower and although they are able to increase the score by performing the same tasks as the successful participants they do it late (after second 38). Another way to see this pattern is shown in Fig. 5.7. In this plot we see that the mean of *personal safety* score for those participants who passed

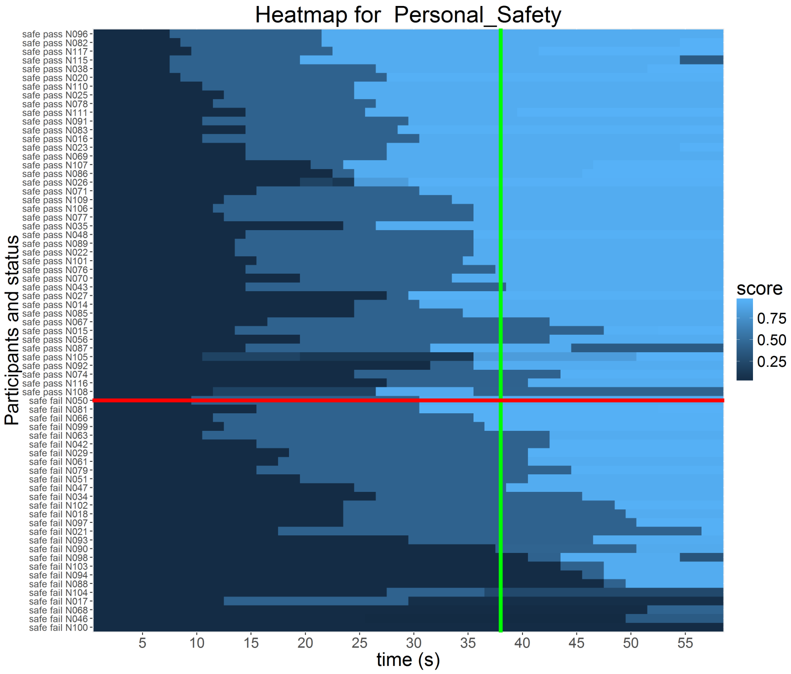


Figure 5.6: Heatmap of the *Personal Safety* Score.

increases sooner than those who failed. Furthermore, not only it increases quicker but it stays above all the way until second 58. We also plot the standard deviation where participants who passed have smaller variation.

The analysis presented above showed that *personal safety* in time was a good predictor of performance. Instead of relying on a single second to make predictions we decided to create features out of this latent node. Fig. 5.8 shows how 10-seconds averages were generated as the new features to predict performance. There were a total of 5 segments.

Once again we used a 10-fold cross-validation to compute performance as well as variable importance. Segment 4 which comprises second 31 to 40 turned out to be the most important followed by segment 5 in a far second place. This confirms our previous analysis where we found that around second 38 the lowest error rate was achieved.

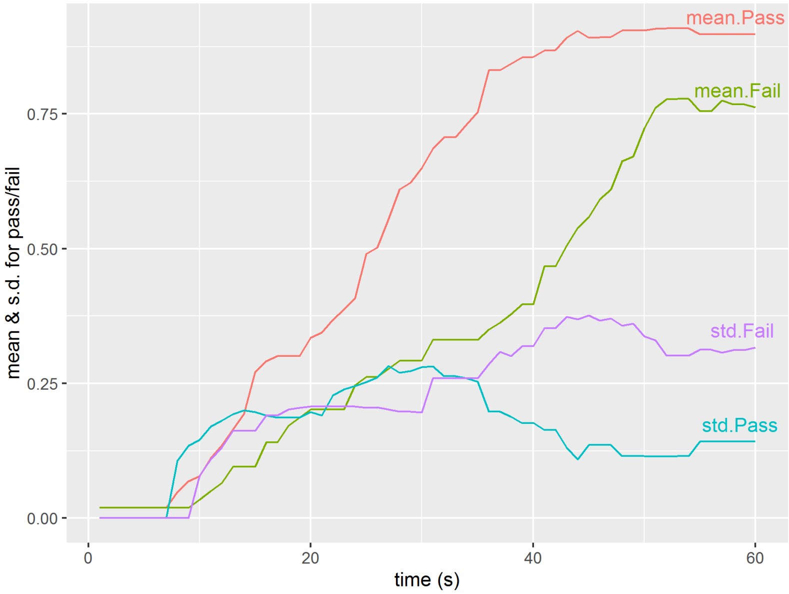


Figure 5.7: Time Series of Mean and Standard Deviation for Pass/Fail on *Personal Safety* Score.

Table 5.2: Confusion Matrix for Pass/Fail Using 10-Seconds Averages as Features.

Predicted / Actual	Pass	Fail
Pass	35	7
Fail	7	20

The cross-validated accuracy was 0.797 with a true positive rate of 0.833 and a true negative rate of 0.741. The confusion matrix is shown in table 5.2 where 7 participants were misclassified for each of the classes.

5.7 Conclusion

Setting up rules with fixed cut offs to predict performance is a challenging task. A lot of research has been devoted to analyze different personal, cultural and social

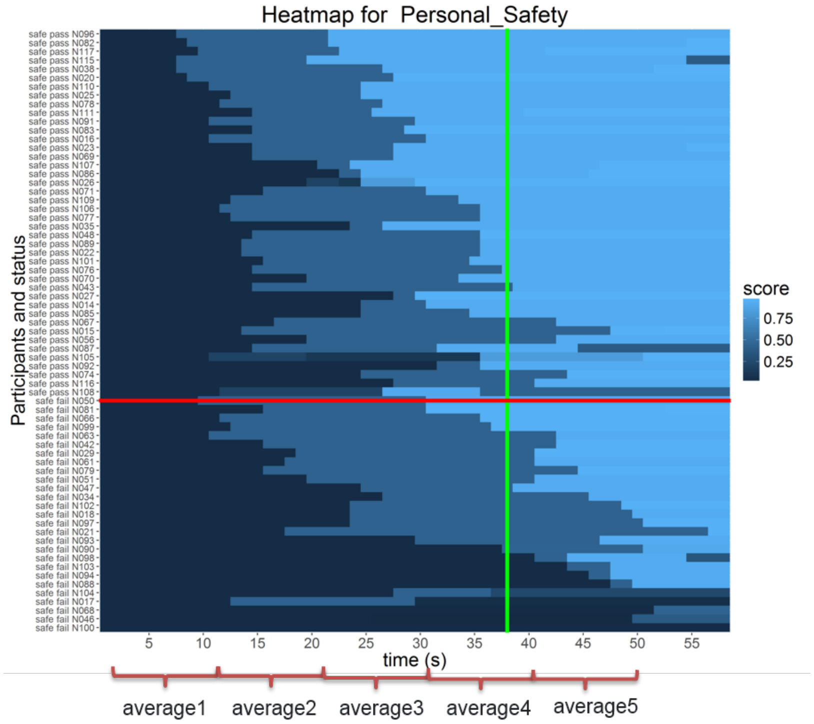


Figure 5.8: Features Created on *Personal Safety* Taking 10-seconds Averages.

information in order to come up with features to make this prediction possible with mixed results [73]. In this contribution we presented a methodology which uses the posterior probabilities of latent nodes in a BBN to predict participants' performance on a learning environment at a given time t. Moreover, this methodology tries to overcome the fact that BBNs are difficult to update and they are normally structured for a very specific goal. If the conditions change or new challenges are introduced the BBN structure and conditional probabilities become outdated [77]. The approach presented here can be used to take advantage of the information which is already embedded in most ITS and use it to predict different outcomes. Performance prediction in the form of a two class classifier for pass/fail was the objective of the applied experiment but other types of goals can be defined. For example, we could be interested in predicting which participants tend to have deficiencies following emergency protocols

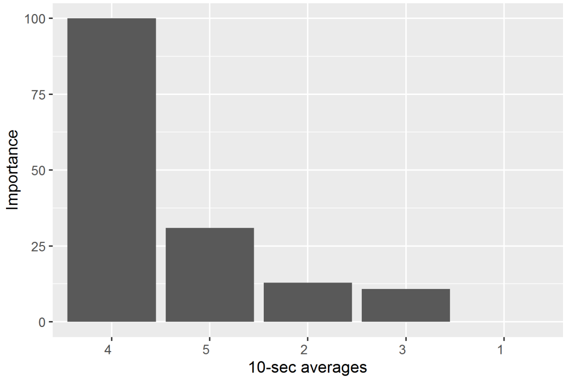


Figure 5.9: Variable Importance for 10-seconds Averages.

or selecting the wrong equipment or what type of subjects are more likely to reach a given objective without monitoring staff health. Most importantly, in our approach we have considered the temporal information provided by the sequence of tasks that provide evidence to update the latent nodes something that is normally neglected in Bayesian modeling.

An applied experiment was performed to illustrate the applicability of this approach were the latent scores of a BayesNet were used as inputs in a logistic regression to predict participants' performance in a Damage Control Simulator. The posterior probabilities were updated in real-time as new evidence was presented. *Personal safety* turned out to be the most important predictor. The observable nodes directly related to *personal safety* according to the BBN are the following tasks: set zebra, evacuate, PPE selection, select SCBA, check equipment, test agent, air on and requests reliefs. This implies that how fast and precise these tasks are performed in the initial seconds into the simulation provide valuable information about the final outcome.

Our current approach assumes that all participants have the same background and prior knowledge so the conditional and prior probabilities are the same. It would be interesting to consider experience in order to better individualize and personalize predictions for different group of subjects. Moreover, different models can be built to consider clusters of students with different set of skills. This approach could be fruitful because finding individualized priors can be expensive [79] and it can also lead to overfitting. Creating an effective BBN is not without challenges, this is because building the structure which encodes the conditional dependence across nodes is not trivial and different configurations can be generated. Therefore, consensus among researchers about the ”ideal” BBN for a specific task is sometimes difficult to reach [75].

Knowledge gained in this contribution can be used to better design ITS by using BBN information already embedded in the learning environment. Furthermore, by using latent nodes scores as inputs to a machine learning algorithm we open the door to explore potential interactions between different latent skills challenging the ”unidimensional” approach that traditionally pychometricians embrace. Ignoring the fact that the majority of skills for a given objective are highly correlated can lead to miss important interactions [29]. Finally, one of the key elements of any ITS is to interpret learners decision in order to enable a model of student learning and reasoning [80]. We believe that this work leads the way into that direction.

Chapter 6

CONCLUSION AND FUTURE WORK

6.1 Conclusions

This dissertation proposes a new set of analytical methods for high dimensional physiological sensors which are applicable to numerous problems in learning science and also in industrial settings where high dimensional signals are present.

The first contribution proposes the event-crossover (ECO) to analyze performance on any learning environment. The ECO is most appropriate for studies where the main objective is to evaluate performance on a learning environment where events are embedded over time with simultaneous sensors or physiological responses being recorded in real-time. The main advantage of our method is that each subject acts as its own control. Therefore, the methodology allows us to avoid the traditional necessity of controlling for other confounded effects such as age, gender, health, skill level, etc. The effectiveness of this methodology was illustrated on an applied experiment where participants played two songs from the video game Guitar Hero with different levels of difficulty. In this contribution the ECO was able to identify the affective constructs of long term excitement, short term excitement and frustration as significant in the hard-expert combination for all type of cases. This implies that the affective state value of the participant for these emotional states is different when the player is making errors (events) than when the player is not making errors (control). A similar result was observed for the easy-expert combination in the > 0 case. Learning scientist now have a new tool that can allow them analyze and understand the cognitive state of participants near specific events.

Contribution 2 introduced analytical methods to study the relationship between a multi-dimensional physiological signal and sentinel events that occur randomly on a learning environment. In the first approach we proposed a multivariate version of the event-crossover where instead of analyzing the different physiological signals independently we use all the information of the input vector distribution near time of events using multivariate methods to draw conclusions. In the second approach we represented different physiological patterns in the form of weight combinations using self-organizing maps (SOM) and analyze correlated proportions of node activations near time of events using different statistical techniques. We proposed a discrete version of the ECO methodology where the high dimensional feature space was lowered to a two-dimensional space. Once the SOM was trained the feature vector was presented to the model and the node activations were tracked. The McNemar's test for participants who passed showed that the output node 1, which is characterized as a combination of very low levels of *distraction*, moderate levels of *workload* and high levels of *engagement* was significantly different from a random chosen interval for the event *report*. Further exploring the node distribution it was found that this pattern is absent during this event. The Generalized McNemar's or Stuart-Maxwell test showed two significant results for *report* and *test agent* when all participants were considered as a whole group.

In the last methodology proposed in this contribution we compared for differences in the physiological signals between two groups at the time of specific events using univariate and multivariate methods. In the multivariate approach it was found that when directly comparing both groups (pass/fail) workload was consistently different for poor performers, where this group were more inclined to have high workload.

The methodologies proposed in this contribution can be used to better understand the decision making process around events in a complex learning environment. The conclusions drawn by applying these tools can enable researchers and educators to improve the design of HCI and ITS by enhancing the user experience and improving learning.

In contribution 3 a methodology was proposed with the goal of extracting features that later could be used on a machine learning model to make performance predictions. The methodology was designed to take a high dimensional physiological signal to train a self-organizing map (SOM) and derive meta affective states which can be seen as a combination of different affective states. The methodology keeps track of the time spent in each meta affective state and this information is later used on a machine learning algorithm to make predictions about performance. The novelty and utility of this methodology relies on the fact that performance prediction is done early on the simulation using only the available information at time $t << T$ where T is the duration of the session unlike other approaches were the information about the session is used in retrospective. The methodology was called *Bag of Affective States* because it resembles a bag-of-words widely used in text and image processing.

The methodology was applied to a damage control simulator where participants required to perform several complex tasks with the objective of putting out a fire on a submarine. Cross-validated results showed an overall accuracy of 80% for a two class pass/fail prediction and where the true negative and true positive rates presented good balance. The first seconds of the session didn't provide enough discriminative information and the machine learning algorithms performed poorly. However, as we consider more time in the analysis performance improved. The model was further reduced to consider only two nodes (6 and 2) and the performance was not degraded. Looking for a more compact model the number of affective constructs used as inputs to train the SOM was reduced while at the same time preserving similar performance. We found that with only four affective constructs (*high engagement, low* *engagement, drowsiness* and *workload average*) similar cross-validated accuracy was achieved. Findings suggest that participants who succeeded the mission were more likely to spent time in an affective state formed by a combination of low levels of *engagement* and *drowsiness* as well as low to moderate levels of workload in contrast with participants who failed who showed lower levels of engagement and higher levels of workload.

In contribution 4 we proposed a methodology to use evidence-driven updates to Bayesian belief networks (BBNs) to predict performance early on considering temporal information. Scores of the latent nodes were used as inputs to a machine learning algorithm in real-time as the observable nodes were updated with new evidence. Furthermore, with this approach it is possible to identify those latent variables which have more discriminative power by means of variable importance. In order to demonstrate the methodology a Bayes belief network (BBN) generated from a training simulator was used as input using information of the first seconds into the session with the objective of predicting participants' performance. Significant results were found early in the learning session which implies that how fast and precise the observable tasks are performed in the initial seconds into the simulation provide valuable information about the final outcome. Moreover, this methodology tried to overcome the fact that BBNs are difficult to update and they are normally structured for a very specific goal. The approach presented here can be used to take advantage of the information which is already embedded in most ITS and use it to predict different outcomes. Most importantly, in our approach we have considered the temporal information provided by the sequence of tasks that provide evidence to update the latent nodes something that is normally neglected in Bayesian modeling.

However, several disadvantages exist with this approach in comparison to the EEG performance prediction from contribution 3. First of all, BBNs are difficult to model because they required expert knowledge and historical information from many participants in order to be reliable. Individualized priors can be expensive [79] and it can also lead to overfitting. Second, the structure which encodes the conditional dependence across nodes is not trivial and different configurations can be generated. Therefore, consensus among researchers about the ”ideal” BBN for a specific task is sometimes difficult to reach [75]. Third, BBNs on ITS normally operate under high uncertainty regarding students' information and this leads to the BBN model to be incomplete and unable to capture all the student's interaction on a given time [76]. Another disadvantage of traditional BBN is that they are generally modeled to do long-term assessment so they are less reliable at the beginning of the tutoring session where there is little evidence available [77]. On the other hand, in terms of portability, reliability and costs EEG turns out to be a very practical tool to use to model affective states which we showed are good predictors of learners' performance. The signals captured by EEG can be associated with different brain processes and they are detected by the synchronization and desynchronization of neurons in specific parts of the brain [20] which makes it suitable to model different affective states as well as measurement of cognitive processes. Finally, we showed that in terms of prediction accuracy EEG compares pretty well with respect to BBNs.

6.2 Future Work

The natural research path to follow up on this work is to extend these methodologies utilizing other physiological devices besides EEG such as: eye-trackers, electrocardiograms (ECG o EKG), electromyogram (EMG) and electrodermal activity (EDA) just to mention a few. The methodologies presented in this work allow combining different types of physiological signals to form real meta affective states and they all together can provide a richer interpretation of the user's cognitive state.

In the second contribution we explored the use of SOMs in order to lower a high dimensional signal into a 2-dimensional grid. Other approaches such as deep autoenconders could be used to lower the dimensionality similarly to that of the PCA with the advantage that deep autoenconders can capture nonlinear relations unlike PCA. Deep autoencoders achieve this by training a multilayer neural network with a hidden layer that is restricted to a few nodes where the goal is to reconstruct the multidimensional input vector. [81]

The bag-of-states approach introduced in the third contribution neglected the order of the affective states. It is logical to think that the sequence of changes between affective states could have important information about a user's cognitive state and therefore about his performance. The different physiological signals could be discretized to form a word representation of the affective state and then word embeddings techniques such as word2vec [82] could be produced using neural networks. This technique not only computes the distribution of events but it is also able to capture information regarding the sequence of events or words and may turn out to be useful for learning scientists looking to understand emotional changes in time.

Future research could also include the study of individual differences in emotion-related cognitive tasks in other types of environments and see if a screening and categorization of participants is achievable. Different configurations could be implemented in ITS to accommodate specific needs based on a profile of a subgroup. EEG raw signals can also be used instead of the affective constructs provided by a commercial, research or medical-grade headset. A large body of literature discusses the different frequencies, bandwidths and ratios of the sensors raw signal that have been shown to be associated with response-inhibition, affective traits and attentional control [22].

In the last contribution using the BayesNet, information about the sequence of events was indirectly considered because the posterior probabilities take into account all previous information. However, the order of the events is not considered. Sequential pattern mining could be applied in this case to come up with a series of rules with a given minimum support that later could be used as inputs for a prediction algorithms.

Finally, one of the issues left out for future research is that of considering student's previous knowledge to be input to the BBN prior to the beginning of the session. Our current approach assumes that all participants have the same background and prior knowledge so the conditional and prior probabilities are the same. It would be interesting to consider experience in order to better individualize and personalize predictions for different group of subjects. Moreover, different models can be built to consider clusters of students with different set of skills. This approach could be beneficial because finding individualized priors can be expensive [79] and it can also lead to overfitting.

REFERENCES

[1] 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.

[2] G. N. Yannakakis and J. Togelius, "Experience-driven procedural content generation," *IEEE Transactions on Affective Computing*, vol. 2, no. 3, pp. 147–161, 2011.

[3] M. A. Nicolaou, H. Gunes, and M. Pantic, "Continuous prediction of spontaneous affect from multiple cues and modalities in valence-arousal space," *IEEE Transactions on Affective Computing*, vol. 2, no. 2, pp. 92–105, 2011.

[4] R. W. Picard, "Affective computing mit media laboratory perceptual computing section technical report no. 321," 1995.

[5] X.-W. Wang, D. Nie, and B.-L. Lu, "Emotional state classification from eeg data using machine learning approach," *Neurocomputing*, vol. 129, pp. 94–106, 2014.

[6] L. A. Schmidt and L. J. Trainor, "Frontal brain electrical activity (eeg) distinguishes valence and intensity of musical emotions," *Cognition & Emotion*, vol. 15, no. 4, pp. 487–500, 2001.

[7] M. Salminen and N. Ravaja, "Increased oscillatory theta activation evoked by violent digital game events," *Neuroscience letters*, vol. 435, no. 1, pp. 69–72, 2008.

[8] G. F. Wilson and F. Fisher, "Cognitive task classification based upon topographic eeg data," *Biological Psychology*, vol. 40, no. 1, pp. 239–250, 1995.

[9] R. R. Johnson, D. P. Popovic, R. E. Olmstead, M. Stikic, D. J. Levendowski, and C. Berka, "Drowsiness/alertness algorithm development and validation using synchronized eeg and cognitive performance to individualize a generalized model," *Biological psychology*, vol. 87, no. 2, pp. 241–250, 2011.

[10] C. Berka, D. J. Levendowski, M. N. Lumicao, A. Yau, G. Davis, V. T. Zivkovic, R. E. Olmstead, P. D. Tremoulet, and P. L. Craven, "Eeg correlates of task engagement and mental workload in vigilance, learning, and memory tasks," *Aviation, space, and environmental medicine*, vol. 78, no. Supplement 1, pp. B231–B244, 2007.

[11] E. W. Anderson, K. C. Potter, L. E. Matzen, J. F. Shepherd, G. A. Preston, and C. T. Silva, "A user study of visualization effectiveness using eeg and cognitive load," in *Computer Graphics Forum*, vol. 30, no. 3. Wiley Online Library, 2011, pp. 791–800.

[12] L. Shen, M. Wang, and R. Shen, “Affective e-learning: Using” emotional” data to improve learning in pervasive learning environment.” *Educational Technology & Society*, vol. 12, no. 2, pp. 176–189, 2009.

[13] X. Li, B. Hu, T. Zhu, J. Yan, and F. Zheng, “Towards affective learning with an eeg feedback approach,” in *Proceedings of the first ACM international workshop on Multimedia technologies for distance learning.* ACM, 2009, pp. 33–38.

[14] M.-K. Kim, M. Kim, E. Oh, and S.-P. Kim, “A review on the computational methods for emotional state estimation from the human eeg,” *Computational and mathematical methods in medicine*, vol. 2013, 2013.

[15] S. H. Fairclough, “Fundamentals of physiological computing,” *Interacting with computers*, vol. 21, no. 1, pp. 133–145, 2009.

[16] J. B. Van Erp, F. Lotte, and M. Tangermann, “Brain-computer interfaces: beyond medical applications,” *Computer*, no. 4, pp. 26–34, 2012.

[17] R. A. Calvo and S. D’Mello, “Affect detection: An interdisciplinary review of models, methods, and their applications,” *Affective Computing, IEEE Transactions on*, vol. 1, no. 1, pp. 18–37, 2010.

[18] J. Allanson and S. H. Fairclough, “A research agenda for physiological computing,” *Interacting with computers*, vol. 16, no. 5, pp. 857–878, 2004.

[19] N. Chumerin, N. V. Manyakov, M. van Vliet, A. Robben, A. Combaz, and M. Van Hulle, “Steady-state visual evoked potential-based computer gaming on a consumer-grade eeg device,” *Computational Intelligence and AI in Games, IEEE Transactions on*, vol. 5, no. 2, pp. 100–110, 2013.

[20] S. K. Hadjidimitriou and L. J. Hadjileontiadis, “Eeg-based classification of music appraisal responses using time-frequency analysis and familiarity ratings,” *Affective Computing, IEEE Transactions on*, vol. 4, no. 2, pp. 161–172, 2013.

[21] W. Klimesch, “Eeg alpha and theta oscillations reflect cognitive and memory performance: a review and analysis,” *Brain research reviews*, vol. 29, no. 2, pp. 169–195, 1999.

[22] P. Putman, J. van Peer, I. Maimari, and S. van der Werff, “Eeg theta/beta ratio in relation to fear-modulated response-inhibition, attentional control, and affective traits,” *Biological psychology*, vol. 83, no. 2, pp. 73–78, 2010.

[23] M. X. Cohen, *Analyzing neural time series data: theory and practice.* MIT Press, 2014.

[24] N.-Y. Liang, P. Saratchandran, G.-B. Huang, and N. Sundararajan, “Classification of mental tasks from eeg signals using extreme learning machine,” *International journal of neural systems*, vol. 16, no. 01, pp. 29–38, 2006.

[25] M. Murugappan, M. Rizon, R. Nagarajan, S. Yaacob, I. Zunaidi, and D. Hazry, "Eeg feature extraction for classifying emotions using fcm and fkm," *International journal of Computers and Communications*, vol. 1, no. 2, pp. 21–25, 2007.

[26] O. Jensen and C. D. Tesche, "Frontal theta activity in humans increases with memory load in a working memory task," *European journal of Neuroscience*, vol. 15, no. 8, pp. 1395–1399, 2002.

[27] P. Zarjam, J. Epps, and F. Chen, "Characterizing working memory load using eeg delta activity," in *Signal Processing Conference, 2011 19th European*. IEEE, 2011, pp. 1554–1558.

[28] Y. Zhang, R. Jin, and Z.-H. Zhou, "Understanding bag-of-words model: a statistical framework," *International Journal of Machine Learning and Cybernetics*, vol. 1, no. 1-4, pp. 43–52, 2010.

[29] Z. A. Pardos, N. T. Heffernan, B. Anderson, C. L. Heffernan, and W. P. Schools, "Using fine-grained skill models to fit student performance with bayesian networks," *Handbook of educational data mining*, vol. 417, 2010.

[30] R. Berta, F. Bellotti, A. De Gloria, D. Pranantha, and C. Schatten, "Electroencephalogram and physiological signal analysis for assessing flow in games," *Computational Intelligence and AI in Games, IEEE Transactions on*, vol. 5, no. 2, pp. 164–175, 2013.

[31] P. C. Petrantonakis and L. J. Hadjileontiadis, "Emotion recognition from brain signals using hybrid adaptive filtering and higher order crossings analysis," *Affective Computing, IEEE Transactions on*, vol. 1, no. 2, pp. 81–97, 2010.

[32] L. Derbali, P. Chalfoun, and C. Frasson, "Assessment of learners attention while overcoming errors and obstacles: An empirical study," in *Artificial Intelligence in Education*. Springer, 2011, pp. 39–46.

[33] G. A. Lujan-Moreno, R. Atkinson, G. Runger, J. Gonzalez-Sanchez, and M. E. Chavez-Echeagaray, "Classification of video game players using eeg and logistic regression with ridge estimator," *xxx*, vol. 29, no. 2, pp. 169–195, 2014.

[34] R. Scherer, G. Moitzi, I. Daly, and G. R. Muller-Putz, "On the use of games for noninvasive eeg-based functional brain mapping," *Computational Intelligence and AI in Games, IEEE Transactions on*, vol. 5, no. 2, pp. 155–163, 2013.

[35] B. P. Woolf, *Building intelligent interactive tutors: Student-centered strategies for revolutionizing e-learning*. Morgan Kaufmann, 2010.

[36] R. S. Baker and K. Yacef, "The state of educational data mining in 2009: A review and future visions," *JEDM-Journal of Educational Data Mining*, vol. 1, no. 1, pp. 3–17, 2009.

[37] J. M. Kivikangas and N. Ravaja, "Emotional responses to victory and defeat as a function of opponent," *Affective Computing, IEEE Transactions on*, vol. 4, no. 2, pp. 173–182, 2013.

[38] D. Kerr, “Using data mining results to improve educational video game design,” *JEDM-Journal of Educational Data Mining*, vol. 7, no. 3, pp. 1–17, 2015.

[39] M. Maclure, Mittleman, and MA, “Should we use a case-crossover design?” *Annual review of public health*, vol. 21, no. 1, pp. 193–221, 2000.

[40] D. W. Hosmer Jr, S. Lemeshow, and R. X. Sturdivant, *Applied logistic regression.* John Wiley & Sons, 2013, vol. 398.

[41] 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, p. e38, 2013.

[42] B. Goldberg, K. W. Brawner, and H. K. Holden, “Efficacy of measuring engagement during computer-based training with low-cost electroencephalogram (eeg) sensor outputs,” in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, vol. 56, no. 1. SAGE Publications, 2012, pp. 198–202.

[43] M. Duvinage, T. Castermans, M. Petieau, T. Hoellinger, G. Cheron, and T. Dutoit, “Performance of the emotiv epoc headset for p300-based applications,” *Biomedical engineering online*, vol. 12, no. 1, p. 56, 2013.

[44] I. Arroyo, K. Ferguson, J. Johns, T. Dragon, H. Meheranian, D. Fisher, A. Barto, S. Mahadevan, and B. P. Woolf, “Repairing disengagement with non-invasive interventions,” in *AIED*, vol. 2007, 2007, pp. 195–202.

[45] V. A. Aleven and K. R. Koedinger, “An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor,” *Cognitive science*, vol. 26, no. 2, pp. 147–179, 2002.

[46] L. D. Miller, L.-K. Soh, A. Samal, K. Kupzyk, and G. Nugent, “A comparison of educational statistics and data mining approaches to identify characteristics that impact online learning,” *JEDM-Journal of Educational Data Mining*, vol. 7, no. 3, pp. 117–150, 2015.

[47] M. Helander, T. Landauer, and P. Prabhu, “Ecological information systems and support of learning: coupling work domain information to user characteristics,” *Handbook of human-computer interaction*, p. 315, 1997.

[48] C. R. Beal, R. Walles, I. Arroyo, and B. P. Woolf, “On-line tutoring for math achievement testing: A controlled evaluation,” *Journal of Interactive Online Learning*, vol. 6, no. 1, pp. 43–55, 2007.

[49] P. Rani, C. Liu, N. Sarkar, and E. Vanman, “An empirical study of machine learning techniques for affect recognition in human–robot interaction,” *Pattern Analysis and Applications*, vol. 9, no. 1, pp. 58–69, 2006.

[50] C. Liu, P. Agrawal, N. Sarkar, and S. Chen, “Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback,” *International Journal of Human-Computer Interaction*, vol. 25, no. 6, pp. 506–529, 2009.

[51] C. Berka, D. Levendowski, P. Westbrook, G. Davis, M. N. Lumicao, C. Ramsey, M. M. Petrovic, V. T. Zivkovic, and R. E. Olmstead, "Implementation of a closed-loop real-time eeg-based drowsiness detection system: Effects of feedback alarms on performance in a driving simulator," in *1st International Conference on Augmented Cognition, Las Vegas, NV*, 2005, pp. 151–170.

[52] M. Stikic, C. Berka, D. J. Levendowski, R. F. Rubio, V. Tan, S. Korszen, D. Barba, and D. Wurzer, "Modeling temporal sequences of cognitive state changes based on a combination of eeg-engagement, eeg-workload, and heart rate metrics," *Using Neurophysiological Signals that Reflect Cognitive or Affective State*, p. 242, 2015.

[53] G. Lujan-Moreno, R. Atkison, and G. Runger, "Affective state assessment in a learning environment with physiological signals and event-crossover analytics," *Journal of Educational Data Mining*, 2016.

[54] T. Hastie, R. Tibshirani, J. Friedman, and J. Franklin, "The elements of statistical learning: data mining, inference and prediction," *The Mathematical Intelligencer*, vol. 27, no. 2, pp. 83–85, 2005.

[55] T. Kohonen, "The self-organizing map," *Proceedings of the IEEE*, vol. 78, no. 9, pp. 1464–1480, 1990.

[56] P.-N. Tan, M. Steinbach, V. Kumar *et al.*, *Introduction to data mining.* Pearson Addison Wesley Boston, 2006, vol. 1.

[57] R. Stevens, T. Galloway, P. Wang, C. Berka, V. Tan, T. Wohlgemuth, J. Lamb, and R. Buckles, "Modeling the neurodynamic complexity of submarine navigation teams," *Computational and mathematical organization theory*, vol. 19, no. 3, pp. 346–369, 2013.

[58] M. Kantardzic, *Data mining: concepts, models, methods, and algorithms.* John Wiley & Sons, 2011.

[59] E. Cutrell and D. Tan, "Bci for passive input in hci," in *Proceedings of CHI*, vol. 8. Citeseer, 2008, pp. 1–3.

[60] Q. McNemar, "Note on the sampling error of the difference between correlated proportions or percentages," *Psychometrika*, vol. 12, no. 2, pp. 153–157, 1947.

[61] A. Stuart, "A test for homogeneity of the marginal distributions in a two-way classification," *Biometrika*, vol. 42, no. 3/4, pp. 412–416, 1955.

[62] A. E. Maxwell, "Comparing the classification of subjects by two independent judges," *The British Journal of Psychiatry*, vol. 116, no. 535, pp. 651–655, 1970.

[63] I. Arroyo, C. Beal, T. Murray, R. Walles, and B. P. Woolf, "Web-based intelligent multimedia tutoring for high stakes achievement tests," in *Intelligent Tutoring Systems.* Springer, 2004, pp. 468–477.

[64] G. Rupert Jr *et al.*, *Simultaneous statistical inference.* Springer Science & Business Media, 2012.

[65] J. P. Shaffer, "Multiple hypothesis testing," *Annual review of psychology*, vol. 46, p. 561, 1995.

[66] D. Marshall, D. Coyle, S. Wilson, and M. Callaghan, "Games, gameplay, and BCI: the state of the art," *Computational Intelligence and AI in Games, IEEE Transactions on*, vol. 5, no. 2, pp. 82–99, 2013.

[67] M. G. Helander, *Handbook of human-computer interaction.* Elsevier, 2014.

[68] T. D. Marcotte, R. A. Meyer, T. Hendrix, and R. Johnson, "The relationship between realtime eeg engagement, distraction and workload estimates and simulator-based driving performance," in *Proceedings of the Seventh International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design, Bolton Landing, NY*, 2013.

[69] H. M. Wallach, "Topic modeling: beyond bag-of-words," in *Proceedings of the 23rd international conference on Machine learning.* ACM, 2006, pp. 977–984.

[70] A. McCallum, K. Nigam *et al.*, "A comparison of event models for naive bayes text classification," in *AAAI-98 workshop on learning for text categorization*, vol. 752. Citeseer, 1998, pp. 41–48.

[71] T. Joachims, "Text categorization with support vector machines: Learning with many relevant features," in *European conference on machine learning.* Springer, 1998, pp. 137–142.

[72] M. Stikic, R. R. Johnson, D. J. Levendowski, D. P. Popovic, R. E. Olmstead, and C. Berka, "Eeg-derived estimators of present and future cognitive performance," *Frontiers in human neuroscience*, vol. 5, p. 70, 2011.

[73] R. Bekele and W. Menzel, "A bayesian approach to predict performance of a student (bapps): A case with ethiopian students," *algorithms*, vol. 22, no. 23, p. 24, 2005.

[74] M. Xenos, "Prediction and assessment of student behaviour in open and distance education in computers using bayesian networks," *Computers & Education*, vol. 43, no. 4, pp. 345–359, 2004.

[75] J. Cheng and R. Greiner, "Learning bayesian belief network classifiers: Algorithms and system," in *Conference of the Canadian Society for Computational Studies of Intelligence.* Springer, 2001, pp. 141–151.

[76] M. Mayo and A. Mitrovic, "Optimising its behaviour with bayesian networks and decision theory," 2001.

[77] C. Conati, A. S. Gertner, K. VanLehn, and M. J. Druzdzel, "On-line student modeling for coached problem solving using bayesian networks," in *User Modeling.* Springer, 1997, pp. 231–242.

[78] R. O. Duda, P. E. Hart, and D. G. Stork, *Pattern classification.* John Wiley & Sons, 2012.

[79] Z. A. Pardos and N. T. Heffernan, "Modeling individualization in a bayesian networks implementation of knowledge tracing," in *International Conference on User Modeling, Adaptation, and Personalization.* Springer, 2010, pp. 255–266.

[80] C. Conati, A. Gertner, and K. Vanlehn, "Using bayesian networks to manage uncertainty in student modeling," *User modeling and user-adapted interaction*, vol. 12, no. 4, pp. 371–417, 2002.

[81] G. E. Hinton and R. R. Salakhutdinov, "Reducing the dimensionality of data with neural networks," *science*, vol. 313, no. 5786, pp. 504–507, 2006.

[82] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in neural information processing systems*, 2013, pp. 3111–3119.