Inferring Mental Workload Changes of Subjects Unfamiliar with a Touch Screen Game through Physiological and Behavioral Measurements

A Dissertation Presented
by

Payam Parsinejad

to

The Department of Mechanical and Industrial Engineering

in partial fulfillment of the requirements
for the degree of

Doctor of Philosophy

in

Interdisciplinary Engineering

Northeastern University
Boston, Massachusetts

June 2016

Dedicated to my parents

Homa and *Manouchehr Parsinejad*

and to my brother

Farzan Parsinejad

For their unfailing love and support

Contents

List of Figures

List of Tables

Acknowledgments

> **"Is it not a strange fate that we should suffer so much fear and doubt for so small a thing? So small a thing! . . ."**
>
> *– J.R.R. Tolkien,*
> *The Fellowship of the Ring*

It is my extreme pleasure to thank my advisor Professor Rifat Sipahi for his guidance, patience, encouragement and unfailing moral support during the entire course of this work. Professor Sipahi you have been a great academic mentor to me and I feel honored and blessed at the same time to have the opportunity to work with you.

I also would like to sincerely thank other members of the committee, Professor Yingzli Lin and Professor Jose Martinez Lorenzo for their invaluable advice, challenging questions, comments and excellent feed back all of which aimed at making my thesis better.

Next I must thank my colleagues: Dr. Payam Nia for his constructive discussions and support on the common area we were doing research, Dr. Nai Qian Zhi for her help in some parts of literature review in Chapter 6, and Dr. Yolanda Vaqueiro for her great contribution in constructing Chapter 5.

My next thanks go to the Department of Mechanical and Industrial Engineering and College of Engineering for giving me the opportunity to fulfill my dreams in obtaining this degree and facilitating means in the best way possible. Professor Hameed Metghalchi: thank you for your kind support and sharing many great thoughts on problem solving in different areas. Lebaron Briggs, Jeffery Doughty, Kevin McCue, Tom Olson, Dr. Jim Papadopoulos, and Dr. Bridget Smyser: thank you for always being there to help me and make things work while I was frustrated. Sara Belotti, Joyce Crain, Noah Japhet and Katherine Swan: thank you for helping me with all the departmental work.

I also would like to thank my great friends, Dr. Baktash Babadi, Dr. Behtash Babadi, and Dr. Mehrtash Babadi for providing the best moral and emotional support one could wish.

Particularly I would like to thank my parents and my brother for their constant encouragement, continuous emotional support and for their limitless love.

This study is supported in part by DARPA Young Faculty Award #N66001-11-1-41617, Sipahi's College of Engineering Faculty Fellow funds, and Mechanical and Industrial Engineering Department at Northeastern University. The human subjects experiments in this study were conducted under IRB# 11-11-19 at Northeastern University. Any opinions in this dissertation are those of the author and do not reflect viewpoints of funding agencies.

Abstract of the Dissertation

Inferring Mental Workload Changes of Subjects Unfamiliar with a Touch Screen Game through Physiological and Behavioral Measurements

by
Payam Parsinejad
Doctor of Philosophy in Interdisciplinary Engineering
Northeastern University, June 2016
Rifat Sipahi, Adviser

Many tasks can be demanding for human operators, including operating an aircraft, driving a vehicle, and making decisions in an air traffic control setting. These tasks, depending on their complexity, cause increased mental workload on humans, which could then lead to human errors. Understanding the interaction dynamics between the human operators and tasks requires effectively detecting and carefully evaluating human mental states. If done successfully, this would help design ways by which the machine can infer mental states and respond intelligently to the operator in a way to assist with the objective to reduce the probability of human error in a task.

Even if the operators are experts in many tasks, when they are faced with a challenging situation they are not familiar with, then their task execution may not be perfect and human error may still be inevitable. To understand this phenomenon and design an inference scheme to detect it via a machine, a touch-screen air traffic management game is designed with two unique difficulty levels, easy and difficult, requiring different mental workload levels. While volunteering subjects are trained and are hence familiar with the easy level of the game, they are only knowledgeable of the difficult game without any training experience.

Two main results of this dissertation are as follows: (*a*) Outcome of the experiments indicates that data collected from subjects' heart rate and skin conductance as well as subjects' finger-stroke patterns on the touch screen can all help flag unfamiliarity of subjects in the difficult game. (*b*) Subjects' behavioral patterns are used to create models using machine learning techniques, whereby the models can autonomously predict the game difficulty solely by tracking subjects' movements in real time. Experiments with newly recruited subjects indicate that such models can indeed predict what game difficulty the subjects are encountering, even if we have no priori knowledge of the game level.

Results obtained in this dissertation point out many future opportunities in synergistic human-machine systems, and pave the way toward real-time adaptive machines that can perform inferences to evaluate the probability of a human error in critical tasks, and can in turn provide a set of assistance modalities to the humans, with the aim to minimize such errors.

Chapter 1

INTRODUCTION

1.1 Objective of Research

Under excessive mental workload humans have difficulty to process the information they perceive from the environment. Mental workload can be induced to humans especially when they encounter unfamiliar situations or they suffer from lack of training in accomplishing a specific task. Such unfamiliarity or inexperience leads the operators to be prone to error in decision making which may cause catastrophes and casualties in many scenarios. Therefore, a machine that can infer subjects' inexperience in real-time would be valuable, as the machine could provide assistance when subjects are face with unexpected challenging situations.

Affective computing and analysis of human's behavioral patterns could help with effectively measuring / evaluating operators' mental states. Such evaluation schemes can then be implemented in machines along with algorithms, to provide intelligent assistance to the subjects in real-time (adaptive machine).

The goal of this research is to investigate how certain anomalies in human operators' physiological measurements and in their behavioral patterns manifest themselves when operators face with an unfamiliar and a challenging situation, even if operators are experts in doing the same tasks in a familiar environment. Further, we would like to study whether or not existing tools in affective computing could be used to identify such inexperience in a timely, efficient and reliable manner. If this could be done, then the machine could be made intelligent and could detect such anomalies in real-time, and then respond to humans with some assistance to prevent possible decision making errors.

1.2 Background and Problem Statement

Many tasks can be demanding for human operators, including operating an aircraft [131], driving a vehicle [89], and making decisions in an air traffic control setting [66]. These tasks, depending on their complexity, cause increased mental workload on humans, which could then lead to human errors.

"Workload" as an important concept within the field of human factors and ergonomics [89], is defined as "...set of task demands, as operator effort, or as activity or performance" [42]. When interacting with a machine, under excessive "mental workload", as human operators' capacity is limited, they may not be able to fully process information they perceive from the environment. For example, air traffic control (ATC) is a demanding activity for human operators [16, 98]. This activity increases mental workload, which could lead to human errors [28]. In contrast, under much lower mental workload, human operators may experience boredom and they tend to make mistakes [116]. Even if the operators are experts in many tasks, when they are faced with a challenging situation they are not familiar with, their task execution may not be perfect either, and human error may still be inevitable. Thus, if mental workload level is at its optimal level, then this is expected to help human operators improve their performances [58]. With this aim, understanding the interaction dynamics between the human operators and tasks is necessary, and requires effectively detecting and carefully evaluating human mental states. If done successfully, this would help design ways by which the machine can infer mental states [104].

In order to successfully quantify and measure humans' mental states, various *subjective*, and *objective* measures have been widely used [84, 116]. Subjective methods, such as NASA-TLX [49] questionnaire, that can perform detailed inferences concerning operators' mental workload; is convenient, and easy to practice [84]. However, subjects' responses may not always correspond to actual mental state evaluations [84]. And further, they may not accurately reflect human operators' experiences [128]. The more objective mental workload evaluation schemes such as physiological measures based on bio-physiological sensors, appear to be more suitable for practical applications since they can provide a relatively continuous record of data over time [56].

Physiological measurements of human operators' mental states are aimed at evaluating the autonomic nervous system (ANS) activity [79]. ANS has two branches: the sympathetic nervous system (SNS) and the parasympathetic nervous system (PNS), which regulate the body's major physiological activities, including the heart's electrical activity, gland secretion, blood pressure, and respiration. Moreover, SNS triggers the body's resources for action under excessive mental workload. In contrast, PNS relaxes the body and stabilizes it into steady state [122, 79]. Physiological measurements include brain-related measures, e.g., functional magnetic resonance imaging (fMRI), electroencephalography (EEG) [43]; eye-related measures, e.g., electrooculography (EOG) [116]; muscle-related measures, e.g., electromyography (EMG) [108]; heart-related or cardiovascular measures, e.g., electrocardiography (ECG), blood volume pulse (BVP), [79]; electrodermal activity measures, e.g., skin conductance (SC) [13]. Heart rate (HR) and heart rate variability (HRV) are linked with the state of ANS [122]. To measure HR, a Blood Volume Pulse (BVP) sensor [3] could be used. The HR data is then processed to calculate HRV [83], where HRV is calculated by two different techniques: *(a)* time domain, based on inter-beat interval (IBI) time series [79]; and in *(b)* frequency domain, based on spectral analysis of the amplitudes of the IBI signal at various frequencies. HRV is used extensively in the literature to study mental effort and cognitive workload [115, 132, 133]. In general, when subjects are under excessive mental workload, HR increases [89], and HRV deceases [92, 129], although some exceptions exist [115]. Moreover, galvanic skin response (GSR) is a measure of the electrical resistance of the skin [13]. When task load increases -which leads to mental workload increase, sweat glands are activated, increasing skin conductance [122]. Since the sweat glands are also controlled by the SNS, skin conductance acts as an indicator for sympathetic activation [122] correlating to cognitive activity [89, 91, 90, 13]. GSR is a widely used tool for measuring skin conductivity -by utilizing skin conductance (SC) sensor [13]

Affective computing [104] which investigates the correlation between subjects' mental states and their performance levels in a task, offers many opportunities for the study of human mental states through physiological sensors [107], such as heart rate (HR) sensor [89, 90], skin conductance (SC) sensor [118, 90, 89, 50, 107], ECG [115] and EEG sensors [116]. Using measurements from these sensors, one can infer, for instance, stress [118, 50], cognitive workload [13, 90, 89], and arousal in subjects [82, 84], as demonstrated in car driving [50, 63, 89], office work-space [118], as well as in human computer interactions through real-world simulations [90, 66] and computer games [115, 82, 84, 116]. In [66], for instance, heart rate (HR) and heart rate variability (HRV) metrics are used to examine cognitive state of human operators during simulated air traffic tasks. Given the importance of functional state of the human operator to optimal system performance, in [132], EEG and ECG were used to monitor the functional state of subjects in real time while they performed the Multi-Attribute Task Battery with two levels of task difficulty. Further, operator functional state assessment in real time led to performance improvement when included in closed loop adaptive automation with a complex task while subjects performed an uninhabited aerial vehicle task [134]. The sensitivity of HR and SC as a measure of mental workload was also investigated, e.g., in a simulated driving environment [90]. Further, the sensitivity of these measures for differentiating tasks with presumed differences in mental workload were evaluated in real-world driving tasks [89]. In [115], ECG sensor is used to investigate subjects' mental states while they play a simulated air traffic game. In [83, 84], the efficacy of physiological measures, SC, HR, and EMG, as evaluators of collaborative entertainment technologies, i.e., in commercialized games was tested. Findings are then followed by [82] where a novel fuzzy logic model method for continuously modeling user emotional state during play experiences through SC, HR, and EMG measurements was presented.

In all the cited studies, environment, e.g., the game, plays a key role in probing mental states in order to study human-machine interactions. A large number of studies have shown that physiological measures such as SC, HR, EMG, can indeed be used to infer emotional and cognitive responses while humans are playing a game [71]; suggesting that careful design of game environments can provide a scientific platform to study many aspects of human machine systems.

Another important parameter in this research is regarding a user's expertise in the game as investigated in human factors field where many studies were devoted to understand how novices and experts perform certain tasks [94]. For example, affective computing is used to correlate between subjects' mental states and their performance levels in a task [90]. It was reported that the level of workload experienced while performing a specific task can be affected by the level of experience and skills of subjects [11, 30], and novice and expert subjects clearly experience different levels of workload when performing the same task [30]. In [135], HR and respiration measures were recorded, and the performance of experienced and novice military pilots in a F-7 jet trainer was investigated; see [11] for an extensive overview on studies utilizing affective computing (EEG, EOG, and HR) in pilots/drivers.

What we know is that experts generally are able to acquire, integrate and respond to task related information more efficiently and more effectively than novices [135]. Therefore, expert operators generally have almost *perfect* task performance, while novices at early stage of skill acquisition perform less accurately and less efficiently than experts, especially in complicated operational environments [127]. The aforementioned studies provide rich information regarding how experts make decisions, what the learning process is in novices as they become proficient with practice, how novices by practice develop proficiency to better handle such tasks [8], and how one can create virtual scenarios in order to investigate these research questions from the perspectives of neuroscience and motor control [34, 35]. With same line of thought, virtual scenarios [35] are also used in order to investigate how increased task accuracy and performance correlate to training [32], and enhance motor skills [52], and how acquired skills are retained over time [30]. In [134], subjects' *lack of training* in performing complex tasks has been investigated where inexperienced subjects benefited considerably much less from computer guided assistance in a game, compared to a group of subjects who had sufficient experience with the same game [134].

Understanding how a subject responds to situations of unexpected nature, and handles a scenario with which the subject has little to no experience, is of great importance as this knowledge could be valuable in many real-world applications involving humans. In such situations, the subject may fail to rapidly and accurately formulate a decision, and/or rush to make a decision without properly evaluating all the parameters contributing to the situation. Either way, such decisions may be poor, or arrive too late, leading to catastrophes. To remedy this, it would be extremely useful to have a computerized utility that could infer a subject's inexperience indirectly through mental states in real-time, and accordingly provide the subject optional decisions, with the aim to alleviate the subject's mental workload in this unexpected challenging situation [60] (adaptive aiding).

Adaptive interaction between humans and machines have been envisioned in many studies [50, 134, 132]. The promise in this direction is to render the machine sensitive to human's mental states in order to both create a comfortable experience for the human, and to provide assistance from the machine to the human whenever the human might need help. For example, affective computing using psychophysiological measures, i.e., EEG, is used to infer mental workload changes in real-time [132, 134] where a classification algorithms is trained to recognize low and high mental workload based upon EEG and EOG (blink interval) [132]; and further similar trained algorithms with EEG and ECG (HR) are used in a game to provide intelligent assistance to improve the performance of subjects [134].

Studies with the same line of thought record multiple bio-physiological signals to better evaluate the mental states of the human operator by combining their individual information [107, 116, 89]. This is simply motivated by the fact that *(a)* no single physiological measurement could provide sufficient information to gain insights into human operator mental states [89, 116], and *(b)* fusing multiple metrics that correlate with mental workload is practical, and could be used to easily program a machine to calculate a subject's mental workload fluctuations [134, 82].

However, bio-sensors are generally sensitive to physical movements [81]. For instance, a BVP sensor is attached to the tip of the finger using straps. These movements should be avoided as much as possible so that the process of collecting the raw HR signal and eventually detecting the individual heartbeat intervals in HRV analysis can be reliably performed. Other factors that can undesirably affect accurate estimation of HRV from IBI include: errors in data acquisition, and miss-detecting heartbeat peaks [10]. In many studies that are concerned with HR sensor artifacts, errors can be diminished either by reducing the noise within the HR signal, namely, smoothing with a moving average window, low-pass filtering [81], or designing algorithms to detect errors in heartbeats and to better extract the IBI [73, 77, 39, 9]. In the cited studies, the IBI error detection is done by utilizing various post-processing algorithms, which also rely on the analysis of the entire recorded data. Recently, in [112] an online IBI error detection algorithm is also proposed; the process is performed in real-time regardless of the HR sensor artifacts, making such an approach appealing. All the approaches surveyed above are based on the use of IBI data, which is generated by finding the maxima points in the HR signal; and if it is not performed carefully, significant errors may be generated in the estimated IBI sequence. Hence, an alternative approach in extracting the IBI data, for example, an algorithm based on the Short Time Fourier Transform (STFT) would be beneficial; since it has been shown that analyzing signals, especially when they are embedded in noise, can be more reliably done using combined 2D time-frequency processing techniques [23, 21]. Moreover, as bio-sensors are generally sensitive to environmental conditions [41], i.e., sensor movements [81], human operators' performance [115, 33], if quantifiable, could be used as another objective indicator of human's mental state.

There are some limitations observed when affective computing is used to provide real-time assistance. For instance, the accuracy of the physiologically driven classifier is biased toward subjects' proficiency in performing a task in a fixed level [134]. This suggests that optimal adaptive aiding will be achieved if the capabilities of each subject are determined and used to provide intelligent assistance. Also physiological measures, i.e., HR, fluctuate rapidly in a short amount of time. This directly affects the performance of the physiologically driven classifier where it produces the presentation and withdrawal of adaptive aiding at a too-rapid rate [134]. Moreover, analysis of physiological sensors in real-time would require recording of multiple HR cycles before an inference can be made [12]. This could could be troublesome when a type of assistance is needed quickly. Finally, physiological measurements of subjects with lack of experience in a task may not correlate well with training algorithms and real-time detection schemes [134].

Mental workload changes manifested by subjects' inexperience can also be inferred by studying subjects' behavioral patterns. Such patterns are affected when subjects are engaged with tasks that demand different levels of mental effort. These patterns are correlated with the strategies subjects develop (or are unable to develop) while trying to cope with task difficulty [132]. Moreover, if one could relate subjects' behavioral patterns to subjects' level of coping with a task, and there is some evidence supporting this in different contexts [41, 33], this would be a valuable "sensor" as such detections can be made much faster with higher bandwidth and in real time, through high-rate measurements such as via touchscreen mouse clicks and light-weight accelerometers attached to human body. Some studies already investigated behavioral patterns, especially hand gestures in touch-screen applications with iPads using subjects' decision making times [41] and keystroke dynamics [33] as metrics with device usability as the research focus; others assessed decision making times with respect to a number of choices needed to be evaluated by the subjects, see the well-known Hick-Hyman Law [55, 62] and Fitts Law [38]; and others implemented experimental studies in which subjects are trained in the tasks to be performed [115, 134, 116].

From the above viewpoint, it is critical to emphasize that behavioral patterns that are different from conventional performance metric have been widely investigated in human computer interaction (HCI) applications, i.e., in gaming [32, 41, 134, 31], in real-world car driving [89, 107] and in simulated laboratory setting applications [33, 90, 116, 115, 93]. The reason for this is that these behavioral patterns are end results of mental workload changes as dictated by inexperience and/or task difficulty, and are hence only loosely coupled with the specifics of a task being performed. On the other hand, performance metrics are directly and tightly associated with specific task outcomes as the very nature of definition of performance. Undoubtedly, behavioral patterns and performance metrics are related, yet these patterns provide higher-level information more directly related to mental workload and strategy development, and less dependent on the specifics of a task. This makes the analysis of such patterns attractive, as features of these patterns can be used directly to study and compare subjects even across different tasks. In this line of thought, "touch behavior" as an affective modality can indicate affective states of human operator [54], including in gaming applications [32] and as an indicator of emotions [41].

Although a large body of literature is available encompassing human factors engineering, affective computing, neuroscience, operational psychology, and human machine systems, to the best of our knowledge, there exists a number of key scientific problems that were so far not addressed in the open literature. First of all, making inferences to assess subjects' inexperience, and using this inference toward an intelligent machine assistance scheme has so far not been studied. Existing work focuses on training all subjects in all game tasks, while limiting the training duration, thereby creating novice vs. expert groups. However, this does not consider the situations of inexperience, and lack of training in an unexpected situation for which even experts fail to successfully perform. Moreover, since behavioral patterns of novices during task execution are unpredictable and variable, while those of experts are more strategic and systematic following consistent patterns, fitting a computational or mathematical model (or training a classifier) on novice subjects' behavioral patterns is challenging, and was not reported to be highly accurate. Moreover, in most of the lab level tasks in studies mentioned above, the task itself may not be directly linked to real-world applications, i.e., arithmetic operations or dual task choices, etc., limiting a subject's choices, and not allowing subjects to explore their task spaces. Further, while real-world applications using touch-based medium are becoming ubiquitous, such infrastructure in conjunction with human mental workload via affective computing have not yet been studied. Furthermore, in many of the cited studies in the area of adaptive aiding, the fitted model trained based on data for a set of subjects is tested on the very same subjects, which calls for improvement by studying the models on different subjects. Last but not least, implementation of a trained classifier or algorithms on a set of new inexperienced subjects' touch behavior data in real-time has not been explored.

1.3 Significance of this Research

Motivated by the above discussions the three main aims of this research are to:

- carefully infer the mental workload changes of human operator using well known affective computing tools (subjective and objective measurements) such as physiological measurements as well as behavioral and performance metrics, while keeping subjects' inexperience as the key parameter.
- test and verify the reliability of the touch behavior measurements as an indicator of human subjects' mental states by using known subjective and objective mental workload evaluation techniques, i.e., affective computing tools, performance, and decision making times.
- use different classification techniques which are trained using different sets of human subjects, verify the sensitivity of such trained classifier based on touched based behavioral pattern in real-time on different sets of subjects.

If successful, results obtained from this research will add to the knowledge pool toward real-time adaptive machines that can perform inferences to evaluate the probability of a human error in critical tasks, e.g., whenever workload exceeds a certain threshold, to regulate machine behavior and further provide a set of assistance modalities to the humans, with the aim to reduce humans operators' mental workload and to minimize such errors. The main aims of this research will be accomplished by delivering the following steps:

- Design and develop an open source air traffic (AT) game as well as a strategic experimental protocol which could elicit different levels of mental workload by probing the inexperience aspect of volunteering human subjects. The game should meet the scientific aspects of game design of virtual environments, with elements relevant in real-world applications where human interacts with a machine, for example in air traffic control management [87].

- Procure and incorporate the bio-physiological sensors, i.e., BVP and SC, which are non-invasively attached to human body, into our open source game.

- Use known affective computing tools to infer and quantify human's mental workload using the sensor readings and performance metrics (touch behavior) as an indicator of subjects' inexperience with a game while the subjects transition from a well-known task to a task they are not experienced within the game.

- With the advantage of recording BVP and SC data, fuse these sensory data together in order to better evaluate human operator mental states without any complex training procedures. Verify this new method on a set of new subjects.

- Design ways by which HRV metric can be estimated even without computing IBI time series, and regardless of existence of noise in sensory data acquisition.

- Assess and analyze human subjects' behavioral patterns by using known and proposing new measurements which are directly related to subjects performance, effort, and decision making times in the AT game in order to detect the effect of inexperience with game difficulty on human subjects' performance, and further to study inexperienced subjects' mental workload changes.

- Develop training (calibration) models using the touch based behavioral measures and use these models in experiments with new subjects to infer subject's inexperience in real-time. Test the sensitivity of the model and develop techniques to re-construct the associated tasks with certain reliability.

1.4 Structure of Dissertation

This dissertation is structured in the following manner:

Chapter 2 provides an overview of affective computing tools. Further, it describes the basic properties of some physiological signals, specifically heart rate and electrodermal activity related signals in relation with autonomic nervous system (ANS), ways to quantify and extract known features from these signals, followed by the detailed background on the use of these tools.

Chapter 3 describes the experimental methods followed in this research, from the design and development of a game eliciting mental workload (air traffic (AT) game) to the mechanism of signal acquisition. Experimental protocol, procedures and the pool of the subjects follow.

Chapter 4 presents the results of the NASA-TLX, and BVP and SC features analysis; the total of four important features extracted from the signals, in inferring the changes in mental workload manifested by subjects' inexperience in the game. Further, we propose a new metric, called combined metric score (CMS) as the result of fusing the BVP and SC features, and we investigate how well this new metric is able to distinguish different levels of mental workload on different human subjects.

In Chapter 5 we introduce a new approach to accurately calculate a metric called pNN50 from noisy BVP signals. This approach is demonstrated first on synthetic noisy signals and next on BVP signals measured in Chapter 4. Results and effectiveness of the approach are then quantitatively assessed.

In Chapter 6 we introduce behavioral metrics as the results of subjects' interaction with a touch monitor. The power of the proposed behavioral metrics in differentiating different levels of the mental workload is investigated by comparing it with physiological metrics (HRV), NASA-TLX scores as well as task performance metrics that are established and studied in Chapter 4.

Finally, in Chapter 7, we use touch based behavioral measures that are verified in Chapter 6 to train a model (calibration) to infer subject's inexperience in real-time. We use the well known classifier techniques used in the field of human factors and human computer interaction. After finding the best calibrated method which holds the least error in classification results, we test its sensitivity in new sets of experiments with different subjects (re-construction). Further we present the preliminary results of the analysis in order to improve the performance of the classifier in distinguishing the touch based measures in different levels of the game (re-configuration) based on the variability characteristics of the classifiers.

The summary of significant achievements of this research along with the future research directions are presented in Chapter 8.

Chapter 2

BACKGROUND: AFFECTIVE COMPUTING TOOLS

This chapter provides the background information for this dissertation topic. It begins with the introduction of the affective computing concepts, followed by the description of the autonomic nervous system (ANS), which controls the physiological response related to a subject's mental states. Next section presents selected examples of prior work in recognizing mental workload from physiological changes. This chapter also presents two different types of physiological signals used in this research for the purpose of affective sensing: the galvanic skin response (GSR) or electrodermal activity, and the blood volume pulse (BVP). The final section shows some examples of other prior investigations related to experiments designed to detect emotion from physiological signals.

2.1 Affective Computing

In the field of Human Machine/Computer Interctions (HMI/HCI), affective computing [104] is the exploration of methods that enable computer systems to sense and adapt to the affective state of human subjects. Affective computing involves not only emotion detection, but also extends to the implementation of emotions, and attempts to give the computer the ability to recognize and express emotions. Initially proposed by Hudlicka [60], *three* key components in the affective computing process are identified [6]: (*i*) Affect Sensing and Recognition, (*ii*) User Affect Modeling / Machine Affect Modeling; and (*iii*) Machine Affect Expression.

From the figure above, it is clear that affective sensing is a core aspect of affective computing in general. A variety of approaches have been proposed to meet the challenge of affect recognition, including the use of psychophysiological measures.

This dissertation focuses on affective reorganization by monitoring of psychophysiological signals, with the advantage that they reflect autonomic nervous system (ANS) reactions and are thus difficult to intentionally distort or supress.

2.2 Autonomous Nervous System

The autonomic nervous system (ANS) regulates the body's major physiological activities to helps human adapt to changes in the environment [122]. The ANS affects organs, such as heart's electrical activity, gland secretion, blood pressure, and respiration. In addition, the ANS activity mediates stress responses and emotional arousal.

The ANS has two branches: (*i*) the sympathetic nervous system (SNS), and (*ii*) the parasympathetic nervous system (PNS). The SNS mobilizes the bodys resources for action under stressful conditions. When fully activated, this division readies the body for a crisis that may require sudden, intense physical activity, which is commonly known as the "fight-or-flight" response, and corresponds with arousal and energy generation. In contrast to the SNS, the PNS relaxes the body and stabilizes the body into steady state. This division stimulates visceral activity and is associated with the relaxation of the body, which is known as a "rest and digest" response [79].

The autonomic nervous system (ANS) has a direct effects peripheral physiological response in individuals, such as skin resistance, heart rate, digestion, respiration rate, breathing, and etc.. For example, The normal rhythm of the heart is controlled by membrane processes of the cardiac sinoatrial (SA) node, which are modulated by innervation from both the sympathetic and parasympathetic divisions of the autonomic nervous system [18, 76]. Therefore, It is critical to measure and monitor the physiological signals to detect a human subjects' affective state.

2.3 Physiological Signals

2.3.1 Skin Electrodermal Activity

Electrodermal activity refers to the electrical properties of the skin. Also called the galvanic skin response (GSR) is easily measured as either skin resistivity or skin conductance, although the most common use is as a measure of skin conductance [13]. Electrodermal activity is one of the most commonly used physiological responses in psychophysiological research and in computing systems that integrate body responses [29].

The electrodermal response is divided into two main types: (*i*) the tonic baseline and (*ii*) the short term phasic responses on the baseline [120, 84]. The tonic baseline refers to the general conductance of the skin (which is measured in this work), while the phasic responses are deviations from the baseline resulting from a stimulus [13]. There are specific sweat glands, called the eccrine sweat glands, which are used for measuring GSR [36]. Located in the palms of the hands and soles of the feet, these sweat glands respond to psychic stimulation instead of simply to temperature changes in the body [74].

Skin conductance response is correlated to changes in cognitive activity [13] as well as arousal [75], and has been used as an indicator of mental workload in many applications [89, 90]. It is considered the most sensitive response used in the detection of deception (lie detectors) [13].

Devices used to measure skin conductance come is a wide range [13]. Interests in using "wearable devices" have been increased in recently [105]. The MIT Media Lab has designed a glove called the galvactivator, GSR rings and bracelets, and GSR shoes [106]. For scientific research, Thought Technologies produces a number of physiological sensor units and the accompanying skin conductance sensor. In this study, skin conductance is measured by using surface electrodes sewn in straps that were placed around two fingers on the same hand, see Section 3.2.2.3. It is known that finger clips were as responsive to skin conductance response as on the feet [85, 74].

2.3.2 Skin Conductance Features

SC measurement can be analyzed following the studies in [36, 29, 13]. Important features of Skin Conductance include the DC level, or **Skin conductance level (SCL)**, which is the tonic level of electrical conductivity of the screen [29]. The **mean SCL**, which is the average of skin conductance response over a period of time. It is common for SCL to gradually decrease while subjects are at rest [29], rapidly increase when novel stimulation is introduced [13], and then gradually decrease again after the stimulus is repeated. SCL is reported to increase with increase in cognitive load and task demand [89, 90].

The distinctive short waveforms (SCR), Figure 2.1, is usually called the skin conductance response (SCR) and is considered to be useful as it signifies a response to internal/external stimuli [13, 69]. Some SCR important features are: **SCR amplitude**, which increase in conductance shortly

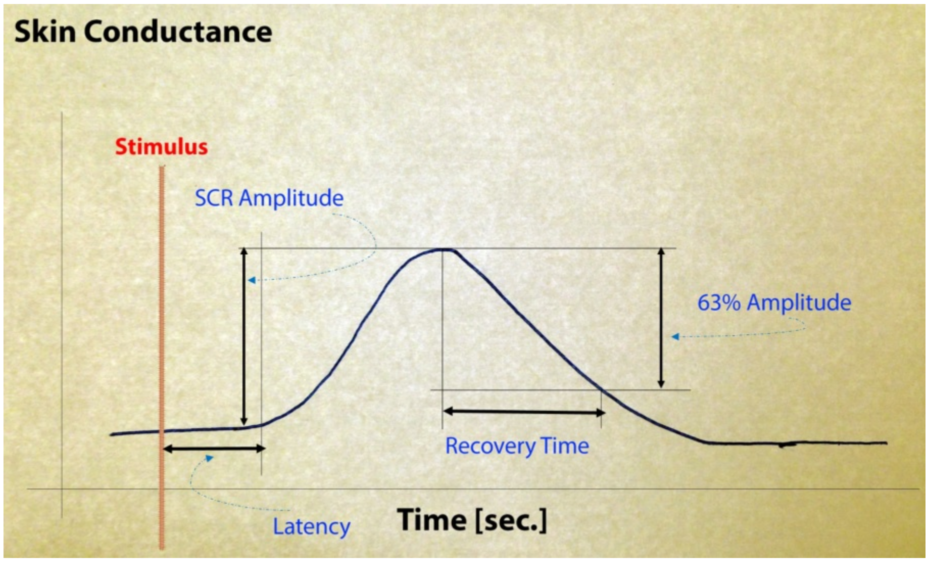


Figure 2.1: **Ideal Skin Conductance Response with typical computed features.**

following stimulus onset [29]; **SCR latency**, which is the temporal interval between stimulus onset and SCR initiation; **SCR rise time**, which is the temporal interval between SCR initiation and SCR peak; and **SCR half recovery time**, which is the temporal interval between SCR peak and point of 63% recovery of SCR amplitude.

2.3.3 Cardiovascular System: Blood Volume and Pulse Volume

The organs that regulate blood flow through the body are the cardiovascular system. Measures of cardiovascular activity include heart rate (HR), interbeat interval (IBI), heart rate variability (HRV). Heart rate indicates the number of heart beats each minute, HRV refers to the changes of the interval between consecutive heartbeats, BVP refers to the amount and timing of blood flowing through the periphery of an individual. In this work, HR and further HRV are analyses by measuring the BVP.

BVP is measured using a plethysmograph [3], which relies on the optical properties of the tissues for example, finger [120]. In this technique, a light source is passed through the tissue, and the amount of light bounced back is measured by a photoelectric transducer [120, 3]

The cardiac cycle in BVP signal have two main periods, (*i*) systole and (*ii*) diastole, Figure 2.2. Systole corresponds to the period where the heart contracts and expels the arterial blood to the arteries, and diastole to the period where the heart is relaxed and receives venous blood from the

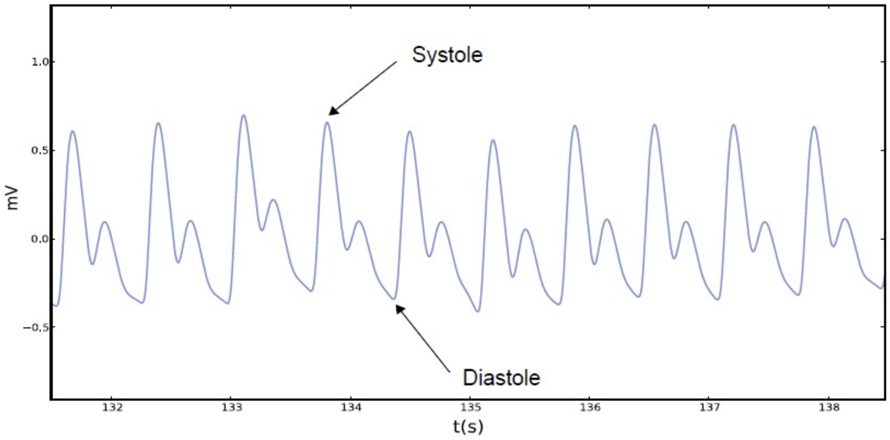


Figure 2.2: **Sample of the BVP (PPG) Signal Recorded in our Laboratory.**

veins. Since the amount of light that is detected by a BVP sensor is proportional to the volume of blood arriving from the heart in each moment, it is easy to understand that these periods will cause different levels of light detection [46].

Heart rate (HR) indicates the number of contractions of the heart each minute. HR has been used to differentiate between different levels of mental workload [91, 90].

Heart rate variability (HRV) refers to temporal variation in the intervals between consecutive heartbeats in sinus rhythm. It is found that heart rate irregularity is gradually supressed when task difficulty in creases [67]. HRV has been used extensively as an indication of mental effort and stress in adults. In high stress environments such as air traffic control [115], HRV is found to be a very useful measure. When subjects are under stress, HRV is suppressed and when they are relaxed, HRV emerges [82]. Similarly, HRV decreases with increases in mental effort [115] and cognitive workload [129], but as the mental effort needed for a task increases beyond the capacity of working memory, HRV will increase [114, 115]. Many researchers have found significant differences in HRV as a function of mental workload (meshkati1988heart, mulder1979sinusarrhythmia).

2.3.4 HR and HRV Features

Data series used in HR/HRV analysis are time series containing beat-to-beat intervals (IBI) extracted from BVP signal. Figure 2.3 shows a hypothetical BVP signal and and how IBI's are determined based on the consecutive heart beats. NN_k and NN_{k+1} represent the k^{th} and $(k+1)^{th}$

data point of the IBI time series signal. The IBI time series of an BVP segment containing n beats is given by

$$IBI_k = NN_{k+1} - NN_k, : 1 \leq k \leq n \qquad (2.1)$$

where NN_k is the time location of the k^{th} beat.

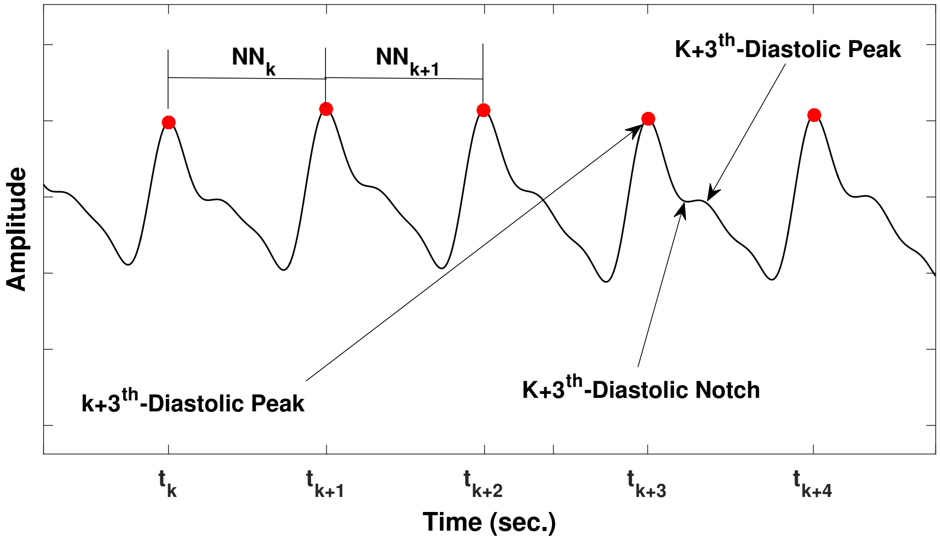


Figure 2.3: **Determination of IBI.**

Pre-processing of IBI time series data is frequently required before HRV analysis to reduce analysis errors. The three primary types of IBI pre-processing are ectopic beat/interval correction[125], detrending [136], and IBI re-sampling [119], see after [111].

HR and HRV analysis can be categorized into (*i*) time-domain and (*ii*) frequency-domain analysis: In time domain analysis, from the pre-processed IBI time series a number of parameters are calculated [19]: **HR:**is the average estimation of the heart rate in beats per minutes, which is known to increase with increasing task demand and mental workload [90, 89]; **RMSSD**: the root mean square successive difference of intervals, which is inversely proportional to stress [99]; and **pNN50%**: the number of successive difference of intervals which differ by more than 50 ms expressed as a percentage of the total number of BVP cycles analyzed, and is which is significantly lower in a mental task than during rest [123].

Fluctuations in HR are often thought to be periodic and occurring on many time scales [25]. Therefore, quantifying these fluctuations within the IBI time series can be done by calculating the power spectrum density (PSD). That is, in frequency domain, analysis of HRV is conducted on the amplitudes of the cardiac interval signal at various frequencies [116]. The HRV power spectrum can be classified into three major frequency bands, which are each associated with different functional influences on the IBI, namely (*i*) the **low-frequency (LF) band (0.02–0.06 Hz)**, associated with the regulation of body temperature [19], (2) the **mid-frequency (MF) band (0.07–0.14 Hz)**, associated with the short-term regulation of arterial pressures [19], and (3) the **high frequency (HF) band (0.15–0.50 Hz)**, reflecting the effects of respiratory activity on the cardiac interval signal [95]. Many studies showed that an increase in mental effort was typically related to a reduction in the power associated with the mid-frequency band in the HRV power spectrum, implying a temporary suppression of normal arterial pressure regulation [1, 95, 126]. The frequency range sensitive to changes in mental effort is between 0.06 and 0.14Hz [126], while the area between 0.22 and 0.4Hz reflects activity related to respiration [95, 65]. Integrating the power in the band related to mental effort provides a measure of HRV [116].

2.3.5 Additional Sensing Modalities

Besides the SC and BVP signals, there are other sensing modalities that can be monitored for affective sensing, mental workload such as Electroencephalography (EEG), blood pressure (BP), and Pupillometry (PD), respiration, electromyogram (EMG).

EEG is a technique for recording electrical activity from the scalp related to cortical activity which has been successfully demonstrated as an indicator of mental workload in [116, 132]. Pupillometry is the study of the dilation of the pupil [120], which is a useful measure changes in mental effort [109]. Respiration can be measured as the rate or volume at which an individual exchanges air in their lungs. Emotional arousal, and increase in mental workload increases respiration rate while rest and relaxation decrease respiration rate [120, 129]. Blood pressure indicates how much pressure is needed to push blood through the system of arteries, veins, and capillaries. Blood pressure is sensitive to highly stressful situations [120].

2.4 Summary

In this chapter, the affective computing concepts is introduced followed by the basic description of the autonomic nervous system. Next, the physiological background and the basic properties of some physiological signals, specifically heart rate and electrodermal activity related signals in relation with autonomic nervous system (ANS) are described. We detail ways to quantify and extract known feature from the physiological sensor data and their correlation with human subjects' mental workload changes.

We hypothesis that data collected from subjects' heart rate and skin conductance could successfully indicate subjects' inexperience in a unfamiliar situation. If successful, this funding can serve as a good foundation toward validating the correlation between subject's touch behavioral pattern and their performance in familiar/unfamiliar situations (changes in mental workload) with subjects' physiological.

In Chapter 3, the experimental setup integrating the software, the sensory measurements devices and experimental protocol are demonstrated.

Chapter 3

EXPERIMENT DESIGN

The overall goal of this research is to overall goal of this research is to investigate how a certain anomalies in human operators' physiological measurements or in their behavioral pattern are presented when they face with an unfamiliar and a challenging situations, even if they are experts in doing the same tasks in familiar environment. And further, we would like to study whether or not there exist tools in affective computing which could be used to identify such inexperience in a timely, efficient and reliable manner. If this could be done, then the machine could be made smart and intelligent which can detect such anomalies in real-time and respond to humans with some assistance to humans to prevents their error.

For this, an experimental setup and a corresponding protocol are defined and implemented for the study to:

- Provide an appropriate stimulus, capable of eliciting different levels of mental workload in the subjects who participating in the experiment.

- Record desired biophysiological signals accurately and synchronously while subjects interact with the machine for further analysis.

- Record the subjects behavioral patterns while interaction with machine.

What follows is complete implementation of the experimental environment together with the designed software, and the hardware components for sensory data measurements.

3.1 Software Development:Air Traffic (AT) Game

In line with the main purpose of the research, to induce mental workload on the subjects who plays interacts with a machine we designed and open-source air traffic (AT) game.

3.1.1 Rationale of the Game Design

In the design of the easy and the difficult levels of the game, the following thought process was taken. First of all, the difficult game should be challenging and non-trivial such that subjects' inexperience can truly be probed. This challenge should be based on affecting mental workload and channel subjects toward complex decision making. Secondly, since for the purpose of this research the types of real-world tasks cited in the introduction section do not require extensive physical demand, the game here should not be physically challenging either, e.g., with increased game speed. Nevertheless, we also do not want to design an extremely slow game where subjects can take too much time to develop their moves. Hence, some level of time pressure is still needed, but at a pace reasonable and comfortable for the subjects. We therefore suggest that a good level of compromise between speed and cognitive load is to create the difficult level of the game without forcing subjects' to make rapid decisions but in parallel require them to evaluate the situation, and to make a decision within a reasonable amount of time considering the time they will need to perceive, formulate a decision, and act. Furthermore, it is important that the subjects remain engaged, putting emphasis on playing the game to the best they can.

Accordingly, the main focus in the difficult game is to trigger primarily mental workload and game performance dimensions with minimal changes in the game speed. Other key considerations include allowing subjects complete degree of freedom in performing the task, without limiting their hand, arm movements with discrete number of selection options, while at the same time having the game linked to scenarios with real-world elements, and incorporation of some randomization to prevent learning.

3.1.2 Game Design

Keeping in mind the rational discussed above, the game can be designed in a number of unique ways. In this research, our game is inspired from flight control games and resembles those produced by Firemonkeys Studios, or Flight Control HD as well as air traffic management tasks found in real-world simulations [115, 16, 66]. For future developments and customization, the game is designed as open architecture in MATLAB environment using Psychophysics Toolbox (PTB) software [14, 103], and was briefly summarized in [101]. PTB interacts between MATLAB and the computer hardware providing the ability to MATLAB to attain full control of the hardware in representing various objects in the display. Such approach not only provides us extensive support for numerical calculations but also allow us to have access to Open Graphics Library (OpenGL) commands, which are used in designing graphical interactive games and laboratory experiments [14, 103]. We can also benefit from its flexible and relatively easy to learn environment, provided by rich and extensive documents, in designing our game. This allows us to further develop and modify the game as desired.

In the game, the airfield is seen from top, and the airplanes merge into the screen from random locations with constant speeds and move along random directions. The subject should touch the screen and draw trajectories for the airplanes to follow and finally land on the runways/landing areas. The airplanes that land disappear from the screen. The game consists of two different difficulty levels, namely, easy and difficult. The design of the AT game comes with simple graphics [48].

- **Game Environment (airfield):** The size of the game environment is dependent on the size of the screen and its resolution. This will be realized by using methods provided by PTB in the beginning of the game.

- **Airports:** The airports are represented by square objects. The location of the airports are fixed in the middle of the screen.

- **Airplanes:** The airports are represented by circle objects. Airplanes always move at a constant speed of 20 pixels/sec. To give a measure of speed, an airplane would need 96 seconds to horizontally travel from one side to the other side of a screen with a resolution of 1920×1080 pixels. To give a measure of speed, an airplane would need 96 seconds to horizontally travel from one side to the other side of a screen with a resolution of 1920×1080 pixels.

  Airplanes arrives into the screen every three seconds. In the beginning of the game, a variable is defined by a constant number, here is 3, in unit of second, which governs how often each airplane merges into the screen. Assume $t_0 = 0$ (beginning of the trial) and merging time for airplane$_1$, airplane$_2$, and airplane$_3$ are t_1, t_2, and t_3, respectively. In this case we have:

$$t_{k+1} - t_k = \text{Constant}; k = 1, 2, 3 \ldots n. \tag{3.1}$$

Further, the direction and positions of the airplanes at the time of entrance into the screen are randomized accordingly, based on which side of the screen they enter the screen, Figure 3.1. However, no two consecutive airplanes enter from the same side. Moreover, airplanes entering the screen are programmed to move toward the center of the screen. Therefore, no airplane exits the screen quickly, and subjects have enough time to set trajectories for the airplanes.

In order to calculate the direction, we randomly choose a location within the screen (ending location), calculate the line equation passes through the starting and ending locations. By calculating the slope of the line we get the direction, along which the airplane moves. To avoid the situation at which two or more airplanes arrives into the screen from one location, with the following technique, the arrival point of the consecutive airplanes are randomly changes around the screen.

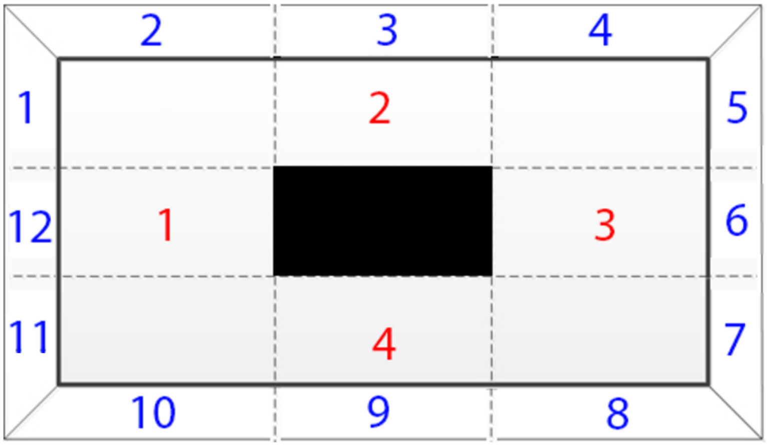


Figure 3.1: **Random location generator.**
Random location generator divides outside of the screen into 12 sections and also divides inside the screen (game environment) into 4 sections. The blue numbers are the labels for each starting zone and the red numbers are the labels for each ending zone. Two random locations are picked within a starting and ending zone to define the merge of each airplane into the screen from a starting location and to an ending location.

First, the outside of the screen is divided into twelve sections (starting zone). The area inside the screen is also divided into four sections (ending zone). We labeled each section within starting and ending zones. Two sections as starting and ending zones are randomly selected and then within each area, a random location is chosen. The two locations are assigned to an airplane. Therefore, airplane starts moving from a point in starting zone and follows a straight line ends somewhere in ending zone. The last starting zone label will be remembered. Hence, for the next airplane emerging into the screen another starting location will be selected randomly.

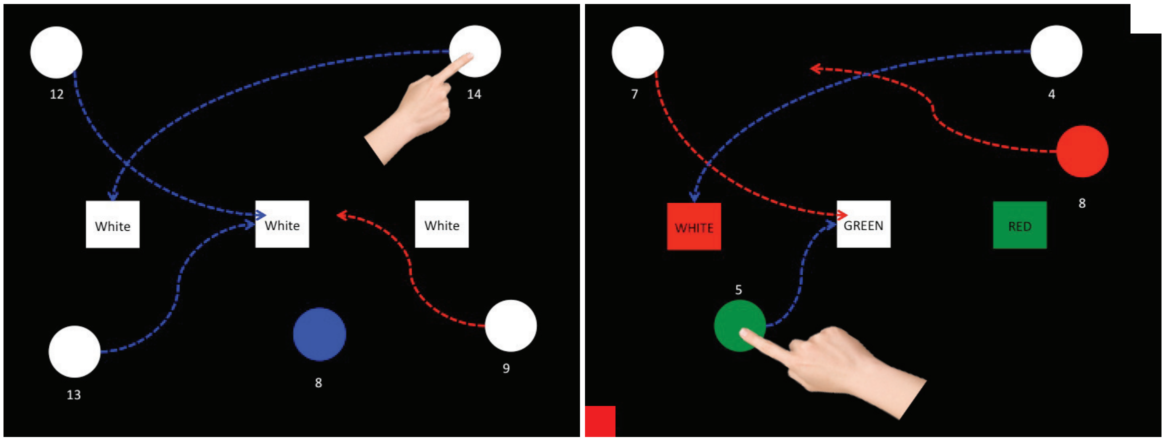
- **Human Interaction (Drawing Trajectories):** A trajectory can be generated for an airplane only by the human subject, by touching one airplane at a time on the screen. A trajectory is *assigned* if it is drawn from an airplane location and ends in the vicinity of a runway, otherwise it is *un-assigned*. When the human subject is drawing a trajectory for an airplane, the airplane follows that new trajectory without changing its speed. Once a trajectory is successfully assigned, the airplane will then follow the shortest path towards its assigned runway.

A trajectory can be generated for an airplane only by the human subject, by touching one airplane at a time on the screen. A trajectory is *assigned* if it is drawn from an airplane location and ends in the vicinity of a runway, otherwise it is *un-assigned*. When the human subject is drawing a trajectory for an airplane, the airplane follows that new trajectory without changing its speed. Once a trajectory is successfully assigned, the airplane will then follow the shortest path towards its assigned runway. When a subject fails to assign an airplane to the runway, he/she can take multiple trials to re-assign the airplane correctly. Once an airplane is assigned, its color turns “blue” indicating that this airplane is already assigned. Finally, the subjects are not instructed to prevent crash, and airplanes may cross through each other without any consequences. If left un-assigned airplanes leave the screen they will disappear and they will not bounce back.

3.1.3 Eliciting Mental Workload

To elicit different levels of mental workload induced to human subjects, the game consists of two different difficulty levels, namely, easy and difficult, Figs. 3.2a–3.2b. To create two distinct levels of difficulty in the game, the following concepts are implemented “only” in the difficult level (Figure 3.2b) of AT game:

- Stroop Color-Word Interference Test [121]. Three airports are displayed on the screen, and each one has a different color with the text of one of the color names written on each airport but not necessarily the one matching with the airport color. In other words, a color name can match with the airport color by chance. The subject has to assign the colored airplanes



(a) **Easy level.**
Airplanes (circles) are in white colors, and an assigned one is blue. Airports (squares) are in white color with text "white". "Red" and "Blue" dashed arrows are sample assigned and un-assigned trajectories, respectively, just to explain the concept. Subjects when playing the game do not see these trajectories.

(b) **Difficult level.**
Airplanes (circles) are in different colors. Airports (squares) are in different colors with different texts that switch randomly at a comfortable pace. Color indicators appear randomly on any two corners of the screen (small squares). "Red" and "Blue" dashed arrows are sample assigned and un-assigned trajectories, respectively, just to explain the concept. Subjects when playing the game do not see these trajectories.

Figure 3.2: **Different levels of AT game.**

to an airport with the matching text. The color markers, and texts on the airports switch randomly either after each trajectory assignment or five seconds after previous switch, to prevent learning, and to keep the challenge steady.

- As we noticed while piloting the game, the subjects prefer to heavily focus on the center of the screen, without paying attention to all the information displayed on the screen. To prevent this to happen, and to also distract the subjects, two small rectangles with different colors are displayed in two randomly selected *corners* of the screen. The subject should keep track of these two color indicators, and select the airplane with the color that is not indicated by the color indicators. With this setting, the subject has to also pay attention to the corners of the screen. These "indicators" appear on the screen in synchrony with the stroop color-word interference scheme described above. Whenever they switch, their locations and colors may also randomly change.

- An audio effect is played 2–3 seconds before the color indicators change their locations, in order to alert the user to select an airplane. This is expected to add higher physiological arousal [124], increased stress [51], and heart rate [72].

3.1.4 Scenario

To summarize, the two difficulty levels of the game are as follows:

- **Easy:** Arriving airplanes are in white color, merging into the screen from random locations with constant speeds and move along random straight trajectories. Three landing areas are presented with white background color and with text "white". The subject can select any airplanes and draw a trajectory toward any of the three runways (Figure 3.2a).
- **Difficult:** Different from the easy level, here *(i)* airports have different colors, Figure 3.2b; *(ii)* stroop test is active; *(iii)* color indicators must be followed to decide which color airplane can be selected; and *(iv)* an audio warning is played.

Throughout the game, the locations and the number of airports remain fixed. Also in both levels of the game, once an airplane is assigned, its color turns "blue" indicating that this airplane is already assigned. Finally, the subjects are not instructed to prevent crash, and airplanes may cross through each other without any crashes.

3.2 Hardware Setup Design

The experiments are performed using a Dell PC machine running a 32 bit Windows 7 operating system. In following the hardware setup components, namely, the display monitor, and the physiological measurement setup are detailed.

3.2.1 Presenting the Stimuli: Touch-Screen Display

The At game requires subjects navigate the airplanes to the airport by touching the screen, drawing a trajectory for the selected airplane throughout the game play. In order to have such interaction dynamic, a 21.5 Dell$^{\text{TM}}$ ST2220T multi-touch monitor with 1920×1080 resolution, and at 60 Hz frame-rate is used for displaying the game is provided for our experimental setup.

3.2.2 Physiological Measurement System

The goal of the systems designed in this research is to provide a continuous digital signal recording of the physiological variables monitored for analysis. In order to record the affective response of the skin conductance and blood volume pulse data a hardware system was integrated in order to record the physiological signals clearly and accurately. The physiological measurement system is utilized in the experiments is composed of four sub-systems: (*i*) the encoder, (*ii*) the blood volume pulse sensor, (*i*) the skin conductance sensor, and (*iv*) an API to embed the sensory data recording within the AT game.

3.2.2.1 ProComp5 Infinity Encoder

ProComp5 Infinity Encoder produced by Thought Technology Ltd[1] is used to record the seignals for the experiment. The microprocessor-powered encoder, Figure 3.3, has 5 protected pin sensor inputs; 2 channels that read data at 2048 samples/second (channels A and B), and 3 channels that read it at 256 samples/second (channels C, D and E). It is able to render a wide and comprehensive range of objective physiological signals used in clinical observation and biofeedback, and can act as an adjunct to client evaluation, assessment, prognosis, and rehabilitation.

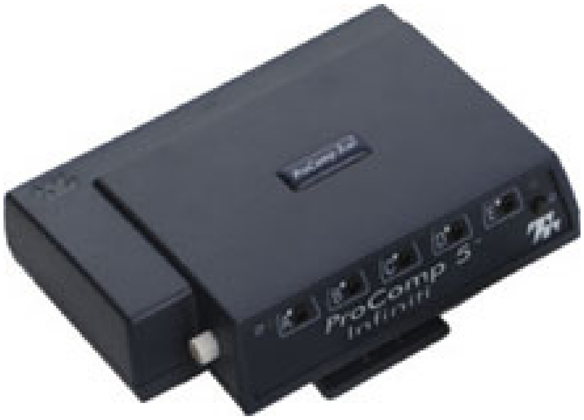


Figure 3.3: **ProComp5 Infinity Encoder.**
ProComp5 Infinity Encoder has 5 channels; 2 channels that read data at 2048 samples/second (channels A and B), and 3 channels that read it at 256 samples/second (channels C, D and E).

The encoder samples the incoming signals, digitizes, encodes, and transmits the sampled data to the ProComp Infiniti USB Adapter (TT-USB) unit, Figure 3.4. A fiber optic cable is used

[1] http://thoughttechnology.com/

for transmission to the TT-USB, providing maximum freedom of movement, signal fidelity, and electrical isolation. The TT-USB interface unit is connected to one of the host computer's USB ports. It receives the data arriving from the encoder in optical form and converts it into the USB format to communicate with the software.

Figure 3.4: **ProComp Infiniti USB Adapter (TT-USB).**
It connects to the fiber optic cable to optically isolate client from the computer.

3.2.2.2 Blood Volume Pulse (BVP) Sensor

The heart rate/blood volume pulse sensor (P/N: SA9308M) is a blood volume pulse (BVP) detection sensor, also known as a photoplethysmography (PPG) sensor, comes as a small finger worn package, Figure 3.5. Hence one could measure heart rate (HR), BVP amplitude, BVP waveform, HR and heart rate variability (HRV) feedback. BVP sensor is connected to the ProComp5 Infiniti Encoder by protected pin cables, and measure biofeedback responses and send the raw signals to the encoder. It can be used on all channels of the ProComp5 Infiniti encoder but channels A and B are preferred because they allow a higher sampling rate.

The BVP sensor does not require skin preparation as it is placed directly in contact with the skin. The sensor is placed against the fleshy part of the first joint of middle finger and tightened its position using the elastic strap.

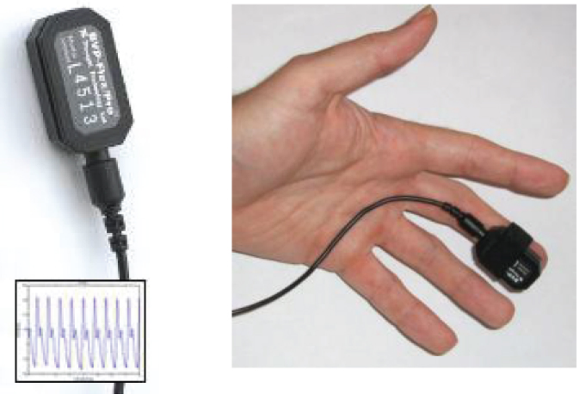


Figure 3.5: **The blood volume pulse sensor.**

3.2.2.3 Skin Conductance (SC) Sensor

The Skin Conductance sensor (P/N: SA9309M), is supplied with two finger bands, and it measures the conductance across the skin, and is normally connected to the fingers or toes, Figure 3.6. The standard measurement unit for conductance is called **Siemens**. Skin conductance is measured in micro-Siemens. Some biofeedback systems display skin conductance in micro-ohms (μm) which is the inverse of an ohm, and is the measure of resistance. These two measures, μs and μm, are equivalent. Normal readings, for skin conductance, in a relaxed state are around 2 μs, but readings can vary greatly with environmental factors and skin type.

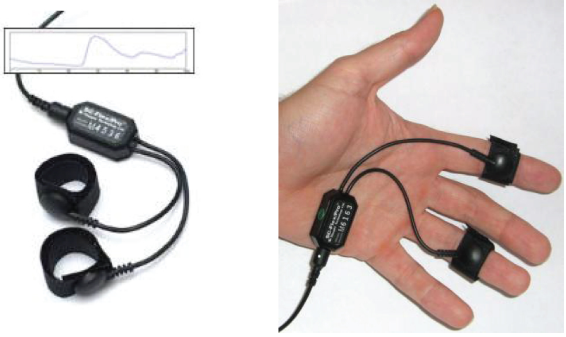


Figure 3.6: **The skin conductance sensor.**

The skin conductance sensor has two short leads that extend from the circuit box. At the end of each lead is an electrode snap similar to those on the extender cables. The SC sensor uses two replaceable electrodes that are sewn inside velcro straps. The electrode strap must be fastened around a finger tightly enough so the electrode surface is in contact with the finger pad but not so tightly that it limits blood circulation. No conductive paste should be used on the electrodes. We need to clean the electrodes with an alcohol wipe between clients. These AG/AG/CL electrode snaps should be replaced after about 50 uses or when wear is apparent.

3.2.3 Synchronization of Sensory Data within the Game

In order to embed the device (ProComp5 Infinity Encoder) sensory recording into our AT game within MATLAB environment, the TTL API SDK provided by Thought Technology Ltd.is mplemented into our software. TTL API provides an interface between encoder and client applications running on Windows. Most functions are ActiveX-compliant and can be used in a wide range of windows development environments. TTL API is supported on Windows 2000, XP, Vista and Windows 7. The SDK includes documentation demonstrating the use of TTL API with numerous windows application platforms (i.e., MATLAB). TTL API consists of two ActiveX-compliant controls, which allow access to live data streams.

Acquiring the live data from a connected encoder is achieved via the following tasks:

- Create and release a pointer object to interface with the device (encoder) within MATLAB environment. The live data via this pointer object is acquired.
- Detect and open encoder connection(s).
- Close any unwanted connections.
- Define the encoder enumeration order by assigning an encoder handler. The list of all physical channels is assessed via an assigned encoder handler.
- Create logical channels corresponding to all encoder physical channels.
- Define which physical channels are active. Active channel is a channel to which a sensor is connected.
- Synchronously start all active channels to read data from them.
- Retrieve data periodically from all active channels: (*i*) check how many samples available (which are stored temporarily in the buffer), and (*ii*) read the data samples available in the buffer.
- When finished, close all connections.

3.2.4 Overall Experimental Setup

In summary, the visual stimuli for the subject (the AT game), described in Section 3.1.2, is displayed on a A 21.5 Dell$^{\text{TM}}$ ST2220T multi-touch monitor with 1920 × 1080 resolution and 60 Hz frame-rate is used for displaying the game. While playing the AT game, the subject has the SC and BVP sensors attached to his/her left hand. The two signals are recorded in Matlab at rate of 256–1024 samples/second by using a multi-channel DAQ system (ProComp5 Infinity Encoder). Participant is wearing a headphone to hear auditatory alarm sound played during the game. Fig 3.7 shows a subject playing the game while he is wearing a headphone, and SC and BVP sensors are attached to him during the experiment.

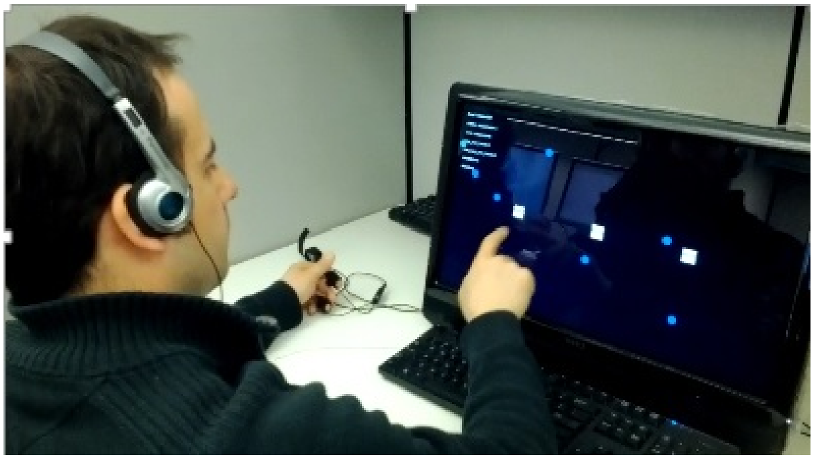


Figure 3.7: **Prof. Sipahi demonstrates the game.**

3.3 Pool of Experimental Subjects

Four sets of experiments are conducted.

- **Experiment One (Data Set 1):** In the first experiments, thirteen subjects (1 female and 12 male; age = 26.5 ± 2.3) participated in the first experiment [101].

- **Experiment Two (Data Set 2):** After approximately six weeks later, twelve of the 13 same subjects (1 female and 11 male) with a mean age of 26.4 ± 2.4 years, again participated in the experiments with the same experimental protocol (Data Set 2).

- **Experiment Three (Data Set 3):** Another twelve subjects (4 female and 8 male) with a mean age of 24.7 ± 3.3 years participate in the third experiment with slight difference in experiment protocol as explained in Section 3.4.
- **Experiment Four (Data Set 4):** Eleven subjects (6 female and 5 male) with a mean age of 24.18 ± 2.27 years participate in the fourth experiment following the porotocol as explained in Section 3.4.

All the subjects were from diverse ethnic backgrounds. Due to the nature of the tasks, the following considerations were made when choosing the participants:

- Participants had to be fluent in English (to avoid difficulty in understanding the instructions for the computer game).
- Participants should have good general health (no hearing or sensing problems).
- All of the participants were right-handed.

All the subjects reported that they have prior experience in playing computer games. While this does not specifically express what type of game they had experience with, subjects having had prior experience with computer games indicates that they may have some familiarity with similar games.

3.4 Experimental Protocol

The subjects sit comfortably in front of the touch monitor, which is positioned vertically. Prior to the experiment, each subject plays the easy level of the game for two minutes to practice the game environment, the touch monitor, and drawing trajectories. With this, all subjects are expected to reach the same level of proficiency in the easy game level. Moreover, all the subjects are also instructed and presented with visual elements the rules and challenges of the difficult game, to familiarize them with this game level. This briefing is necessary because the rules in the difficult game are so unique that the subjects need to know what they will be facing and how to handle the challenges of the game. Specifically, here screen shots of the game levels, similar to Figure 3.2 are shown to the subjects, to explain them the conditions when decisions are correct or wrong including the rules of color indicators switching their positions and their colors; and on how to assign the airplanes to the runways. Experimental personnel also confirms that the subjects learn the rules by asking them questions on how to play the difficult game. Subjects however do *not* practice the difficult game.

We should note that all the subjects in Data Set 2 have past experience playing both game levels (from Data Set 1) under identical experimental protocol [101]. In other words, each subject in Data Set 2 had played 2 sessions of the easy game and 1 session of the difficult game already in Data Set 1. Six weeks later, on average, Data Set 2 was collected. We assume that subjects in Data Set 2 have not retained much of their experiences with the difficult game, and can also be deemed inexperienced with the difficult game. All the subjects in Data Set 3 have had no prior experience with any of the games. The reasoning behind the order of game segments is as follows: (*i*) Firstly, to reset and stabilize subjects' physiological states, all participants relaxed during R1. (*ii*) In many real world scenarios, subjects will be managing their tasks comfortably until a challenging scenario is encountered. Analogously, it is therefore of interest to understand the transition from an easy game to a difficult one, i.e., from E1 to D (Data Sets 1–2), and further the transition from a challenging scenario to a relaxing one, i.e., from D to E1 (Data Set 3). (*iii*) It is of strong interest to compare subjects' physiological states and performance in both easy games, and investigate whether or not the difficult game has any left over impact on the subsequent easy game. (*iv*) Lastly, it is of interest to study whether or not subjects' specific physiological states and performance found in the difficult level or in E1 are independent of the game order. In this manuscript, we will focus on the research questions (*ii*)–(*iv*) related to subjects' active play time.

In Data Set 4, subjects are instructed to play the game 10 times (trials), where in each trial, each subject plays one easy and one difficult game. Similar to Data Sets 1–3, each game lasts 1 minutes. In each trial, the order of the games are randomly defined. In other words, the subjects are not familiar with the order of the game. For example, at 1st trial one plays easy game first, and then difficult level, and in the second trial, he/she might play the game with the same order or not. This continues till the 10th (last) trial. Similar to Data Set 2, the participants are not instructed to fill out the NASA-TLX questionnaire. Also no bio-physiological data (BVP, SC) is recorded in Data Set 4.

3.5 Summary

This chapter describes the development and designing software, AT game, and integrating it with designated hardware of a experimental setup which is capable of (*i*) providing reliable stimulation to elicit different levels of mental workload in subjects, and (*ii*) record the two physiological signals, namely SC and BVP, that are used to provide comparison of the mental workload recognition performance. The experimental procedure and pool of subjects were also described.

In Chapter 4 the data analysis tools are demonstrated, followed by results of the analysis of the recording sensory data in order to understand how successful bio-physiological data are in inferring and differentiating the different levels of mental workload.

Chapter 4

ANALYSIS OF EXPERIMENTAL RESULTS

4.1 Introuduction

Affective computing [104] offers many opportunities for the study of human mental states through physiological sensors [107], such as heart rate (HR) sensor [89, 90], skin conductance (SC) sensor [118, 90, 89, 50, 107], ECG [115] and EEG sensors [116]. Using measurements from these sensors, one can infer, for instance, stress [118, 50], cognitive workload [13, 90, 89], and arousal in subjects [82, 84], as demonstrated in car driving [50, 63, 89], office work-space [118], as well as in human computer interactions through real-world simulations [90, 66] and computer games [115, 82, 84, 116]. In [66], for instance, heart rate (HR) and heart rate variability (HRV) metrics are used to examine cognitive state of human operators during simulated air traffic tasks. Given the importance of functional state of the human operator to optimal system performance, in [132], EEG and ECG were used to monitor the functional state of subjects in real time while they performed the Multi-Attribute Task Battery with two levels of task difficulty. Further, operator functional state assessment in real time led to performance improvement when included in closed loop adaptive automation with a complex task while subjects performed an uninhabited aerial vehicle task [134]. The sensitivity of HR and SC as a measure of mental workload was also investigated, e.g., in a simulated driving environment [90]. Further, the sensitivity of these measures for differentiating tasks with presumed differences in mental workload were evaluated in real-world driving tasks [89]. In [115], ECG sensor is used to investigate subjects' mental states while they play a simulated air traffic game. In [83, 84], the efficacy of SC, HR, and EMG as evaluators of collaborative entertainment technologies, i.e., in commercialized games, was tested. Findings are then followed by [82] where a novel fuzzy logic model method for continuously modeling user emotional state during play experiences through SC, HR, and EMG measurements was presented.

In all the cited studies, environment, e.g., the game, plays a key role in probing mental states in order to study human-machine interactions. A large number of studies have shown that physiological measures such as SC, HR, EMG, can indeed be used to infer emotional and cognitive responses while humans are playing a game [71]; suggesting that careful design of game environments can provide a scientific platform to study many aspects of human machine systems.

Many studies in neuroscience and human factors fields regarding a user's expertise in the game were devoted to the investigation of how novices and experts perform certain tasks [94]. Along these lines, studies utilize affective computing tools, and investigate the correlation between subjects' mental states and their performance levels in a task. It was reported that the level of workload experienced while performing a specific task can be affected by the level of experience and skills of subjects [11]. For example, novice and expert subjects clearly experience different levels of workload when performing the same task [30]. In [135], HR and respiration measures were recorded, and the performance of experienced and novice military pilots in a F-7 jet trainer was investigated; see [11] for an extensive overview on studies utilizing affective computing in pilots/drivers. What we know is that experts generally are able to acquire, integrate and respond to task-related information more efficiently and more effectively than novices [135]. Therefore, expert operators generally have almost *perfect* task performance, while novices at early stage of skill acquisition perform less accurately and less efficiently than experts, especially in complicated operational environments [127].

The aforementioned studies provide rich information regarding how experts make decisions, what the learning process is in novices as they become proficient with practice, how novices by practice develop proficiency to better handle such tasks [8], and how one can create virtual scenarios in order to investigate these research questions from the perspectives of neuroscience and motor control [34, 35]. In summary, the cited studies are based on an overarching research question that seeks to put light on how increased task accuracy and performance correlate to training [32], and enhance motor skills [52], and how acquired skills are retained over time [30].

Another research question of strong interest, and is complimentary to the above, is the investigation of subjects' *lack of training* in complex tasks. This is in line with a recent study where inexperienced subjects benefited considerably much less from computer guided assistance in a game, compared to a group of subjects who had sufficient experience with the same game [134]. Understanding therefore how a subject responds to situations of unexpected nature, and handles a scenario with which the subject has little to no experience, is of great importance as this knowledge could be valuable in many real-world applications involving humans. In such situations, the subject may fail to rapidly and accurately formulate a decision, and/or rush to make a decision without properly evaluating all the parameters contributing to the situation. Either way, such decisions may be poor, or arrive too late, leading to catastrophes. For these reasons, it would be extremely useful to have a computerized utility that could infer a subject's inexperience indirectly through mental states in real time, and accordingly provide the subject optional decisions, with the aim to alleviate the subject's mental workload in this unexpected challenging situation. This research is motivated by these observations.

In order to grasp the fundamentals of the above described problem, here, we first focus on the forward problem of analyzing subjects' inexperience with a challenging game as the subjects transition from a well-known task in the game to a task they are not experienced with. For this, we take an affective computing approach and utilize a type of air traffic (AT) game on a touch screen, developed by our team, aiming to capture a human-machine interaction scenario. Specifically, we report human-subjects experiments based on two levels of difficulty, easy vs. difficult, where subjects hands-on practiced the easy game level, but only received a briefing with visual elements regarding the rules of the difficult level.

Besides affective computing tools, the study here also utilizes post-experimental NASA-TLX surveys to assess subjects' perceived mental workload in the game, post calculates various metrics from HR and SC sensors collected in real time as the subjects played the game, and develops correlations between these metrics as well as with subjects' game performance. Moreover, a combined metric score (CMS) is presented, which combines the metrics all together under a single scalar indicator, with the aim to relate this indicator to subjects' inexperience as manifested through their mental workload changes at the face of a challenging task. CMS is calibrated on a data set corresponding to a group of subjects, and validated with statistical significance in separate experiments both on the very same subjects and on different subjects.

First, we review the physiological indicators of mental workload followed by discussion on prior work related to affective computing. In the subsequent section, we summarize our AT game design concept and experimental protocol, and summarized the data analysis tools utilized in this manuscript. Experimental results, statistical analysis, CMS formulation, and comparisons among data sets are provided next. The article ends with discussions, conclusions, and future research directions.

In Chapter 3, we detailed about are experimental setup. In there, we introduced our AT game, see Section 3.1, followed by our experimental design and procedure. In following sections, we introduce our data analysis tools in order to analyze the recorded bio-physiological signals, namely BVP and SC. Next we will investigate how the results of the analysis improve if we fuse the sensory data together. We define a new measurement, combined metric score (CMS), as a combination of 3 different heart related measures with 1 electrodermal activity related feature. This chapter ends with our conclusion on how well CMS metric could infer subjects' mental workload level when they play the game.

4.2 Data Analysis Tools

Two standard sensors, namely skin conductance (SC: Thought Technologies - model SA9309M) and blood volume pulse (BVP: Thought Technologies - model SA9308M) sensors were utilized. The sensor data were recorded at a rate of 256 samples per second in Data Set 1, and 2048 samples per second in Data Sets 2–3. While 256Hz sampling is more than sufficient for physiological measurements, we chose 2048HZ to enhance the resolution of the HRV, inspired from [19]. In all data sets, the SC signal is down sampled to 60 Hz for further analysis, which is satisfactory as sampling rates as low as 1 Hz are acceptable [13] for studying SCL.

Prior to analyzing the recorded data, in order to eliminate the artifacts, such as high frequency noise caused by variations in electrode contact and unintended movements of the fingers carrying the sensors; the BVP is passed through 512 (in Data Set 1) and 4096 (in Data Set 2 and Data Set 3) point zero-phase low-pass filter with 6 Hz cutoff frequency, following [40]. The SC signal is first down sampled to 60Hz, and then passed through 120 point zero-phase low-pass filter with 1 Hz cutoff frequency [26] for the three data sets. Further, the recorded data is split into 5 sub-data of 60 seconds, each corresponding to a segment of the game, Figure 4.1.

4.2.1 Feature Extraction from BVP Signal

In the low-pass filtered BVP data, we first identify the local maxima points using "findpeaks" routine in MATLAB, see, e.g., [137, 7]. This helps to identify inter-beat intervals (IBI). In order to reduce the error in the analysis, IBI time series are further pre-processed, via a standard beat interval correction algorithm, as suggested in [111, 5].

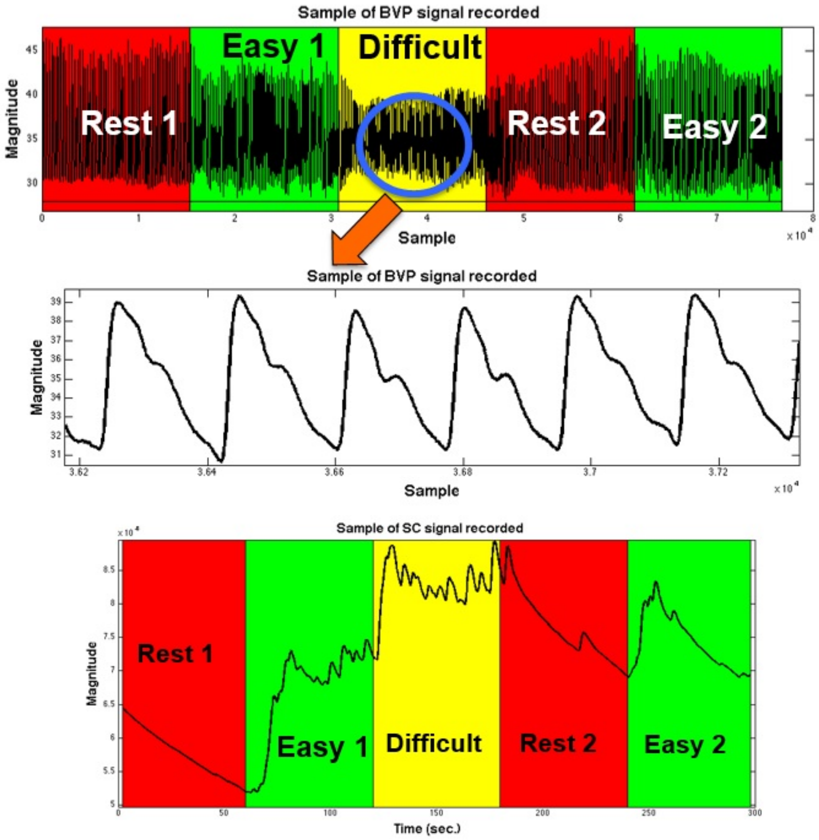


Figure 4.1: **BVP and SC are divided into 5 segments.**
Top and middle is the BVP signal, and bottom is the sample recorded SC signal. Prior to the analysis, both BVP and SC are divided into 5 sub-data, each corresponding to one segment of the game.

In this study, the following three features are then computed from the IBI data in time domain (see the complete overview on BVP features in Section 2.3.3): **mean HR**, which is known to increase with increasing task demand [90, 89]; **RMSSD**, which is inversely proportional to stress [99]; and **pNN50**, which is significantly lower in a mental task than during rest [123].

4.2.2 Feature Extraction from SC Signal

In the low-pass filtered SC data, we first remove the first and last 10 seconds of the signal data. The reason for this is justified as follows: the SC is expected to rapidly increase when a sudden stimulus is introduced at the game transition. We remove the first 10 sec. of the SC to eliminate this effect as it unfavorably supports our hypothesis. Moreover, SC measurements have certain latency [29], and by ignoring the first 10 sec. of each game, we also prevent the memory effects and the influence of the previous game rolling into the current game. We also wish to eliminate the steady state behavior in SC measurements. Although the subjects cannot learn the game due to its randomized events, we remove the last 10 sec. of each game segment to prevent any effects of relaxation of subjects.

One of the most common tonic feature of skin conductance measures, namely, the **mean SCL**, which is also used in [31, 74, 118, 89, 90, 22, 107, 122, 44], is then computed corresponding to each segment of the game (see the complete overview on SC features in Section 2.3.1). SCL was reported to increase with increase in cognitive load and task demand [89, 90].

4.3 Analysis of Experimental Results and Comparisons

Total of 37 subjects (24 unique subjects) participated in this study. While the results presented in the next section are encouraging, it is important to remark that this number can be considered at the lower limit of running statistical analysis (total of 24 unique subjects vs. 32 recommended in general).

In terms of statistical analysis, among all participants, within-subjects statistical comparisons using a repeated measures general linear model (GLM) procedure are conducted on NASA-TLX data obtained from the subjects in Data Set 1 and Data Set 3, on game-performance metric (Data Sets 1-3), and on the metrics extracted from BVP and SC sensors in Data Sets 1–3. A value of $\alpha = 0.05$ is used to define statistical significance. Greenhouse-Geisser adjustments are examined for the above metric distributions; the adjusted degrees of freedom are reported for those that violate the assumption of sphericity, otherwise degrees of freedom are reported as whole values. Whenever significant main effects appeared, post hoc comparisons of paired means were carried out using a least significant difference test (LSD). In this study, statistical analysis is conducted using IBM SPSS$^{\circledR}$ version 20, following the procedures reported in [89, 113].

4.3.1 Hypothesis and Results Summary

Since we have several metrics, comparisons, and data sets, we wish to simplify the presentation with a condensed summary section, leaving the details, interpretations, and discussions to separate sections.

Overall, for all the metrics, we hypothesize that significant differences exist among game levels E1, D, and E2 (*F* test). Further, significant differences are hypothesized in pair-wise comparisons E1 vs. D, and D vs. E2; whereas, no significant difference is expected between the two identical games, E1 and E2.

For all the metric analysis that shows consistency with the above hypothesis is marked with ✓else with ✗. Moreover, green and red colors are used respectively to indicate that a metric in the corresponding data set and game pair repeatedly shows, and does not show statistical significance in pair-wise comparisons [1].

Overall, we find out that except a few cases, all the metrics in all data sets are supporting our hypotheses. Those few cases can be found on Table 4.5 with a sign ✗. Moreover, in almost all the cases, statistical power of the analysis for each parameter was larger than 95%, with three exceptions: RMSSD in Data Set 2 (∼88%) and Data Set 3 (∼77%), and pNN50 in Data Set 1 (∼93%).

We now present all the details of the statistical analysis.

4.3.2 Subjective Workload Evaluation

The NASA Task Load Index uses six dimensions to assess mental workload: mental demand, physical demand, temporal demand, performance, effort, and frustration. Twenty step bipolar scales are used to obtain ratings for these dimensions. A score from 0 to 100 (assigned to the nearest point 5) is obtained on each scale. A weighting procedure is used to combine the six individual scale ratings into a global score by paired comparison task prior to the workload assessments. Paired comparisons require the subject to choose which dimension is more relevant to workload across all pairs of the six dimensions. The number of times a dimension is chosen as more relevant is the weighting of that dimension scale for a given task for that operator. A workload score from 0 to 100 is obtained for each rated task by multiplying the weight by the individual dimension scale score, summing across scales, and dividing by 15 (the total number of paired comparisons) [49].

Mental workload scores based on each subject's response to NASA-TLX questionnaire in Data Set 1 and Data Set 3 are listed on Table 4.1, where the scores range from 0 to 100, with 100 being the highest subjective score for increased mental workload [49]. Part of this analysis for Data Set 1 was summarized in [101].

[1] Only if the metric is found to be statistically significant among all the groups (*F* test).

Data Set 1				Data Set 3			
Subject	E1	D	E2	Subject	D	E1	E2
1	10	92.7	6	1	78	13.3	4.3
2	3	71.3	5	2	78.7	7	6.7
3	5	69.7	5	3	60.3	5	6.3
4	6	81.7	4	4	84.3	11.3	9
5	7	64.7	5	5	82.7	17.7	13
6	9	59.7	11.3	6	78	8.7	5
7	4	92.3	4	7	82	11.7	2.7
8	10.7	78.7	11.7	8	76.7	21.3	6.7
9	5	66.3	8	9	84	17	4.03
10	5	74.7	10	10	71.3	4.3	6.3
11	7.7	85.3	8	11	68.3	3.7	2.3
12	5	54.7	7.3	12	68.7	4.7	5
13	5	80.7	6.3	–	–	–	–

Table 4.1: **NASA-TLX workload scores were assessed for the 13 subjects in Data Set 1 and another 12 subjects in Data Set 3.**
The workload scores were calculated from 0 to 100 for each game (E1, D and E2), where 100 points corresponds to the largest mental workload. Notice the order of the D and E1 is changed in Data Set 3.

In both Data Set 1 and Data Set 3, significant difference is found between three groups E1, D, and E2 ($p < 0.05$), see details on Tables 4.6–4.7. Following this, pair-wise comparison reveals consistent results in both Data Set 1 and Data Set 3 where the average mental workload score for the difficult level is significantly higher than the average mental workload scores for E1 and E2 ($p < 0.05$). Moreover, in Data Set 1, there are no statistically significant differences in mental workload scores of E1 and E2 ($p = 0.308$). On the contrary, in Data Set 3, these scores in E1 are significantly higher than those in E2.

It is important to note that the above analysis is based on six dimensions, namely, mental demand, temporal demand, physical demand, performance, effort, and frustration level which are converted to NASA-TLX scores following standard procedures [49, 20]. We find that, in both Data Sets 1 and 3, mental demand was the dominant dimension followed by temporal demand and performance dimensions, as perceived by the subjects (plots are suppressed). In this sense, subjects' perceptions on average also meet our game design rationale, discussed above.

4.3.3 Objective Workload Evaluation

In order to asses the objective measure of mental workload that the subjects experienced while playing different levels of the game, we extract the following metrics: Task performance by recording subjects' trajectories and airplane assignments; mean HR, RMSSD, and pNN50 from BVP sensor; and mean SCL from SC sensor. These are the same metrics calculated in [101], but only for Data Set 1. Here, we mainly focus on comparing the results, which cover both the same subjects (Data Set 2) and balanced experiments with different subjects (Data Set 3).

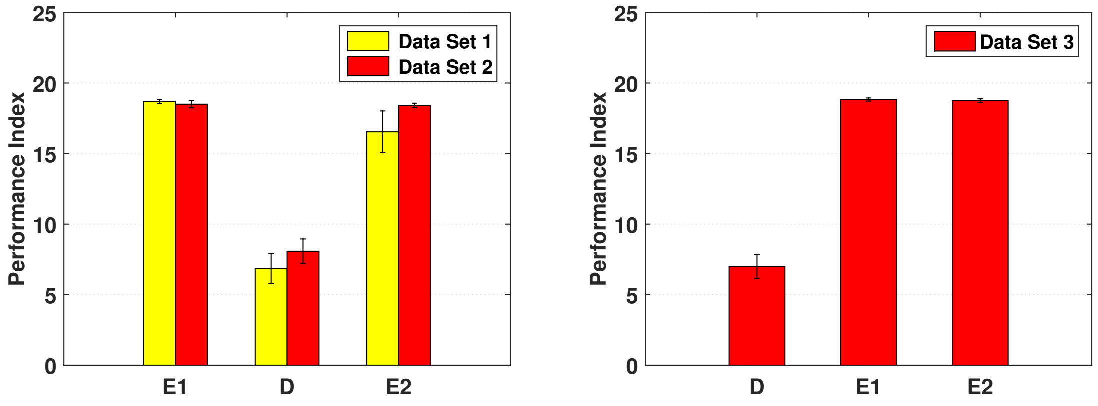
4.3.3.1 Task Performance Analysis with Comparison

Subjects play the game on a touch-screen monitor, and hence all their finger-strokes are recorded. A subject's performance in the AT game with the "goal" of assigning an airplane to an airport can be measured by the number of successful "airplane assignments" to the airports.

It is very important to note the following observation first. In the easy game, an airplane arrives into the screen every 3 sec. With the game length being 60 sec., a subject is therefore presented a total of 20 airplanes. In the difficult game, airplane arrival rate and game duration are the same as in the easy game, and as long as the subject makes correct assignments, color indicators immediately switch without imposing any unwanted delay, giving the subject a chance to react on his/her next assignment. Considering these settings, it is critical to note that a very good player in both easy and difficult game will be presented 20 airplanes in 60 sec. In this sense, the difficult game does not unfavorably reduce the number of assignments a subject can make. On the other hand, due to added challenges in the difficult game, subject's making mistakes eventually reduces the number of successfully assigned airplanes.

In Figure 4.2, the "average" of subjects' performance metric is presented, where we observe that this average is the lowest in the difficult game compared with other segments of the game. We find out that among all the subjects in Data Sets 1–3, subjects' performance is significantly affected by game difficulty, see Tables 4.6–4.7.

Moreover, in Data Set 1, subjects' performance for the difficult game is significantly lower than the average of the same value for E1 and E2 ($p < 0.05$), Figure 4.2a. The tests also indicated that subjects' performance for the E1 level do not significantly differ from the values calculated for



(a)

(b)

Figure 4.2: **Overall performance metric, on average, for each segment of the game across all subjects in Data Sets 1–3.**
Performance Index is defined as the number of successful airplane assignments. The subjects' performance decreases from E1 to difficult (D) game and further increases when the subjects played E2 game. Subjects' performance level for D, on average, is the lowest among three segments of the game. The error-bar represents the standard error.

E2 level ($p = 0.178$). These results were qualitatively the same in Data Set 2 (Figure 4.2a) and Data Set 3 (Figure 4.2b), see Tables 4.6–4.7.

4.3.3.2 BVP Feature Analysis with Comparisons

IIn light of the literature, we expect that mean HR increases; and RMSSD and pNN50 relatively decrease in the difficult game due to subjects' increased mental workload. In Data Set 1, this was consistent in mean HR, RMSSD, and pNN50 in 77%, 92% and 77% of the 13 subjects, see Table 4.2. In Data Set 2, this ratio was respectively, 100%, 83% and 57% of the 12 subjects, and 100%, 67%, and 75% of the 12 subjects in Data Set 3 (see Table 4.2).

Statistical results are also consistent: in Data Sets 1–3 the mean HR, RMSSD, and pNN50 are significantly affected by the changes in game-difficulty ($p < 0.05$), see Tables 4.6–4.7. In Data Sets 1–3, pair-wise comparisons reveal that the average of mean HR for the difficult game is significantly higher than the average of the same metric for E1 and E2 ($p < 0.05$), Figure 4.3a–4.3b. However, in Data Set 1, there was also statistically significant difference between averages of normalized mean HR values for E1 and E2 ($p < 0.05$). In contrast, Data Sets 2–3 did not reveal any

Feature	Data Set 1			Data Set 2			Data Set 3		
	C	IC	CR*	IC	IC	CR**	C	I	CR**
mean HR	10	3	77%	12	0	100%	12	0	100%
RMSSD	12	1	92%	10	2	83%	8	4	67%
pNN50	10	3	77%	7	5	58%	9	3	75%
mean SCL	11	2	85%	11	1	92%	11	1	92%

Table 4.2: **List of the consistency/supporting rate of the features extracted from the BVP and SC signal across all the subjects in Data Sets 1–3.**

The consistency (C) column shows the number of times feature values are consistent with the hypothesis, considering that the values are the highest for the mean SCL and mean HR, and are the lowest for RMSSD and pNN50 in the difficult game; otherwise presented in the inconsistency (IC) column.

*Consistency Rate (CR) in Data Set 1 is [total number of subjects supporting consistency (C) / 13] × 100.

**Consistency Rate (CR) in Data Sets 2–3 is [total number of subjects supporting consistency / 12] × 100.

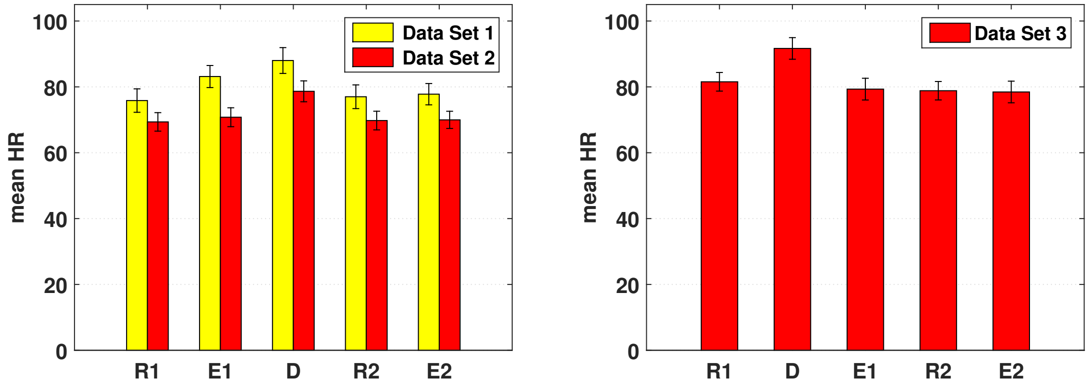
statistically significant difference for the same metric between E1 and E2, see Tables 4.6–4.7.

Moreover, pair-wise comparisons for RMSSD and pNN50 in Data Sets 1–3 also indicate that these metrics for the difficult game are significantly lower than those in E1 and E2 ($p < 0.05$), see Figures 4.4a–4.4d. The tests also indicated that pNN50 for the E1 game does not significantly differ from those calculated for E2. Similar results are obtained for RMSSD but only in Data Sets 1–2, see Tables 4.6–4.7.

4.3.3.3 SC Feature Analysis with Comparisons

We expect that the mean SCL increases much more in the difficult game as the subjects experience higher level of mental workload, partly induced by their inexperience. We observed that the mean SCL is indeed larger in the difficult game, among 85% of the 13 subjects in Data Sets 1 [101], among 92% of the 12 subjects in Data Set 2, and among 92% of the 12 subjects in Data Set 3, see Table 4.2.

In all three data sets, mean SCL values are significantly affected when subjects experienced different levels of difficulty while playing the game ($p < 0.05$), see Tables 4.6–4.7. In Data



(a) (b)

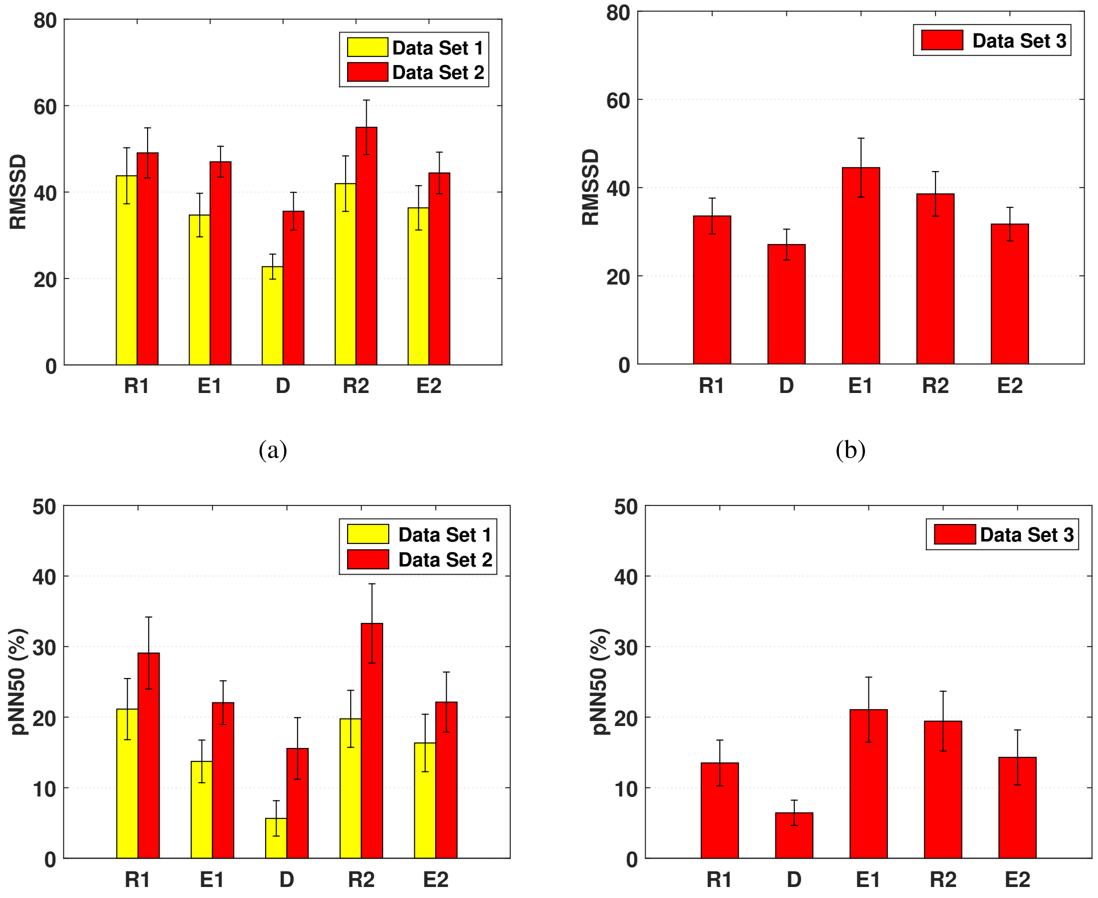
Figure 4.3: **The average values of the mean HR extracted from BVP for each segment of the game across all subjects in Data Sets 1–3.**
In Data Sets 1 and 2 (left) the mean HR increases from R1 to D when the subjects play D game, and further decreases from D to E2 in a linear fashion. Consistent behavior is *again* observed in Data Set 3 (right) where mean HR increases from R1 to D, and further decreases from D to E2 linearly. The error-bar represents the standard error.

Sets 1-3, pair-wise comparisons reveal that the average of mean SCL values in D are significantly higher than E1 and E2 ($p < 0.05$, Figure 4.5). Moreover, in Data Sets 1–2, statistically significant difference is found between the averages of mean SCL for E1 and E2 ($p < 0.05$), see Figure 4.5a, and Tables 4.6–4.7. However in Data Set 3, we observe that mean SCL between E1 and E2 does not render significant differences, see Figure 4.5b, and Tables 4.6–4.7.

4.3.4 Combined Metric Score (CMS)

4.3.4.1 Fusing Skin Conductance and BVP Metrics Using Data Set 1

With the advantage of recording SC and BVP synchronously, we now investigate their combined effects. This is motivated by the fact that *(a)* a single physiological measurement may not always provide sufficient information into subjects' mental states [116, 89]. This is expected with the presence of human subjects bringing variability (here consistency rates range from 60% to 100% of subject populations). *(b)* Having a composite metric that fuses various metrics together could better capture multiple dimensions of how subjects' physiological measurements present themselves against game difficulty and subjects' inexperience with the game.



(a)

(b)

(c)

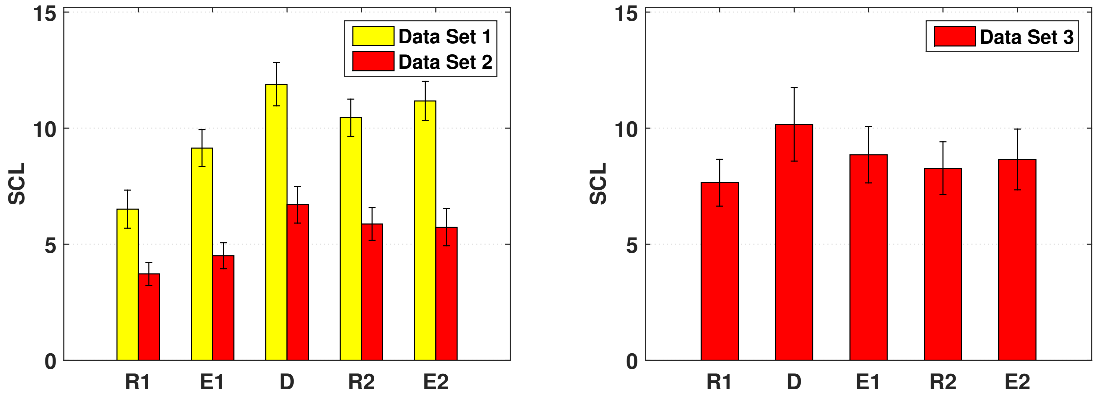
(d)

Figure 4.4: **The average values of the RMSSD and pNN50 extracted from BVP for each segment of the game across all subjects in Data Sets 1–3.**
in Data Sets 1–2, RMSSD (top-left) and pNN50 (bottom-left) decrease from R1 to D when the subjects play and further overshoot in R1 when subjects are at rest. Similarly, in Data Set 3, RMSSD (top-right) and pNN50 (bottom-right) decrease from R1 to D, However overshoot in E1.

The way we fuse the sensory information together is simple yet practical, and does not need any black box machine learning algorithms:

4.3.4.2 1 - Normalization

Since SC and BVP signals are highly dependent on each subject [107], normalization is applied to remove this individual dependency, as is common practice in the literature [82, 29, 122].



(a)

(b)

Figure 4.5: **The average values of the mean SCL extracted from SC for each segment of the game across all subjects in Data Sets 1–3.**
In Data Sets 1–2 (left), the mean SCL increases in a linear fashion from R1 to D, and decreases in E2. Similarly, in Data Set 3 (right), the mean SCL increases from R1 to D game and further decreases when the subjects play E1 and E2. In all Data Sets, mean SCL in D is the highest, on average, among all five segments of the game

This is performed following the standard formula,

$$F_{norm} = \frac{F - min(F)}{max(F) - min(F)} \tag{4.1}$$

where $min(F)$ and $max(F)$ are the minimum and maximum of a metric across all game segments, and F_{norm} is the normalized metric value where its minimum is 0, and maximum is 1.

4.3.4.3 Step 2 - Calibration with Data Set 1

We next use subjects' consistency ratios 77%, 92%, 77%, and 85% of the metric values (mean HR, RMSSD, pNN50, and mean SCL) *only* in Data Set 1, see Table 4.2. These are the ratios describing the $\%$ of subjects where a particular metric presented consistency with our hypothesis. This then yields a composite formula for the "Combined Metric Score (CMS)"

$$\begin{gathered}\text{Combined Metric Score}_{i,j} = 1/4\Big[(1 - \text{normalized mean HR}_{i,j}) \times 0.77 + \\ \text{normalized RMSSD}_{i,j} \times 0.92 + \text{normalized pNN50}_{i,j} \times 0.77 + \\ (1 - \text{normalized mean SCL}_{i,j}) \times 0.85\Big]\end{gathered} \tag{4.2}$$

where consistency ratios multiply associated metrics as weighting coefficients, j is the game segment of the experiment, $j = 1, 2, \ldots 5$, including the rest periods, and i is the subject number, $i = 1, \ldots 13$. Since the total number of normalized features used is four, the scaling is by four in (4.2). With this, the combined metric value in (4.2) is a value between 0 and 1. Moreover, since the mean HR and mean SCL metric values are found to increase with game difficulty, these metrics are subtracted from 1, to have all the features consistently represented. Consequently, in the way we setup (4.2), the lowest value is for Difficult (ideally zero) and the highest value is for R1 or R2 segments (ideally 1).

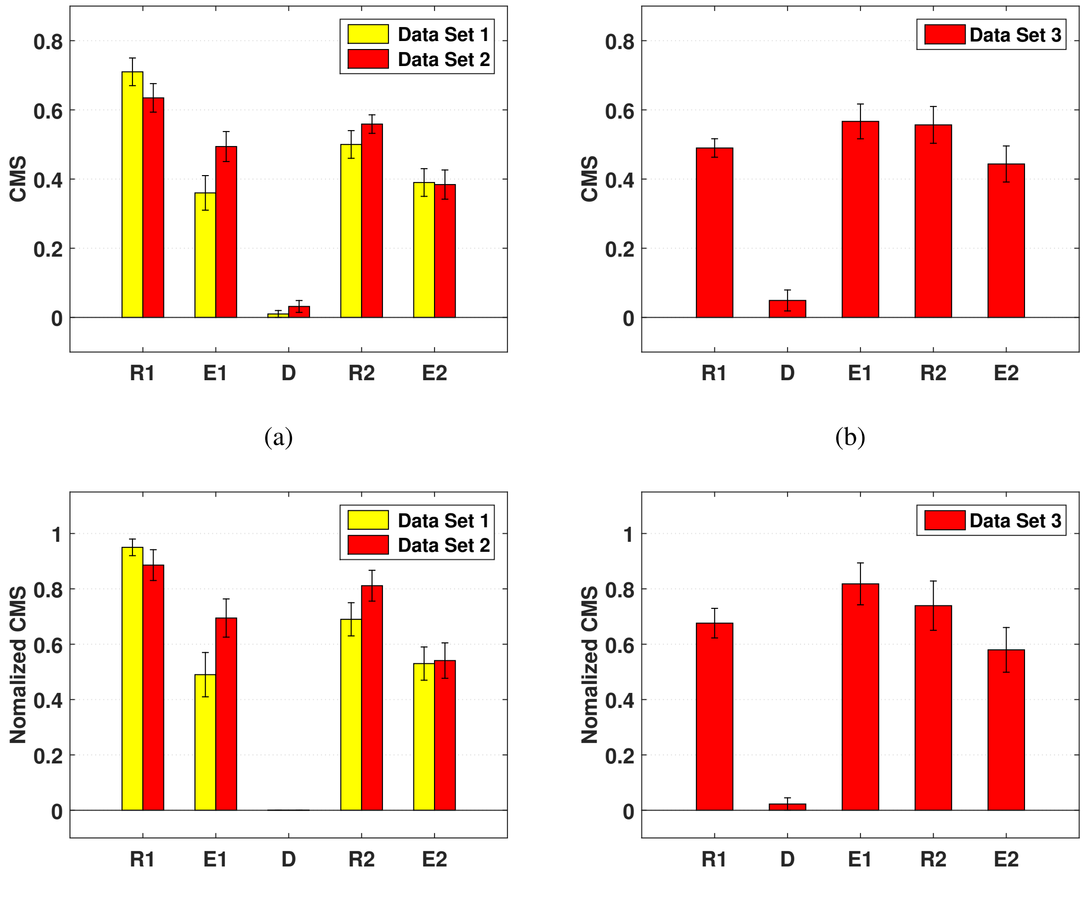
	Data Set 1					Data Set 2					Data Set 3				
Subject #	R1	E1	D	R2	E2	R1	E1	D	R2	E2	R1	D	E1	R2	E2
1	0.83	0.17	0.00	0.46	0.35	0.73	0.62	0.00	0.55	0.23	0.46	0.10	0.52	0.28	0.17
2	0.83	0.38	0.05	0.53	0.15	0.78	0.51	0.00	0.48	0.31	0.37	0.37	0.65	0.26	0.33
3	0.54	0.61	0.00	0.63	0.38	0.36	0.52	0.20	0.63	0.53	0.49	0.00	0.61	0.68	0.54
4	0.58	0.36	0.00	0.65	0.31	0.51	0.48	0.00	0.54	0.23	0.47	0.06	0.72	0.41	0.39
5	0.79	0.32	0.00	0.65	0.37	0.62	0.37	0.00	0.65	0.30	0.57	0.00	0.63	0.72	0.45
6	0.83	0.27	0.00	0.48	0.34	0.57	0.22	0.05	0.66	0.34	0.44	0.03	0.35	0.61	0.31
7	0.82	0.22	0.03	0.61	0.37	0.71	0.35	0.08	0.57	0.48	0.53	0.00	0.41	0.79	0.34
8	0.45	0.10	0.08	0.68	0.22	0.83	0.48	0.04	0.35	0.33	0.59	0.00	0.62	0.40	0.58
9	0.69	0.72	0.00	0.23	0.51	0.80	0.37	0.00	0.55	0.34	0.49	0.00	0.75	0.59	0.40
10	0.69	0.50	0.00	0.42	0.60	0.63	0.57	0.05	0.50	0.64	0.55	0.00	0.62	0.80	0.73
11	0.72	0.10	0.00	0.45	0.63	0.62	0.70	0.00	0.68	0.25	0.30	0.05	0.73	0.59	0.31
12	0.60	0.42	0.02	0.36	0.46	0.46	0.73	0.00	0.53	0.62	0.62	0.02	0.17	0.56	0.78
13	0.83	0.50	0.00	0.37	0.39	–	–	–	–	–	–	–	–	–	–

Table 4.3: **List of Combined Metric Score (CMS) values calculated for each segment of the game across all the subjects in Data Sets 1–3.**
Data Set 1 is used for CMS model fitting (calibration), and it is then validated using Data Sets 2–3. CMS lies between 0 and 1 as it is calculated from the normalized feature values extracted from BVP and SC signal data using Eq. (4.2).

Using CMS formula on Data Set 1, we obtain the results on Table 4.3:Data Set 1, which show statistically significant differences between CMS in the D versus E1 and E2 game levels ($p < 0.05$), see also Tables 4.6–4.7. That is, CMS in D is significantly lower than that of CMS in E1 and E2 ($p < 0.05$), see also Figure 4.4a. This may not be surprising since the Data Set 1 was used to generate (4.2).

4.3.4.4 Step 3 - Validation on Data Sets 2–3

We now investigate the validity of CMS. Without any additional calibration, we calculate CMS using (4.2) but this time on Data Sets 2–3. Recall that these data sets encompass the subjects that already participated in Data Set 1, and a set of completely new subjects.



(a)

(b)

(c)

(d)

Figure 4.6: **The average values of actual and normalized combined metric score (CMS) in Data Sets 1–3.**

In Data Sets 1–2 (left column), and Data Set 3 (right column), on average, CMS value is the lowest for the difficult game. In Data Sets 1–2, the average of normalized CMS for the D game is 0.00 ± 0.00.

Across all the subjects in both data sets, consistent results are obtained where CMS is indeed the lowest for the D game, see Table 4.3:Data Set 2–3. We observe statistically significant differences between CMS values in Data Sets 2–3 ($p < 0.05$), see also Tables 4.6–4.7:Data Set 2–3. Following this, pair-wise comparisons reveal that the average of CMS for the difficult game is significantly lower than for both E1 and E2 ($p < 0.05$). Moreover, there are no statistically significant differences between CMS values for the E1 and E2 games (see Tables 4.6–4.7) which are again consistent with the statistical results obtained for CMS calibrated by only Data Set 1.

Data Set 1			Data Set 2			Data Set 3		
		normalized			normalized			normalized
segment	CMS	CMS	segment	CMS	CMS	segment	CMS	CMS
R1	0.71 ± 0.13	0.95 ± 0.11	R1	0.63 ± 0.14	0.89 ± 0.19	R1	0.49 ± 0.09	0.68 ± 0.19
E1	0.36 ± 0.19	0.49 ± 0.30	E1	0.49 ± 0.15	0.69 ± 0.24	D	0.05 ± 0.10	0.02 ± 0.08
D	0.05 ± 0.02	0.00 ± 0.00	D	0.03 ± 0.06	0.00 ± 0.00	E1	0.57 ± 0.17	0.82 ± 0.26
R2	0.50 ± 0.14	0.69 ± 0.21	R2	0.56 ± 0.09	0.81 ± 0.19	R2	0.56 ± 0.18	0.74 ± 0.31
E2	0.39 ± 0.13	0.53 ± 0.22	E2	0.38 ± 0.15	0.54 ± 0.22	E2	0.44 ± 0.18	0.58 ± 0.28

Table 4.4: **The average values of actual and normalized combined Metric Score (CMS) across all subjects in Data Sets 1–3.**

The combined score values listed here are calculated using Eq. (4.2).

Metric	D and E1			D and E2			E1 and E2		
	Data Set 1	Data Set 2	Data Set 3	Data Set 1	Data Set 2	Data Set 3	Data Set 1	Data Set 2	Data Set 3
NASA-TLX	🟩 ✓	■	🟩 ✓	🟩 ✓	■	🟩 ✓	🟥 ✓	■	🟩 ✗
Performance	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟥 ✓	🟥 ✓	🟥 ✓
mean HR	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✗	🟥 ✓	🟥 ✓
RMSSD	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟥 ✗	🟥 ✓	🟥 ✓	🟩 ✗
pNN50	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟥 ✓	🟥 ✓	🟥 ✓
mean SCL	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✗	🟩 ✗	🟥 ✓
CMS	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟩 ✓	🟥 ✓	🟥 ✓	🟥 ✓

Table 4.5: **Statistical Comparison.**

The “Green background” color represents a statistically significant difference, otherwise it is represented by the “red background” color.

“✓” symbol means that the statistical result is inline with our hypothesis, otherwise it is represented by “✗” symbol.

*: NASA-TLX is not conducted in the Data Set 2.

4.4 Discussions

4.4.1 NASA-TLX Questionnaire

Statistically significant results on NASA-TLX mental workload scores suggest that the difficult level of our AT game indeed required much more mental workload compared with the easy levels of the game in all data sets. Moreover, analysis of the six individual dimensions of NASA-TLX revealed that playing the difficult game has much higher mental demand as well as performance and temporal demand dimensions on subjects, compared with the easy game. This is supporting evidence of our desire to design a game that could create good balance between mental workload, game speed, and performance. Furthermore, observing significant difference in NASA-TLX mental workload scores between E1 and E2 in Data Set 3 can be an indication of residual effects of playing E1 immediately after D level in Data Set 3. We will use this important information in the following discussions.

4.4.2 Task Performance Metric

Consistent statistical results in Data Sets 1–3 on the performance metric indicate that the subjects are challenged by the complexity of the difficult game, and subjects show consistent playing in identical E1 and E2 games, regardless of the order of the game levels. As their performance in the easy games follows similar patterns, subjects can be considered to be proficient with the easy game level. Notice that among all the subjects in Data Sets 1–3, only one subject in Data Set 1 failed to assign any of the airplanes in E2. While we think this could be related to a technical problem in the data acquisition system, as we could not confirm the reasoning of such incident, Subject 1 data is still included in statistical analysis. Furthermore, we verified that all subjects could assign successfully an airplane in the difficult game level, suggesting that pre-experimental briefing was effective. Finally, we note that, the obtained results also have strong correlation with the NASA-TLX questionnaire scores, which again demonstrates that subjects on average experience the highest mental workload while playing the difficult game versus the two identical easy game levels.

4.4.3 BVP and SC Features

All BVP and SC metrics (mean HR, RMSSD, pNN50 and mean SCL) are significantly affected by the cognitive workload changes in different levels of the game [101, 89, 53, 24], which we mainly attribute to increases in subjects' mental workload, induced by their lack of experience with the difficult game. This was clearly captured by mean HR (Figure 4.3) and mean SCL (Figure 4.5), across all subjects in Data Sets 1-3, being the "largest," and RMSSD and pNN50 (Figure 4.4) being the "lowest" in the difficult game than in the other game segments, indicating, on average, that the subjects as a group reflect their arousal and mental workload increases through these metrics [101, 89]. With support of NASA-TLX results and performance metric calculations, one could therefore argue that the subjects as a group allocated higher level of mental effort on average when they played the difficult game, which shows that the subjects as a group present some consistency in their physiological responses, as predicted by these metrics.

However a few inconsistencies is observed when we compare mean HR and mean SCL values in the two identical easy, E1 and E2 games: (*a*) in Data Set 1, mean HR for E1, on average is significantly higher than E2 (Figure 4.3a). One possible explanation of this was that HR in E2 is still influenced by the preceding difficult game even if there is a rest period, R2, making it statistically different from E1 [101]. This is not the case in Data Sets 2–3 where the effect of cognitive workload on mean HR in the two identical E1 and E2 is not significantly different (Figs. 4.3a–4.3b). In Data Set 2, this could be partly attributed to the same subjects with prior experience in Data Set 1 being less affected, and in Data Set 3, E1 and E2 are separated by a rest period instead of D in between them in Data Set 1. An observation we extract from these discussions is that the rest period of 60 sec. may not be long enough for HR to regulate, and whenever a difficult game precedes the rest period, the effect of the difficult game rolls over even after the rest period. (*b*) In Data Set 1, mean SCL for E1, on average is significantly lower than E2 (Figure 4.5a). We believe that this is due to the fact that the SCL measurements are in general affected by increased emotional arousal, specifically in the difficult game, and remain in effect for a while even if the subject enters a rest phase. In other words, SCL measurements have a type of "memory" effect. Similarly, we do not observe such cases in Data Sets 2–3 possibly supported by similar reasoning we provided above regarding mean HR.

Statistical analysis on the above mentioned physiological metrics point out three key messages: (*i*) Since mean HR and mean SCL in D are larger, and RMSSD and pNN50 for D are smaller than those for E1 and E2, this is consistent with the fact that subjects found the difficult game much more difficult than the games E1 and E2, as assessed by NASA-TLX and performance metric. This is also consistent with what is observed in our preliminary results [101]. If we consider only the first two levels of the game, D and E1, we can conclude that subjects' perception of game difficulty as manifested by their inexperience is correlated with all the physiological metrics studied in this work. Further, after the relaxed state R1, the HR and SCL metrics can successfully differentiate the difficult game level D from the easy game level E1, *regardless of the game order.* (*ii*) When all levels of the game are considered, we reveal that the game order has a *strong* nonlinear effect on RMSSD and pNN50 metrics. For example, the relaxation from D to R2 in Data Sets 1–2, or to E1 in Data Set 3 is very pronounced, indicating it is less critical what follows D. This relaxation seems to also present a type of overshoot, as observed from D to E1 in Data Set 3, which then settles in R2 and E2 (Figs. 4.4). This overshoot is visible also in Data Sets 1–2 but may not be clearly noticeable since this time game level D is followed first by R2. This is valuable insight into how RMSSD and pNN50 interplay with game order, which, to the best of our knowledge, has not been studied through subsequent game levels. Therefore, one can use RMSSD and pNN50 metrics to detect subject' leaving the difficult game, regardless of what the subsequent game is, whether an easy game or a rest period. (*iii*) Finally, if E1 and R2 periods were longer, we think that the physiological metrics studied in Data Set 3 would have sufficient time to settle and ultimately be similar to those in Data Set 1–2, but this needs to be confirmed in future experiments.

4.4.4 Combined Metric Score (CMS)

Figure 4.6 and Table 4.4 indicate that on average and across all the subjects in the three Data Sets 1–3, CMS calculated using (4.2) for each segment of the game is the smallest for the difficult game, appearing to be consistent with our expectations. In Data Sets 1–2, the values for E1 and E2 lie between those of difficult and R1/R2; and moreover R1/R2 segments take the largest value, showing consistency, Figure 4.6a. In Data Set 3 however, the strong impact of playing E1 immediately after playing the D game is observable, recalling a type of overshoot described above, where the CMS for E1 is the highest, and further relaxes toward E2, Figure 4.6b.

The above results suggest that CMS proves to be a valid metric for the same subjects in Data Set 2 (except one missing) and with a set of completely different 12 subjects (Data Set 3) where it still reliably predicts mental workload increases as triggered mainly by subjects' inexperience in the difficult game. It also shows that on average, and in support of NASA-TLX results (Data Set 1, Data Set 3) as well as performance metric calculations, subjects' experience, *regardless of the game order*, the highest level of mental workload when they play a challenging game with which they have insufficient experience, supporting our hypothesis; and the lowest level of mental workload when they did not play the game and remained at rest, as was expected and is consistent in all three studies. Moreover, even if the subjects have no exposure to the difficult game level in Data Set 1, and are experienced with both game difficulty levels in Data Set 2, we obtain similar and consistent results, again owing to subjects' lack of training and insufficient exposure to the difficult game.

	Feature	Easy 1			Difficult			Easy 2		
		Mean	SD	SEM	Mean	SD	SEM	Mean	SD	SEM
Data Set 1	NASA-TLX	6.34	2.37	0.66	74.81	11.83	3.28	7.05	2.63	0.73
	Performance	18.69	0.48	0.13	6.85	3.85	1.07	16.54	5.32	1.48
	mean HR	83.14	12.08	3.35	88.00	14.14	3.92	77.78	11.65	3.23
	RMSSD	34.68	18.13	5.03	22.74	10.41	2.89	36.35	18.49	5.13
	pNN50	13.73	10.90	3.02	5.65	9.04	2.51	19.76	14.57	4.04
	mean SCL	9.14	2.86	0.79	11.89	3.36	0.93	11.17	3.05	0.85
	Combined Score	0.36	0.19	0.05	0.05	0.02	0.05	0.39	0.14	0.04
Data Set 2	Performance	18.50	0.90	0.26	8.08	3.03	0.87	18.42	0.51	0.15
	mean HR	70.78	9.93	2.87	78.65	11.00	3.17	69.98	9.05	2.61
	RMSSD	47.01	12.39	3.58	35.56	15.15	4.37	44.42	16.67	4.81
	pNN50	22.05	10.75	3.10	15.56	15.14	4.37	33.28	19.45	5.62
	mean SCL	4.50	1.96	0.56	6.70	2.75	0.79	5.73	2.76	0.80
	Combined Score	0.49	0.15	0.04	0.03	0.06	0.02	0.38	0.15	0.04
Data Set 3	NASA-TLX	10.48	5.91	1.71	76.08	7.45	2.15	5.94	2.90	0.84
	Performance	18.83	0.39	0.11	7.00	2.89	0.83	18.75	0.45	0.13
	mean HR	79.31	11.43	3.30	91.67	11.36	3.28	78.45	11.38	3.28
	RMSSD	44.52	23.15	6.68	27.08	12.10	3.49	31.72	13.15	3.79
	pNN50	21.06	15.96	4.61	6.44	6.21	1.79	14.30	13.49	3.89
	mean SCL	8.85	4.19	1.21	10.16	5.46	1.58	8.65	4.53	1.31
	Combined Score	0.57	0.17	0.05	0.05	0.10	0.03	0.44	0.18	0.05

Table 4.6: **Mean, standard deviation (SD), and standard error of the mean (SEM) for feature values in different games across all the subjects in Data Sets 1–3.**

The descriptive statistics presented are from the calculated normalized values for mean HR, RMSSD and pNN50 from BVP and mean SCL from SC signals in each game. The combined metric score is calculated using (4.2).

	Metric	All Games			Paired Samples Test					
					D vs. E1		D vs. E2		E1 vs. E2	
		F	sig.	Observed Power	Mean Difference	sig.	Mean Difference	sig.	Mean Difference	sig.
Data Set 1	NASA-TLX†	$F(1.043, 12.518) = 393.004$	$p < 0.05^*$	0.999	68.469	$p < 0.05^{**}$	67.762	$p < 0.05^{**}$	−0.708	$p = 0.308$
	Performance	$F(2, 11) = 58.149$	$p < 0.05^*$	0.999	−11.846	$p < 0.05^{**}$	−9.692	$p < 0.05^{**}$	2.154	$p = 0.178$
	mean HR	$F(2, 11) = 14.302$	$p < 0.05^*$	0.999	4.862	$p < 0.05^{**}$	7.670	$p < 0.05^{**}$	5.353	$p < 0.05^{**}$
	RMSSD†	$F(2.624, 31.491) = 10.801$	$p < 0.05^*$	0.995	−11.946	$p < 0.05^{**}$	−13.615	$p < 0.05^{**}$	−1.669	$p = 0.345$
	pNN50	$F(2, 11) = 7.422$	$p < 0.05^*$	0.935	−8.080	$p < 0.05^{**}$	−10.692	$p < 0.05^{**}$	−2.612	$p = 0.130$
	mean SCL†	$F(1.285, 15.420) = 181.161$	$p < 0.05^*$	0.999	2.755	$p < 0.05^{**}$	.722	$p < 0.05^{**}$	−2.032	$p < 0.05^{**}$
	CMS†	$F(2.668, 32.021) = 39.981$	$p < 0.05^*$	0.999	−0.348	$p < 0.05^{**}$	−0.380	$p < 0.05^{**}$	−0.032	$p = 0.595$
Data Set 2	Performance†	$F(1.110, 12.208) = 115.288$	$p < 0.05^*$	0.999	−10.417	$p < 0.05^{**}$	−10.333	$p < 0.05^{**}$	0.083	$p = 0.754$
	mean HR	$F(2, 10) = 23.481$	$p < 0.05^*$	0.999	7.880	$p < 0.05^{**}$	8.678	$p < 0.05^{**}$	0.798	$p = 0.474$
	RMSSD	$F(2, 10) = 6.709$	$p < 0.05^*$	0.887	−11.456	$p < 0.05^{**}$	−8.858	$p < 0.05^{**}$	2.597	$p = 0.261$
	pNN50	$F(2, 10) = 10.667$	$p < 0.05^*$	0.982	−6.492	$p = 0.037^{**}$	−6.575	$p < 0.05^{**}$	−0.08	$p = 0.973$
	mean SCL†	$F(1.464, 16.105) = 25.211$	$p < 0.05^*$	0.999	2.199	$p < 0.05^{**}$	0.968	$p < 0.05^{**}$	−1.231	$p < 0.05^{**}$
	CMS	$F(2, 10) = 126.390$	$p < 0.05^*$	0.999	−0.462	$p < 0.05^{**}$	−0.352	$p < 0.05^{**}$	0.110	$p = 0.068$
Data Set 3	NASA-TLX	$F(2, 10) = 787.828$	$p < 0.05^*$	0.999	65.608	$p < 0.05^{**}$	70.139	$p < 0.05^{**}$	4.531	$p = 0.057^{**}$
	Performance†	$F(1.013, 11.144) = 179.213$	$p < 0.05^*$	0.999	−11.833	$p < 0.05^{**}$	−11.750	$p < 0.05^{**}$	0.083	$p = 0.0.339$
	mean HR†	$F(2.061, 22.673) = 25.143$	$p < 0.05^*$	0.999	12.358	$p < 0.05^{**}$	13.217	$p < 0.05^{**}$	0.858	$p = 0.360$
	RMSSD†	$F(1.923, 21.148) = 5.278$	$p < 0.05^*$	0.767	−17.442	$p < 0.05^{**}$	−4.636	$p = 0.155$	12.806	$p = 0.033^{**}$
	pNN50†	$F(2.866, 31.527) = 8.238$	$p < 0.05^*$	0.981	−14.617	$p < 0.05^{**}$	−7.858	$p < 0.05^{**}$	6.758	$p = 0.076$
	mean SCL†	$F(1.550, 17.051) = 11.841$	$p < 0.05^*$	0.971	1.308	$p < 0.05^{**}$	1.510	$p < 0.05^{**}$	0.202	$p = 0.373$
	CMS	$F(2, 10) = 27.806$	$p < 0.05^*$	0.999	−0.516	$p < 0.05^{**}$	−0.395	$p < 0.05^{**}$	0.121	$p = 0.169$

Table 4.7: **Statistical differentiation of different levels of difficulty using various features in Data Sets 1–3.**

The $\alpha = 0.05$ level is chosen for statistical analysis on metrics extracted from BVP and SC signals.

* indicates that significant difference is found among all three groups feature values. ** indicates that the average difference is significant in pair wise analysis. NASA-TLX is not conducted in the Data Set 2.

4.5 Summary

In this chapter, we study from an affective computing point of view, whether or not subjects' inexperience in an experimental task can manifest itself as obvious variations in physiological measurements obtained from HR and SC sensors.

Although the number of subjects in this study is at the lower limit of running statistical analysis which might add some uncertainty or variability in statistical analysis results, the power of the study seems to be quite favorable, supporting the findings, and calling for future studies in expanded populations. Specifically, we find out that the metrics derived from sensor measurements present consistency, and hence, through affective computing, subjects' inexperience in a challenging task via the ensuing mental workload changes is detectable. The results also show strong correlation with NASA-TLX questionnaire, and subjects' overall performance index. Moreover, the findings suggest that different levels of subjects' overall performance are directly correlated with the physiological data collected. In addition, we find out that lack of experience in the difficult game produced remarkably different physiological responses in all data sets, which were also associated with performance.

Having confirmed the validity of the metrics in all data sets at hand (Data Sets 1–3), Data Set 1 was then used to calibrate the metrics under a combined metric score (CMS). By using the data obtained in Data Sets 2–3, CMS is then put to test in order to study its validity. We find out that CMS still yields consistent and reliable results with the new data sets. Hence, CMS offers the potential to be used in future studies as a single scalar quantity that could be used to make predictions on subjects' inexperience and/or what difficulty levels the subjects are encountering while playing various game levels. These reverse scientific questions, including how to infer inexperience through measurements remain as open problems to be studied in the future, all of which have the potential to pave the way toward adaptive real time assistance schemes in human-machine systems.

Chapter 5

DEVELOPMENT OF A COMBINED TIME-FREQUENCY TECHNIQUE FOR ACCURATE EXTRACTION OF PNN50 METRIC FROM NOISY HEART RATE MEASUREMENT

5.1 Introduction

It is well known that environmental stressors might trigger a *fight or flight* response in humans. Such a mechanism, originated in the brain, passes arousal/stress messages to the nervous and endocrine system. Some of these messages activate the sympathetic branch of the autonomic nervous system (ANS), resulting in, among others, the following physiological responses: pupils dilation, salivation inhibition, bronchial relaxation, adrenaline secretion, gastrointestinal activity reduction, and cardiac activity modulation. This chapter studies how the cardiac activity modulation can be used as an indirect assessment of the ANS activity, which is intimately related to the arousal and stress level of human subjects [19].

Cardiac activity modulation, governed by the systolic and diastolic cycles, can be measured by utilizing standard heart rate (HR) sensors, such as electrocardiogram (ECG), photoplethysmography (PPG) [3] or blood volume pulse (BVP); and it can be quantified through heart-rate variability (HRV) analysis through the study of how instantaneous HR period varies with respect to time. HRV has been broadly utilized to assess the affective states of humans [104, 60, 61], including changes in mental workload [57, 90, 89] and stress levels [122, 50, 99, 123].

Traditional approaches compute the HRV in either time domain or frequency domain. Both approaches are built upon the inter-beat interval (IBI) time series (see Figure 2.3), which is constructed from the measured HR data by *(a)* first detecting the systolic peaks, and *(b)* then calculating the time differences between all consecutive peaks. In time domain, the IBI data is used to extract various HRV metrics. One of them is the so-called pNN50, which is defined as the probability that the time difference between consecutive IBI samples is greater than 50 ms. The pNN50 has been used not only to analyze the condition of the cardiovascular system [78], but also to assess stress level and mental workload changes in human subjects [123, 122]. In frequency domain, the Fourier transform of the IBI data can also be used to extract various HRV metrics, which have been used to assess human stress [122, 50] and mental workload [115, 126]. One of these metrics is based on the ratio between the low frequency (LF) power (0.04Hz-0.15Hz) and high frequency (HF) power (0.15Hz-0.4Hz). It has been reported that an increased LF/HF power ratio indicates a dominance of the sympathetic branch of the ANS, while its reduction is associated with a dominance of the parasympathetic branch of the ANS [19].

The efficacy of the aforementioned HRV techniques, both in time and frequency domain, can be limited when used in a realistic clinical setting. For instance, data acquisition instabilities and motion in HR sensors, like BVP, may degrade the signal to noise ratio in the measured data; and this may ultimately result in miss-detection of the heart beat peaks, causing errors in the IBI time series, and therefore inaccurate estimation of the HRV-based metrics [10, 112] in both time and frequency domain. Several techniques have been developed for reliable extraction of the IBI, which include the following: 1) restricting the human motion in a controlled laboratory setting; 2) smoothing the data with a low-pass filter [81], 3) detecting heartbeat peaks and correcting outliers in the time series via postprocessing [73, 77, 39, 9] and real-time [112] algorithms.

Nevertheless, it has been shown that analyzing signals embedded in noise can be more reliably done using combined 2D time-frequency processing techniques than using 1D time or frequency methods [23, 21]. The main reason behind this improvement is that the energy of the undesired random noise is typically distributed over the whole time-frequency domain, whereas the signal of interest will only concentrate its energy within limited time intervals and/or frequency bands. Moreover, time-frequency techniques provide a suitable framework to analyze transient signals; and, therefore, they can be adapted for the study of inherently-transient HR data. While to the best of our knowledge such an approach on the HR data and pNN50 metric has not been presented in the literature, a closely related work in this context can be found in [80][Chapter 4.4] where the time varying instantaneous frequency is extracted using a time-frequency approach.

In this chapter, we present *(a)* a customized Short Time Fourier Transform (STFT) time-frequency technique to analyze HR data highly corrupted by noise; and *(b)* a novel mathematical formulation to extract the pNN50 metric from noisy HR data in the combined time-frequency domain. By leveraging on the aforementioned advantages of the time-frequency analysis, the proposed method is capable of enhancing the accuracy and robustness of the pNN50 metric extracted from the noise-corrupted HR data. One of the main reasons why this successful outcome is achieved is because the proposed method does not rely on the calculation of the IBI data, which often presents errors when computed with an automatic peak detection algorithm from noisy HR data. Moreover, the proposed approach does not require the use of any filtering, visual evaluation of IBI data and correction of its outliers or any prior knowledge about the noise distribution.

The chapter is organized as follows: in Section 5.2.1, we review the conventional time domain formulation to compute the pNN50 metric, and we describe the proposed mathematical framework needed to extract the pNN50 from the HR data in the combined time-frequency domain, Section 5.2.2; in Section 5.4, we evaluate the performance of the proposed method on a noisy synthetic analytic signal, Section 5.4.2 and on a realistic BVP signal, Section 5.3.3. Results demonstrate that the proposed time-frequency domain approach outperforms conventional time domain methodology in the accuracy of the extracted pNN50 metric, thus providing a better assessment tool to characterize stress levels and mental workload changes in human subjects [123, 122].

5.2 Methods

5.2.1 pNN50 calculation in time domain

The pNN50 is computed by observing temporal changes in the normal-to-normal (NN) intervals, see Figure 2.3. In this figure, a generic PPG signal is depicted with several pressure peaks generated during the systolic and diastolic phases of the cardiac cycle [45]. Let us denote t_k, $k = 1, 2...n_t$, the time instant at which the systolic peak is measured. Then, the NN_k interval at $t = t_k$, $k = 1, 2, \ldots, n_t - 1$ with n_t being the number of systolic peaks, is given by the following equation:

$$NN_k = t_{k+1} - t_k. \tag{5.1}$$

Since the value NN_k captures the time difference between two consecutive peaks in a quasi-periodic signal, it can be considered to be an approximation of the instantaneous period of the signal, T_k^i, at every instant of time t_k. This value can be written as follows:

$$NN_k \approx T_k^i. \tag{5.2}$$

The amount of HRV can be quantified by measuring the variation of the instantaneous period of the PPG signal at every time t_k. This variation, ΔT_k^i, can be approximated as the difference between two consecutive intervals, ΔNN_k, when equation (5.2) is used. That is,

$$\Delta T_k^i \approx \Delta NN_k = NN_{k+1} - NN_k. \tag{5.3}$$

The pNN50 is defined as the probability that ΔNN_k, for $k = 1, 2 \ldots n_t - 2$, is greater than 50 ms, where $n_t - 2$ is the total number of intervals differences. This probability can be extracted using the following estimator [78]:

$$pNN50 \approx \frac{\#NN50}{\#NN}, \tag{5.4}$$

where $\#NN50$ is the number of interval differences ΔNN_k, $k = 1, 2 ... n_t - 2$, greater than 50 ms; and $\#NN = n_t - 2$.

5.2.2 Proposed pNN50 formulation in time-frequency domain

The time domain method described above may incorrectly estimate the instantaneous period and the pNN50 statistic when the PPG data is corrupted by certain amount of noise. This problem mainly occurs because the detection of the systolic peaks cannot be reliably done in the presence of noise. As discussed in the Introduction, the Short-Time Fourier Transform (STFT), which is a combined time-frequency domain method, can reduce the impact of noise in the signal of interest. Hence, it is a suitable tool to reliably calculate the pNN50 metric from noisy PPG data.

The proposed method to compute the pNN50 using the STFT is divided into six steps: 1) Extracting the average IBI of the PPG signal; 2) Computing the STFT of the PPG signal using a sliding window that is customized by the averaged IBI; 3) Calculating the power spectrogram from the STFT; 4) Extracting the instantaneous frequency from the spectrogram; 5) Computing the variation in the NN intervals from the instantaneous frequency; 6) Calculating the pNN50 from the variations in the NN intervals. The implementation of the algorithm is carefully described in the following subsections.

5.2.2.1 Step 1 - Extracting the average IBI

The first step in the proposed algorithm is to estimate the average IBI, denoted by $\widehat{IBI}$, of the PPG signal. The value of the estimated IBI is equivalent to the reciprocal of Heart Rate (HR) estimate, $\widehat{HR}$, and hence it can be expressed as the averaged frequency of the heart rate, $\hat{f}_{HR}$, as

$$\widehat{IBI} = \frac{1}{\widehat{HR}} = \frac{1}{\hat{f}_{HR}}. \tag{5.5}$$

The value of $\hat{f}_{HR}$ is obtained as a weighted average of the dominant components of the PPG signal. For this, a value is defined, 0.8, as the threshold above which the components in the normalized Fourier transform are selected [100]. These components are located around the fundamental frequency, fmax, which is associated with HR (Figure 5.1). That is,

$$\hat{f}_{HR} = \beta_1 \int_{f_l}^{f_u} f \; |X\left(f\right)| \, df, \tag{5.6}$$

where f_l and f_u are the lower and upper bounds for the integration variable f; $|X\left(f\right)|$ is the magnitude of the Fourier Transform of $x(t)$, acting as weighted function; and β_1 is a normalizing constant given by $\beta_1 = 1/\int_{f_l}^{f_u} |X\left(f\right)| \, df$.

5.2.2.2 Step 2 - Computing the STFT of the PPG signal

The second step in the algorithm is to compute the STFT of the PPG signal. The STFT $X\left(t, f\right)$ of the signal of interest $x(t)$ is given by

$$X\left(t, f\right) = \int_{-\infty}^{\infty} x(\tau) w(\tau - t; a) e^{-j2\pi f \tau} d\tau, \tag{5.7}$$

where $w(t; a)$ is a sliding window, of fixed size a seconds, used to extract the spectral components $x(t)$ in different instants of time.

There are three time-domain input parameters to the STFT algorithm: 1) the window size a; 2) the sampling rate of $X\left(t, f\right)$, which is defined by the window overlap region for two adjacent samples; and 3) the shape of the STFT window $w(t; a)$. These three parameters must be carefully chosen to accurately obtain the pNN50 in the time-frequency domain.

The choice of the window size a is a trade-off between time and frequency resolution. A short window can efficiently suppress the signal outside the window and give a localized measure of the spectral density, but the frequency resolution may not be sufficient. A long window can produce a rich spectral transformation, but the time resolution may be poor due to the large number

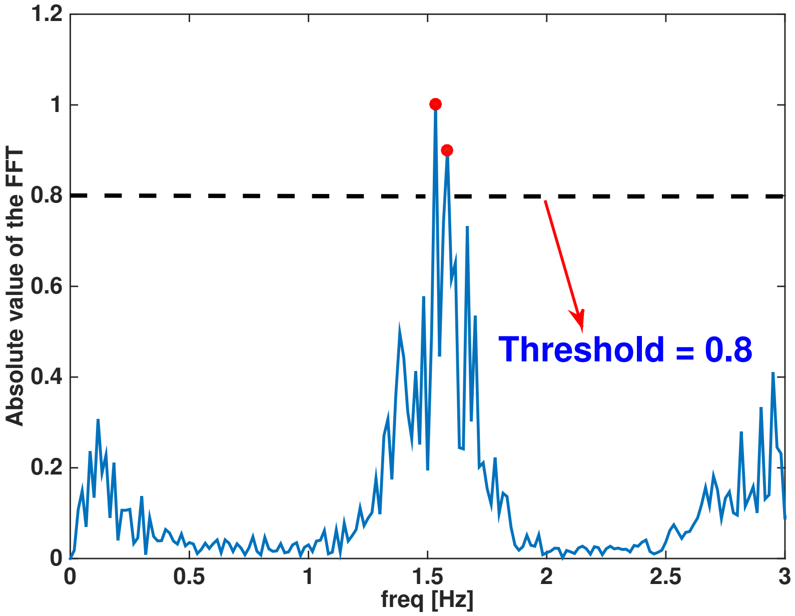


Figure 5.1: **FFT of the BVP measurements.**
The dominant components around $f_{max} = 1.53$ Hz, (i.e. corresponding to the *red dots*) are used to obtain HR for the specific signal data. With threshold $= 0.8$ (dotted line), the frequency 1.58 Hz is located around f_{max}.

of averaged samples. In this study, the window size has been selected to be fixed at $a = 2 \times \widehat{IBI}$. This is the minimum size of the window in the time domain that is required to accurately compute the pNN50, since at least three peaks of the PPG signal are required for this purpose. The window $w(t; a)$ slides along the time axis to calculate the spectral components at different time instants. Consecutive windows are overlapped to ensure the continuity of the resultant signal.

The sampling rate of the STFT in time domain, R_t, depends on the size of the window, a, and the overlap between consecutive windows, ov; and it is given by $R_t = 1/(a - ov)$. The selected sampling rate is given by the inverse of the inter-beat interval: $R_t = 1/\widehat{IBI}$, therefore the overlap is equal to the averaged IBI: $ov = \widehat{IBI}$ seconds. In this way, the value of the sampling rate for the time-frequency domain method is the same as the resolution of the time domain method. The chosen sampling rate generates the sampled signal $X(t_k, f_j)$, for $k = 1, 2 ... n_t$ and $j = 1, 2 ... n_f$.

One additional advantage of computing the pNN50 from the time-frequency method is that it uses all temporal samples in the interval $a = 2 \times \widehat{IBI}$, while the time domain method only uses three discrete data points corresponding to the three local maxima in the same interval. Increasing the sampling rate is possible by increasing the overlap; however, this resolution enhancement in the spectrogram requires additional computational time.

5.2.2.3 Step 3 - Calculating the power spectrogram from the STFT

The spectrogram $P\{\cdot\}$ of the sampled signal $X\left(t_k, f_j\right)$ is calculated by

$$P\{X\left(t_k, f_j\right)\} = \left|X\left(t_k, f_j\right)\right|^2, \tag{5.8}$$

which will be used in the next step to extract the instantaneous frequency of the PPG signal.

5.2.2.4 Step 4 - Extracting the instantaneous frequency from the spectrogram

The instantaneous frequency f_k^i of the PPG signal, which is defined as the reciprocal of the instantaneous period T_k^i, is calculated from the spectrogram at every instant of time t_k as a weighted average of all the values of the spectrogram greater than a given threshold, K:

$$f_k^i = \beta_2 \sum_{j|P\{X(t_k,f_j)\}>K} f_j P\{X\left(t_k, f_j\right)\}, \tag{5.9}$$

where β_2 is also a normalizing constant given by $\beta_2 = 1/\left(\sum_{j|P\{X(t_k,f_j)\}>K} P\left\{X\left(t_k, f_j\right)\right\}\right)$.

5.2.2.5 Step 5 - Computing the variation in the NN intervals from the instantaneous frequency

The variation in the NN intervals, ΔNN_k, in the time-frequency domain is computed as follows:

$$\Delta NN_k \approx \Delta T_k^i = \left.\frac{\mathrm{d}T^i(t)}{\mathrm{d}t}\right|_{t=t_k} \Delta t_k, \tag{5.10}$$

where Δt_k is the time separation between peaks, given by $\Delta t_k = t_{k+1} - t_k = NN_k \approx T_k^i$; $T^i(t)$ is the instantaneous period at time t, which is the reciprocal of the instantaneous frequency $f^i(t)$; and the derivative of the instantaneous period at time $t = t_k$ is calculated by using the chain rule as:

$$\left.\frac{\mathrm{d}T^i(t)}{\mathrm{d}t}\right|_{t=t_k} = \frac{-1}{\left(f^i(t)\right)^2} \left.\frac{\mathrm{d}f^i(t)}{\mathrm{d}t}\right|_{t=t_k}. \tag{5.11}$$

Finally, taking into consideration that $T_k^i = 1/f_k^i$, the variation in the NN intervals, ΔNN_k, at time $t = t_k$ is given by:

$$\Delta NN_k = \frac{-1}{\left(f^i(t)\right)^2} \left.\frac{\mathrm{d}f^i(t)}{\mathrm{d}t}\right|_{t=t_k} \frac{1}{f_k^i}. \tag{5.12}$$

5.2.2.6 Step 6 - Calculating the pNN50 from the variations in the NN intervals

Using (5.12) it is possible to calculate the NN_k interval differences, and consequently the pNN50 given by (5.4), using the instantaneous frequency calculated in the time-frequency domain method.

5.3 Data Analysis

The proposed method has been applied to synthetic and experimental data. In order to verify the validity of this new approach, it was tested using signals with known instantaneous frequency. For these cases, and to quantify the error of the estimation given by the different methods, Root Mean Square (RMS) was used as figure of merit [68].

5.3.1 Test signals with no noise

5.3.1.1 Synthetic sinusoidal signal

The first test signal is a sinusoidal function $\cos(\phi(t))$. The relation between the argument, $\phi(t)$, of a sinusoidal signal and its instantaneous frequency is given by:

$$f^i(t) = \frac{1}{2\pi}\frac{\mathrm{d}\phi(t)}{\mathrm{d}t} \quad \Leftrightarrow \quad \phi(t) = 2\pi\int_0^t f^i(\tau)\mathrm{d}\tau. \tag{5.13}$$

The instantaneous frequency selected for the signal under test follows a sinusoidal variation, given by:

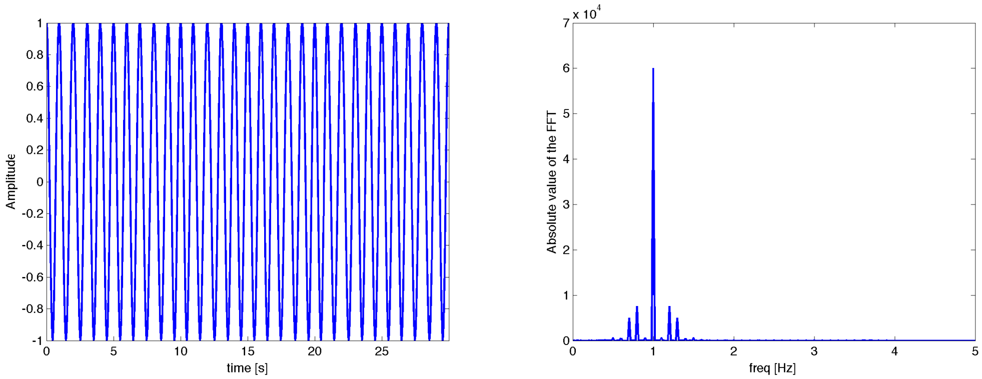
$$f^i(t) = f_0\left(1 + A_1\cos\left(2\pi f_1 t\right) + A_2\cos\left(2\pi f_2 t\right)\right); \tag{5.14}$$

and using (5.13), the argument of the signal under test is calculated as:

$$\phi(t) = 2\pi f_0 t + \frac{A_1 f_0 \sin(2\pi f_1 t)}{f_1} + \frac{A_2 f_0 \sin(2\pi f_2 t)}{f_2}, \tag{5.15}$$

where $f_0 = 1$ Hz is the main frequency of the cosine, $f_1 = 0.2$ Hz and $f_2 = 0.3$ Hz are the frequencies of the modulated signals, and $A_1 = A_2 = 0.05$ are their amplitudes. The resultant signal is similar to a pure cosine, as shown in Figure 5.2(a), but it has several spectral components due to its variable instantaneous frequency, as depicted in Figure 5.2(b).

Following the steps described in Section 5.2.2, $\widehat{IBI}$ is estimated following (5.5)-(5.6). Using $\widehat{IBI}$, the size of the window, a, and the overlap, ov, that defines the sample rate, R_t, of the STFT signal are defined. The value of each parameter is shown on Table 5.1 as Case 1.



(a) (b)

Figure 5.2: **Sinusoidal test signal.**
(a) time domain, and (b) frequency domain.

	Window size a [s]	Overlap between windows ov [s]	Rate of the STFT R_t [Hz]
Case 1 (Section 5.3.1.1)	2.0000	1.0000	1.0000
Case 2 (Section 5.3.1.2)	1.4688	0.7344	1.3617
Game Segment R1	1.9346	0.9673	1.0338
Game Segment E1	1.7266	0.8633	1.1584
Game Segment D	1.5791	0.7896	1.2665
Game Segment R2	2.0508	1.0254	0.9752
Game Segment E2	1.8174	0.9087	1.1005

Table 5.1: **Value of the parameters of the STFT for the different synthetic cases.**

In this work, three different types of windows are studied: rectangular, Hamming and Chebyshev. Figure 5.3(a-c) shows the spectrogram of the test signal for the different windowing cases. The rectangular window has a narrower main lobe, but the level of the secondary lobes is higher, as can be observed in Figure 5.3(a). Secondary lobes are much less pronounced in the case of Hamming and Chebyshev windows.

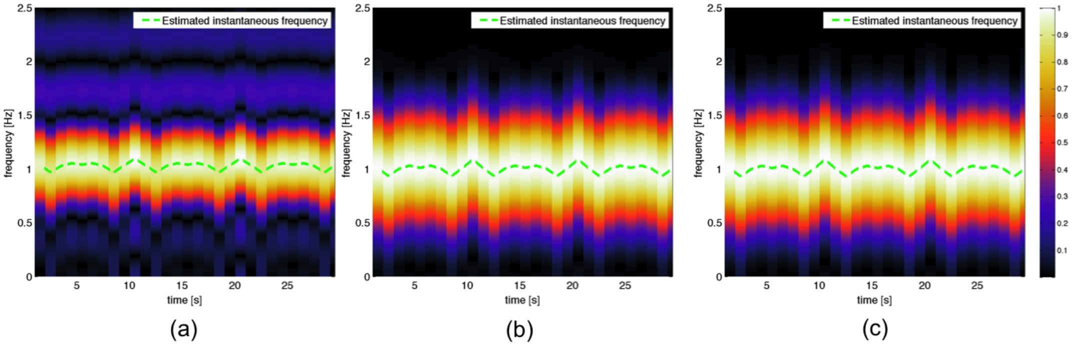


Figure 5.3: **Spectrogram of a cosine with variable instantaneous frequency.** It is using window of size $a = 2$ s, overlap between windows $ov = 1$ s and type of window (a) rectangular, (b) Hamming, and (c) Chebyshev.

5.3.1.2 Synthetic signal with several spectral components

The second test signal used was created as a linear combination of different spectral components with different amplitudes. Figure 5.4 shows the resulting signal in time domain and frequency domain.

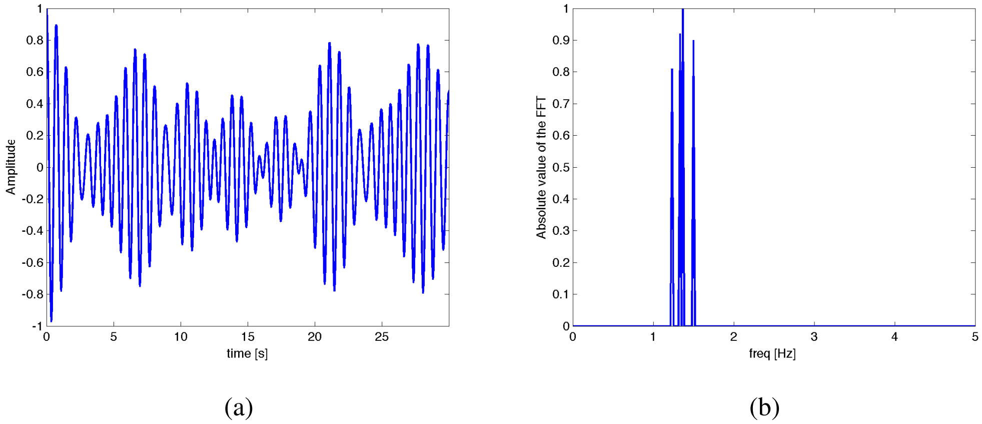


Figure 5.4: **Signal under test (a) time domain, and (b) frequency domain.**

In this case it is not possible to estimate the instantaneous frequency as in (5.13). Instead, the Hilbert transform was used to derive the analytic function of the signal, resulting in the ability to define the instantaneous frequency for every point on the function [59]. This was possible since the signal at hand is a narrowband signal[1]. The instantaneous frequency calculated using the Hilbert transform is used as *ground truth* in this section.

As described in the previous section, the $\widehat{IBI}$ of the signal is estimated and the size of the window and the rate of the STFT are defined next. STFT is performed using the three considered windows. Table 5.1 and Figure 5.5 respectively show the value of the parameters calculated for this signal denoted as Case 2 and the spectrogram for the three cases.

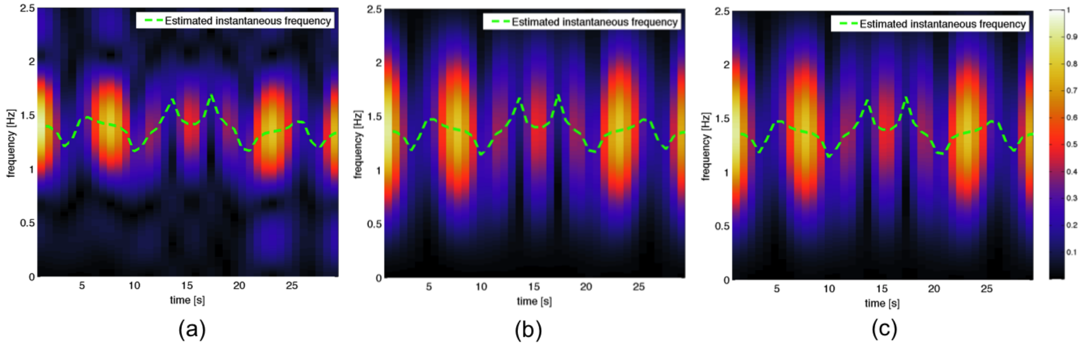


Figure 5.5: **Spectrogram of a signal under test.**
Using window of size $a = 2$ s, overlap between windows $ov = 1$ s and type of window (a) rectangular, (b) Hamming, and (c) Chebyshev.

5.3.2 Test signal with noise

It is well known that various noise levels are always present in real systems. In implementing the above presented methods, different values of Signal to Noise Ratio (SNR) were considered, using the test signal presented in Section 5.3.1.2. The noise is characterized by a Gaussian distribution of zero mean and standard deviation given by the SNR.

The actual value of the instantaneous frequency, calculated through the Hilbert transform for a noiseless case from the previous section, was compared with the results obtained when a noisy signal is analyzed with the time domain and time-frequency domain methods. For the latter, only Chebyshev window was considered.

[1] One precaution here is that Hilbert transform cannot be applied to broadband signals, such as real data [59]. See further remarks in Section 5.3.2.

5.3.3 Case Study: Experimental Data

The proposed time-frequency analysis technique is now applied to the *raw data* obtained the recorded BVP sensor measurements from the 12 subjects participated in the mental-workload experiment [101, 100], Data Set 2; see Chapter 3.1 for more details on the experimental setup and procedure.

In the experiment, 12 subjects played the following games in this particular order: Rest 1 (R1), Easy 1 (E1), Difficult (D), Rest 2 (R2), and Easy 2 (E2), each lasting one minute, where E1 and E2 are identical easy levels of the game, and R1 and R2 are rest periods when there is no game displayed on the screen and subjects rest. One of the sensors utilized in the experiment is a blood volume pulse (BVP) sensor recorded at 2048 Hz, from which heart rate (HR) and inter-beat interval (IBI) can be calculated. The BVP data is recorded during the rest period where subjects were at rest.

For each BVP signal data recorded corresponding to each segment of the game, pNN50 in time domain using (5.4)is first calculated as was done similarly in [101], and also by applying the proposed time-frequency analysis technique. Here, the pNN50 metric is calculated when a person is at rest or playing a game with different levels of difficulty. Specifically, the pNN50 is computed on five different temporal segments for a given player: two identical easy-level game segments (E1 and E2), one difficult-level game segment, and two identical no-game rest segments (R1 and R2). In this analysis, no low-pass-filter pre-processing is applied to the data although this is a common practice [50, 40, 3]. In this way, one can gain better insight regarding the consistency of the pNN50 calculated in time and time-frequency domain for different SNR.

As various noise levels are always present in real systems, different values of Signal to Noise Ratio (SNR), namely, 0dB, 5dB, 10dB, 15dB, 20dB, and 40dB, are considered in implementing the above presented methods, STFT algorithm. The noise is characterized by a Gaussian distribution of zero mean and standard deviation given by the SNR. Prior to the STFT analysis, a noise level is artificially added to the experimental BVP data[2].

Finally, for each simulated signal to noise ratio, SNR = 0dB, 5dB, 10dB, 15dB, 20dB, and 40dB, corresponding to a segment of the experiment, the pNN50 in time domain using (5.4) *again*[3], and by using proposed time-frequency analysis technique is calculated. For STFT, three different window options are used, namely, Rectangular, Hamming, and Chebyshev with 50dB of attenuation.

[2] Notice that the raw BVP data is considered as the baseline signal here, although it may be inherently noisy. Noisy BVP data however refers to the raw BVP data with artificially added noise.

[3] This time on the noisy BVP data.

In this way, one can gain better insight regarding the consistency of the pNN50 calculated in time and time-frequency domain for different SNR.

Notice, the ground truth IBI and pNN50 data are not available in this experimental example; and, for this reason, a human operator extracts the IBI data manually via detecting the time occurrences of maxima points in the HR data, in order to compute the ground truth pNN50. To gain additional confidence about this decision, the pNN50 values calculated for the all 12 BVP data, used in our previous work [101] (under low pass filtering), are compared with the ones that are calculated here on the raw signal. The absolute difference between these two sets of PNN50$\%$ values has a mean of 5.20% and a standard deviation of 3.50%, across the five segments of the game and for all 12 subjects. Since the average difference within each segment is sufficiently small, the pNN50 values calculated in time-domain based on the raw signal could be used as a ground truth in this case study.

The next step is to investigate how well the time-frequency domain method computes the pNN50 values, compared with those that are computed using the conventional method, see Section 5.2.1. Specifically, the following questions are of interest: *(a)* how consistent is the proposed method in calculating pNN50 across various noise levels added to the raw signal? and *(b)* which windowing technique is more consistent across those noise levels?

5.4 Results and Discussion

5.4.1 Test signals with no noise

5.4.1.1 Synthetic sinusoidal signal

Figure 5.6 depicts the comparison among the actual value of the instantaneous frequency given by (5.14), the estimation extracted from the spectrograms, and the value calculated in time domain as inverse of the NN intervals.

Using the result in (5.12), the NN interval differences are calculated directly from the estimation of the instantaneous frequency in the time-frequency domain. This result is used to compute the pNN50 as described in (5.4). Table 5.2 summarizes the value of the RMS of the differences between the actual value and the different estimations of the instantaneous frequencies, and the pNN50 calculated with each method.

The pNN50 values calculated using the time-frequency domain method are closer to the ground truth, given by equation 5.14, than those given by time domain method. Nonetheless, since

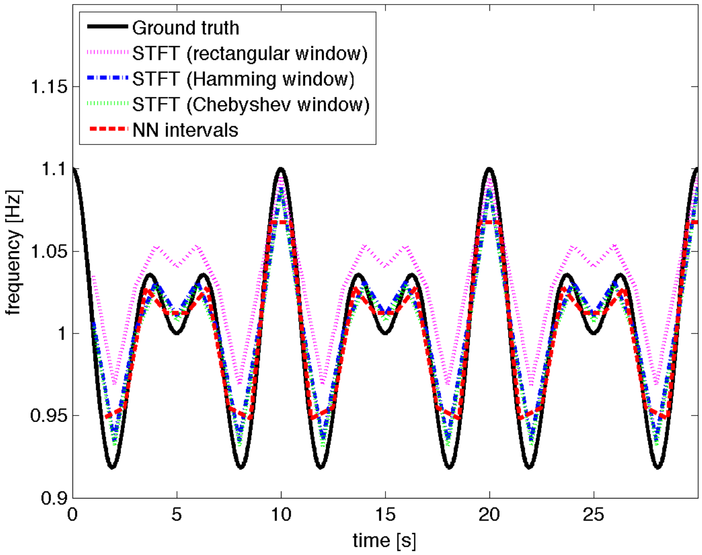


Figure 5.6: **Instantaneous frequency of the signal.**

	RMS	pNN50
Ground truth	0	0.5415
STFT (rectangular window)	0.0344	0.5000
STFT (Hamming window)	0.0089	0.5862
STFT (Chebyshev window)	0.0073	0.5862
Time domain	0.0053	0.3860

Table 5.2: **RMS of the estimation of the instantaneous frequency and pNN50 for a sinusoidal signal.**

there is no noise corrupting the signal, both time domain and time-frequency domain techniques, using a judiciously selected Chebyshev or Hamming window, are suitable to detect the instantaneous frequency. The uniform rectangular window is, however, not suitable in this particular case due to its inherently high side lobe levels.

5.4.1.2 Synthetic signal with several spectral components

Figure 5.7 shows the instantaneous frequency calculated with the Hilbert transform (ground truth), and the estimations obtained using the time-frequency domain method and the time-domain method. Table 5.3 summarizes the RMS of the differences between the ground truth and the estimations, and shows the values of pNN50 values calculated using each method.

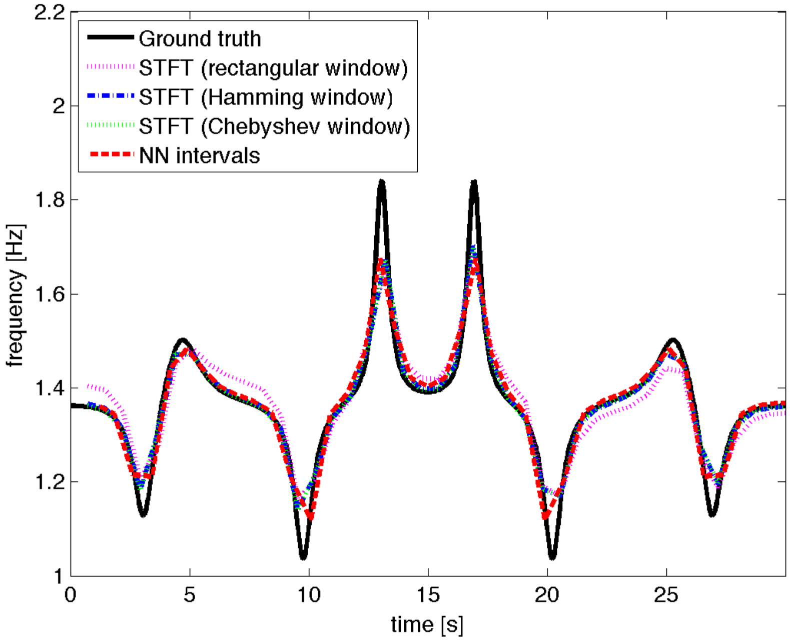


Figure 5.7: **Instantaneous frequency of the signal under test.**

It is demonstrated that both the original time domain method and the proposed time-frequency method, accurately estimate the instantaneous frequency and the pNN50 value for two synthetic signals when no noise is present. Next section shows the results of both methods when the data is corrupted by noise.

	RMS	pNN50
Ground truth	0	0.3666
STFT (rectangular window)	0.0425	0.3418
STFT (Hamming window)	0.0270	0.3734
STFT (Chebyshev window)	0.0268	0.3734
Time domain	0.0357	0.3544

Table 5.3: RMS **of the estimation of the instantaneous frequency for the test signal.**

5.4.2 Test signal with noise

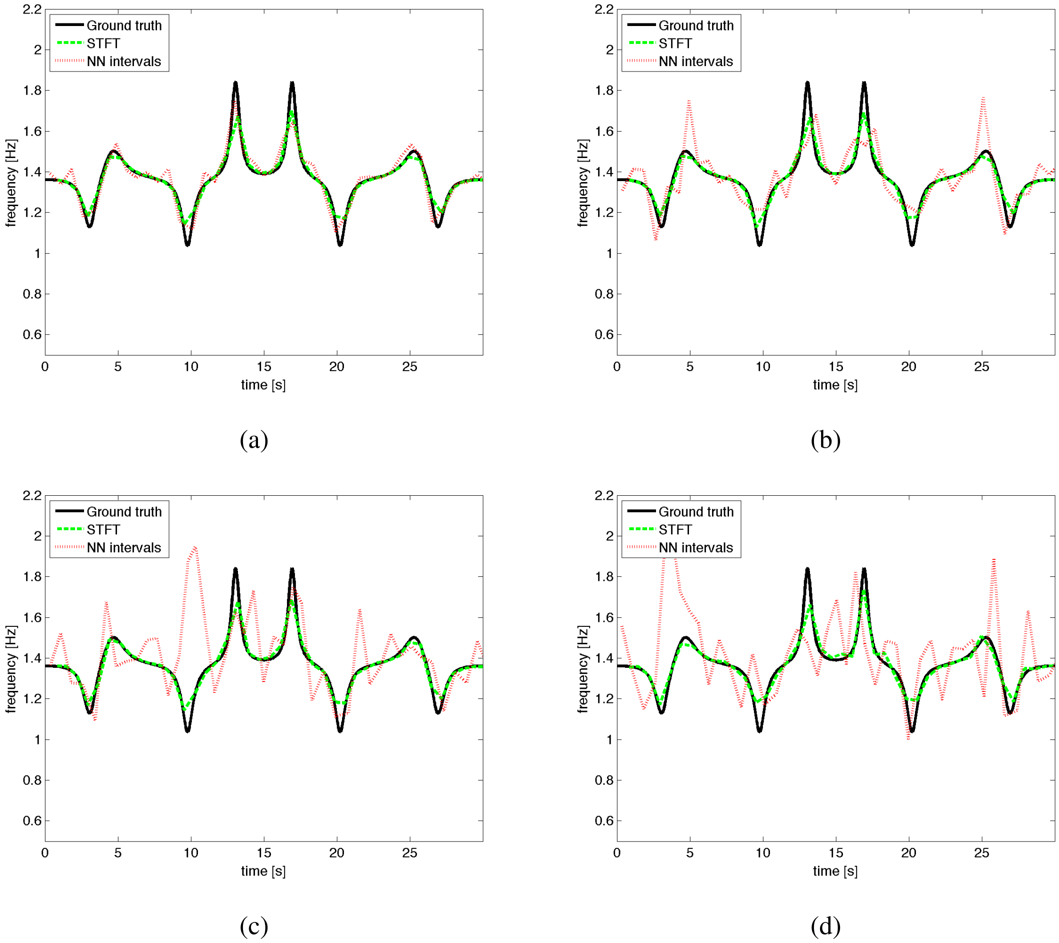
Figure 5.8(a-d) shows the actual and estimated values of instantaneous frequency for several values of SNR. Time domain method performs worse as the noise level increases and masks the correct positions of the maxima points in the HR data[4]. Nevertheless, the instantaneous frequency calculated with the time-frequency domain method is similar to the noiseless case, even for values of SNR as high as 0dB.

Figure 5.9(a) shows the RMS of the error between the estimation of each method and the ground truth value, and Figure 5.9(b) compares the estimations of the pNN50 statistics calculated with both methods. For each value of SNR, 100 simulations with random noise have been performed. The value for RMS and pNN50 were calculated as the mean value of all the simulations for each level of noise, and the uncertainty interval is given by the standard deviation obtained. We clearly observe that for SNR levels lower than 40dB, the pNN50 values calculated in time domain are unreliable, whereas the estimations made by the time-frequency domain method are consistent and accurate.

5.4.3 Case Study: Experimental Data

For all the 12 subjects, the pNN50 values corresponding to each segment of the game are calculated in time-frequency domain implementing the proposed STFT method and using three windowing techniques across various noise levels added to the raw signal. The results are also

[4] Notice that with added noise, the arising signal is no longer narrow band hence Hilbert transform can no longer be utilized. Indeed, with this transform the prediction of instantaneous frequency is erroneous and hence suppressed from the figures.



(a)

(b)

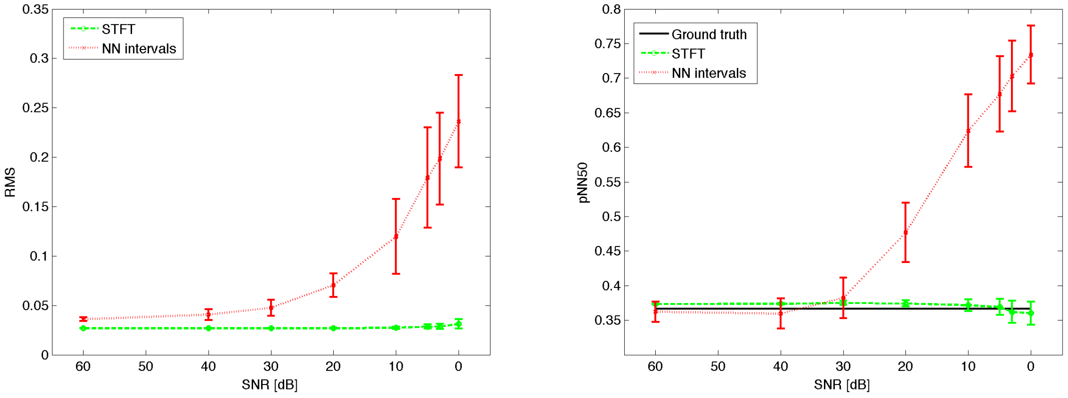
(c)

(d)

Figure 5.8: **Instantaneous frequency for different values for SNR.**
SNR $=$ 30dB, 10dB, 5dB, and 0dB. Computations with Hilbert transform are suppressed as they lead to erroneous results.

compared with the corresponding ones in time domain that are defined as the ground truth values for each subject and in each segment of the game. Here, two sample results of two random subjects are presented in Figs. 5.10 and 5.21.

As depicted in Figs. 5.10–5.21a, the pNN50 values computed using the traditional time-domain technique are degraded for low SNR values. However, the proposed time-frequency domain technique accurately computes the pNN50 even for low SNR values for the corresponding segment of the game (Figs. 5.10–5.21b, 5.10–5.21c, and 5.10–5.21d for Uniform, Hamming, and Chebychev, respectively).



(a)

(b)

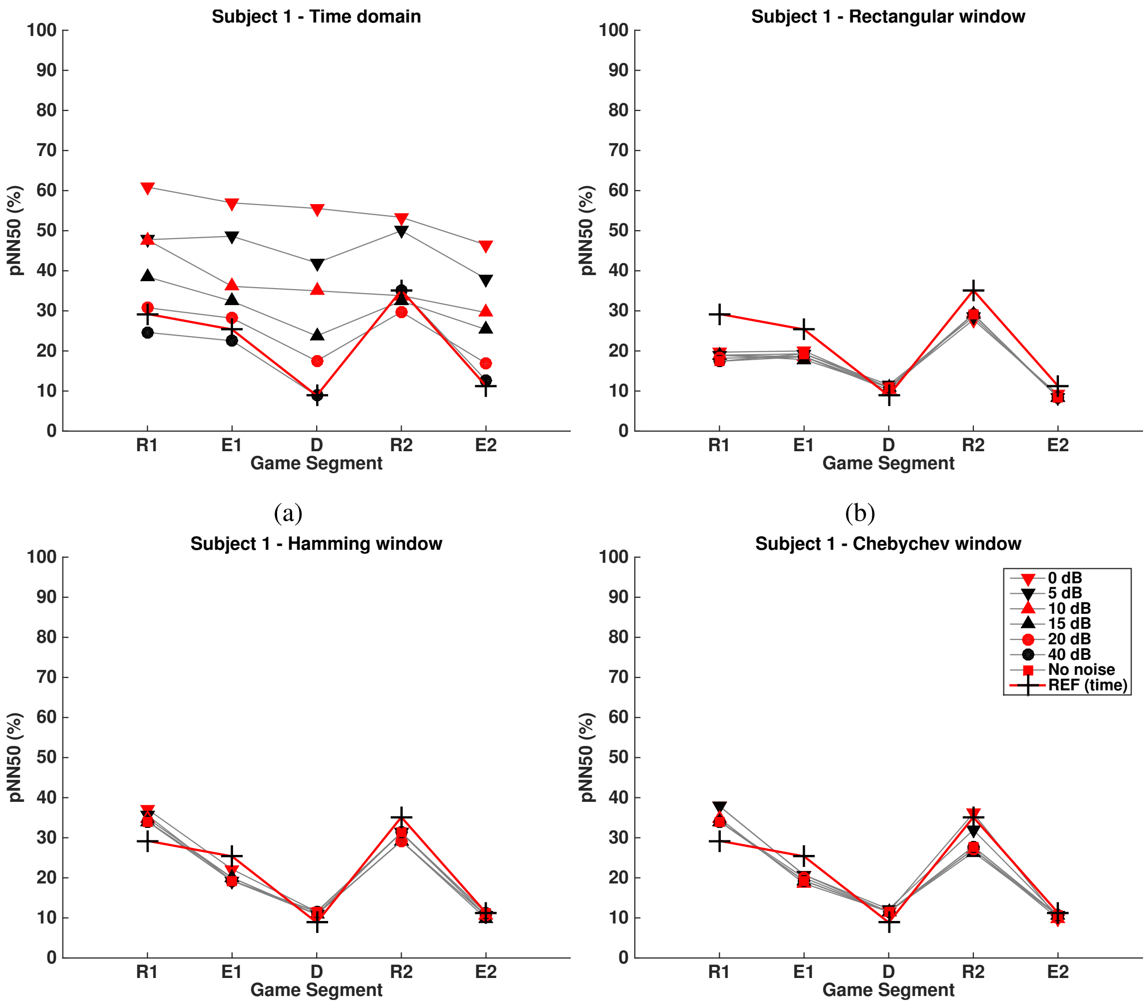
Figure 5.9: **Results of the simulations with noisy signals**
(a) SNR vs. RMS, and (b) SNR vs. pNN50.

To investigate how well the time-frequency domain method computes the pNN50 values, the corresponding pNN50 values calculated in time domain, and also by using the proposed technique in time-frequency domain across various noise levels are compared with the defined ground truth, and then the absolute *difference* is observed.

For each subject, the average of pairwise comparisons between the pNN50 values corresponding to each segments of the game in time domain, and the ones that are calculated in time-frequency domain, with the ground truth are presented on Table 5.4, as well as in Figure 5.22. For each subjects, the average of absolute *difference* is calculated out of 30 data points (pNN50) in time domain (6 noise levels × 5 segments of the game) and out of 35 data points (1 no noise + 6 noise levels × 5 game segments) for each window in time-frequency domain corresponding to all game segments.

Figure 5.23 represents the average of mean absolute difference presented in Table 5.4 for time and time-frequency domain approach in calculating the pNN50 across all the subjects.

Clearly, the time domain approach in calculating the pNN50 values on the BVP signal with noise carries the *highest* amount of difference on average, 17.30 ± 5.43, across all five game segments. The results clearly show the impact of strong noise level within the signal, which stems from corrupted IBI data. Results obtained based on STFT suggest that implementing Hamming and Chebychev windows results in “less amount” of difference in pNN50 values, 5.38 ± 3.41 and 5.66 ± 3.53 on average respectively, when compared with the corresponding values calculated by



(a)

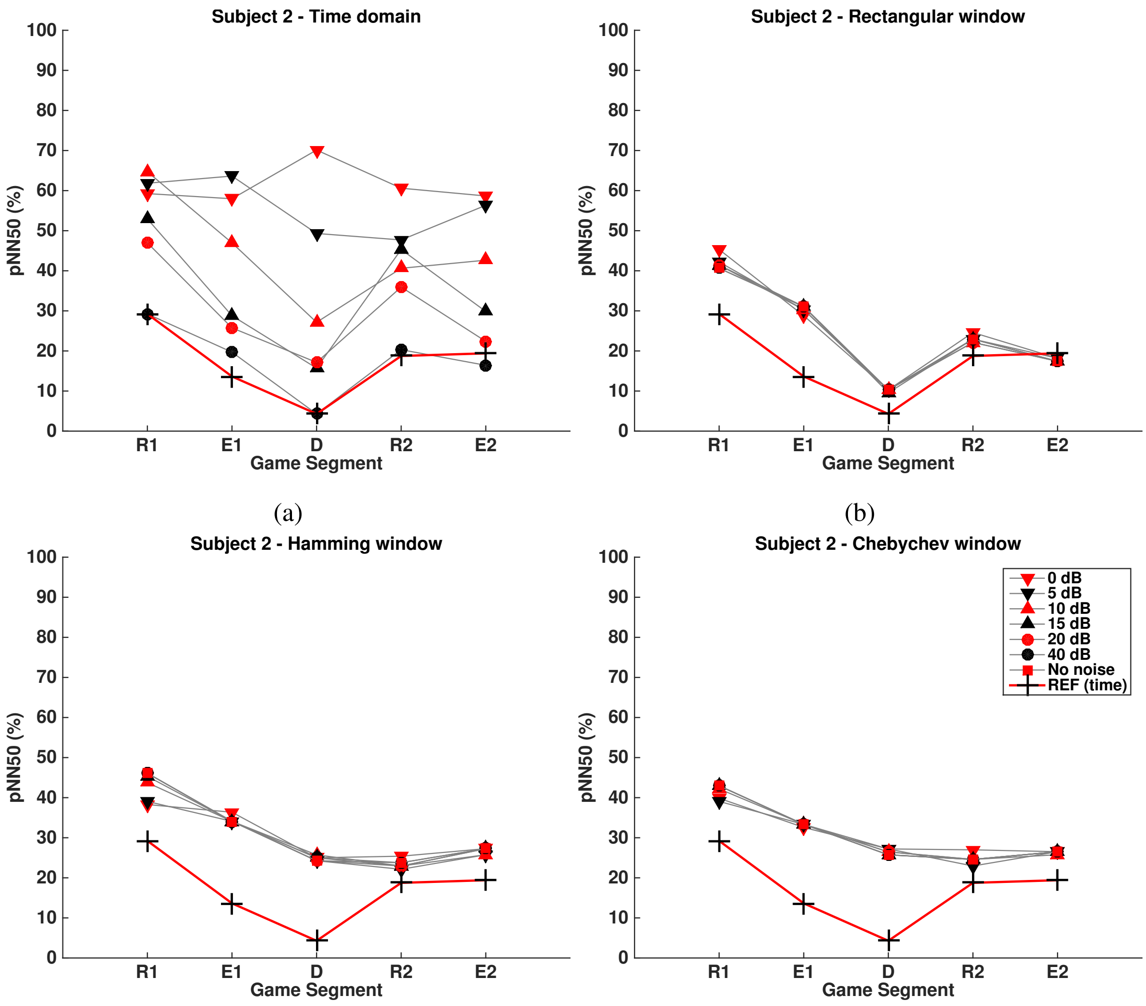
(b)

(c)

(d)

Figure 5.10: **pNN50 corresponding to all segments of the game, across different noise levels artificially added to the 2048 Hz signal for subject 1.**
Time domain (top-left), Rectangular window (top-right), Hamming window (bottom-left), and Chebychev window (bottom-right).

time-domain approach. This is consistent with what is obtained in [100] where the analytical signals Hamming and Chebyshev windowing produced superior results, compared with the corresponding values calculated by time domain approach.



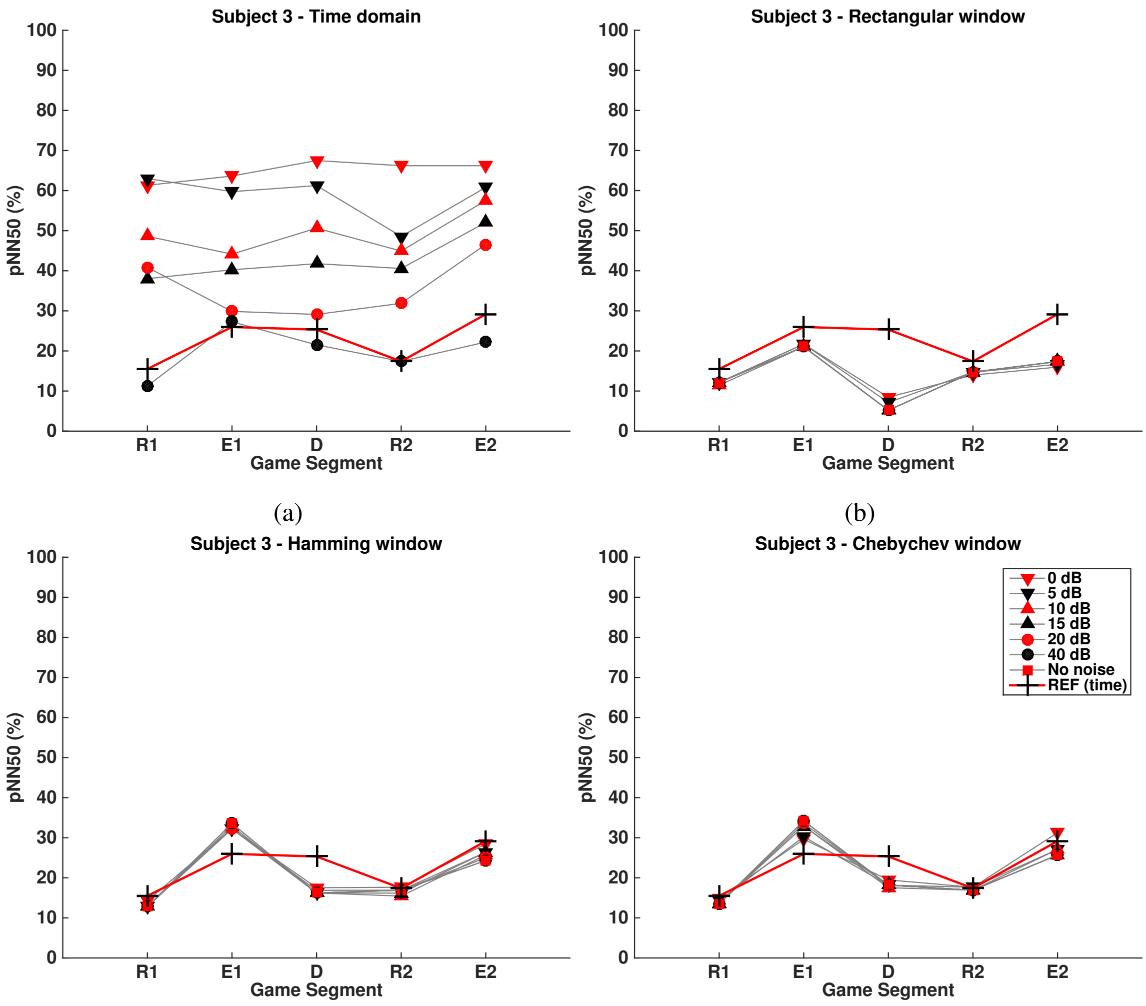
(a) (b)

(c) (d)

Figure 5.11: **pNN50 corresponding to all segments of the game, across different noise levels artificially added to the 2048 Hz signal for subject 2.**
Time domain (top-left), Rectangular window (top-right), Hamming window (bottom-left), and Chebychev window (bottom-right).

5.5 Summary

Calculation of pNN50 in time-domain is traditionally based on the IBI data, which needs to be constructed from detected maxima peaks in the BVP signal. Results suggest that in the case when this signal is noisy, it is difficult to detect the peaks, which ultimately affects pNN50 calculations when performed in time domain (Figs. 5.14a and 5.19a). To accurately calculate pNN50,



(a)

(b)

(c)

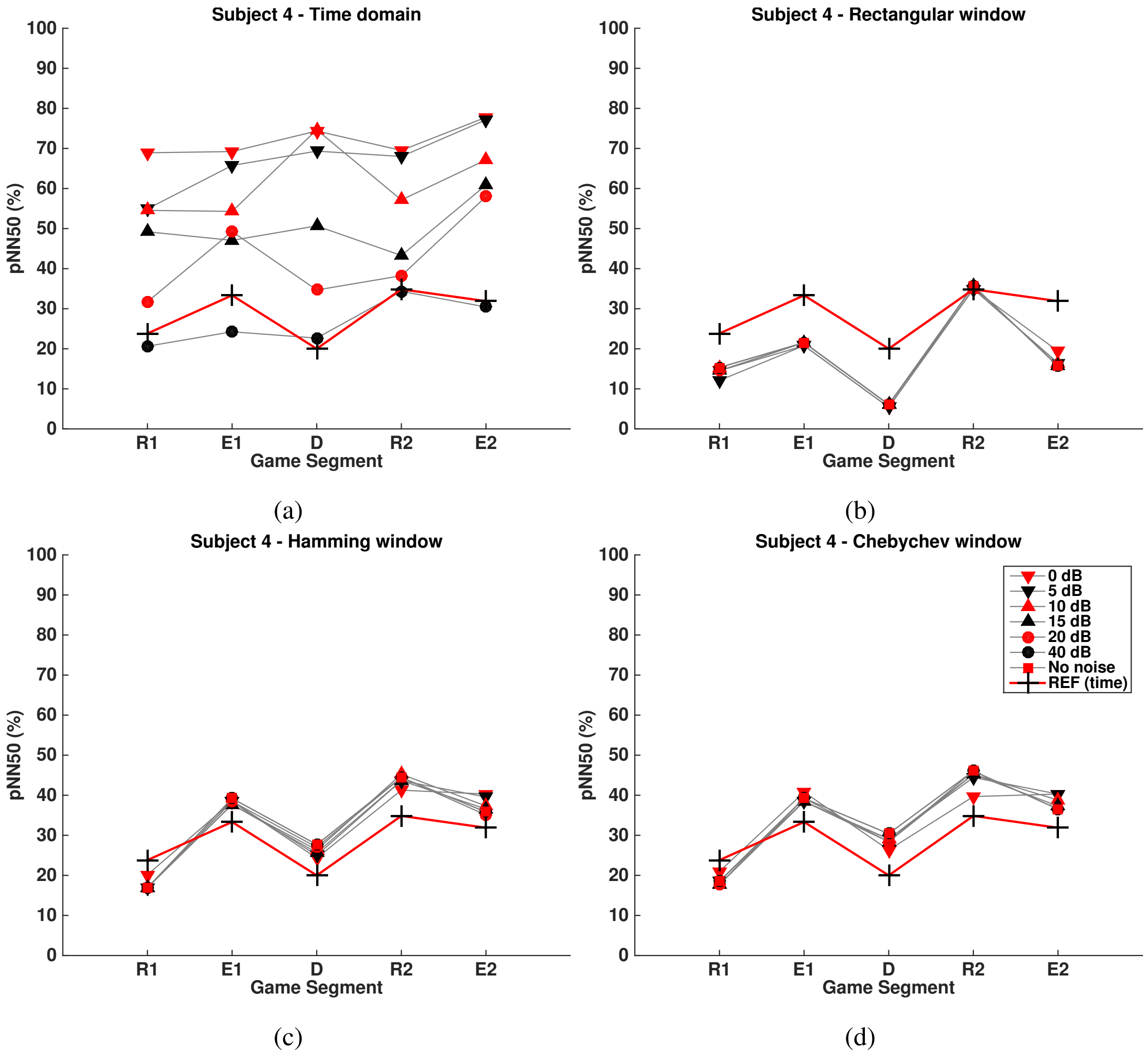
(d)

Figure 5.12: **pNN50 corresponding to all segments of the game, across different noise levels artificially added to the 2048 Hz signal for subject 3.**

Time domain (top-left), Rectangular window (top-right), Hamming window (bottom-left), and Chebychev window (bottom-right).

especially from signals with noise, and whose IBI cannot be reliably computed, the proposed time-frequency domain approach render more accurate results, as demonstrated. STFT approach also avoids many time-domain based post processing efforts otherwise needed to reliably extract the IBI, thereby allowing a more robust means of computing pNN50.

For comparison purposes, classical approach based on time domain analysis is also taken, both on the raw BVP signal and this signal with various artificially added noises. In the case when



(a)

(b)

(c)

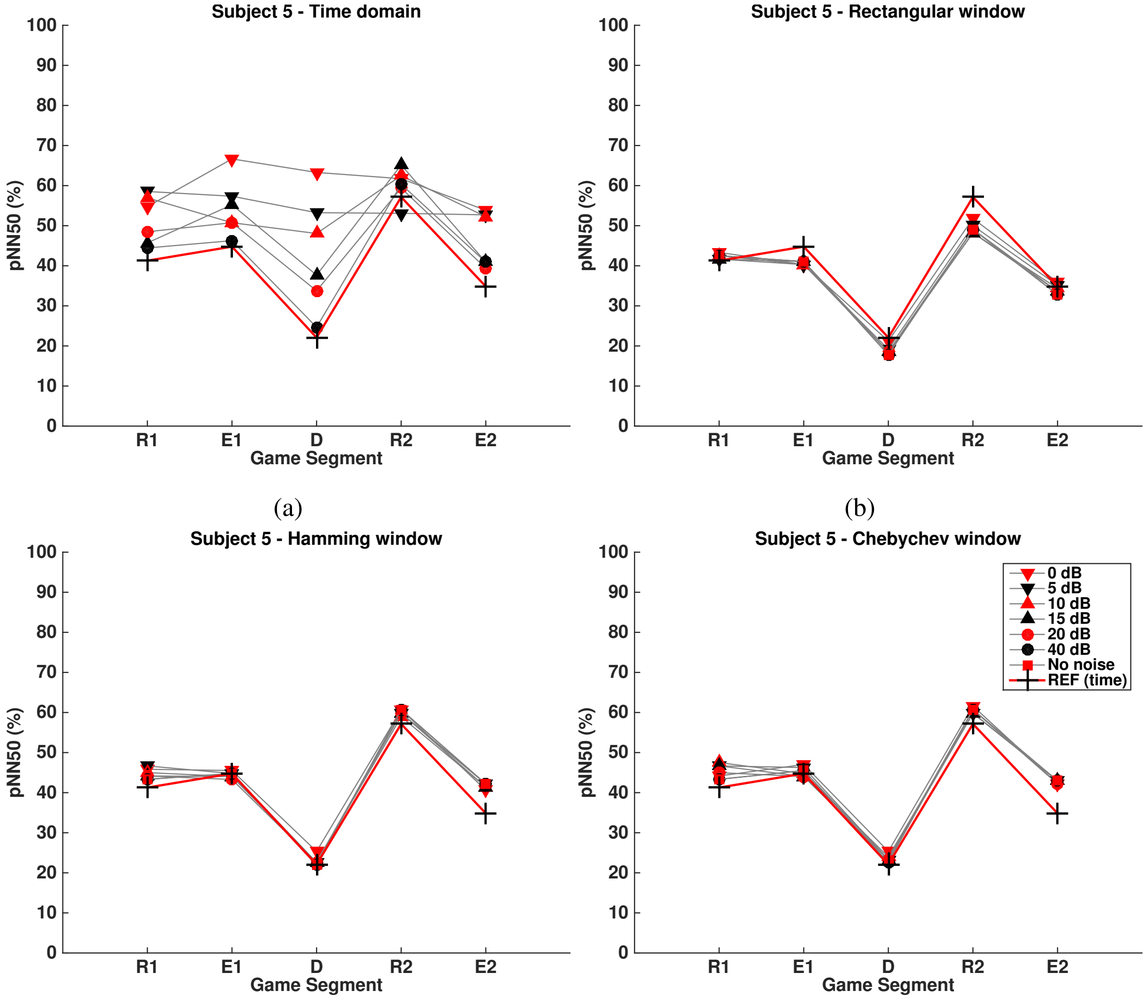
(d)

Figure 5.13: **pNN50 corresponding to all segments of the game, across different noise levels artificially added to the 2048 Hz signal for subject 4.**

Time domain (top-left), Rectangular window (top-right), Hamming window (bottom-left), and Chebychev window (bottom-right).

the signal is noisy, as expected it is challenging to detect the exact locations of the maxima points of the signal in time-domain, making it difficult to accurately compute pNN50 values. With pNN50 computed using the raw signal serving as ground truth, the results of the two approaches (time and time-frequency domain) in calculating pNN50 are compared.

(a) calculating the pNN50 values in time domain carries the "highest" amount of difference on average when compared with the pNN50 values calculated via the time-frequency domain



(a)

(b)

(c)

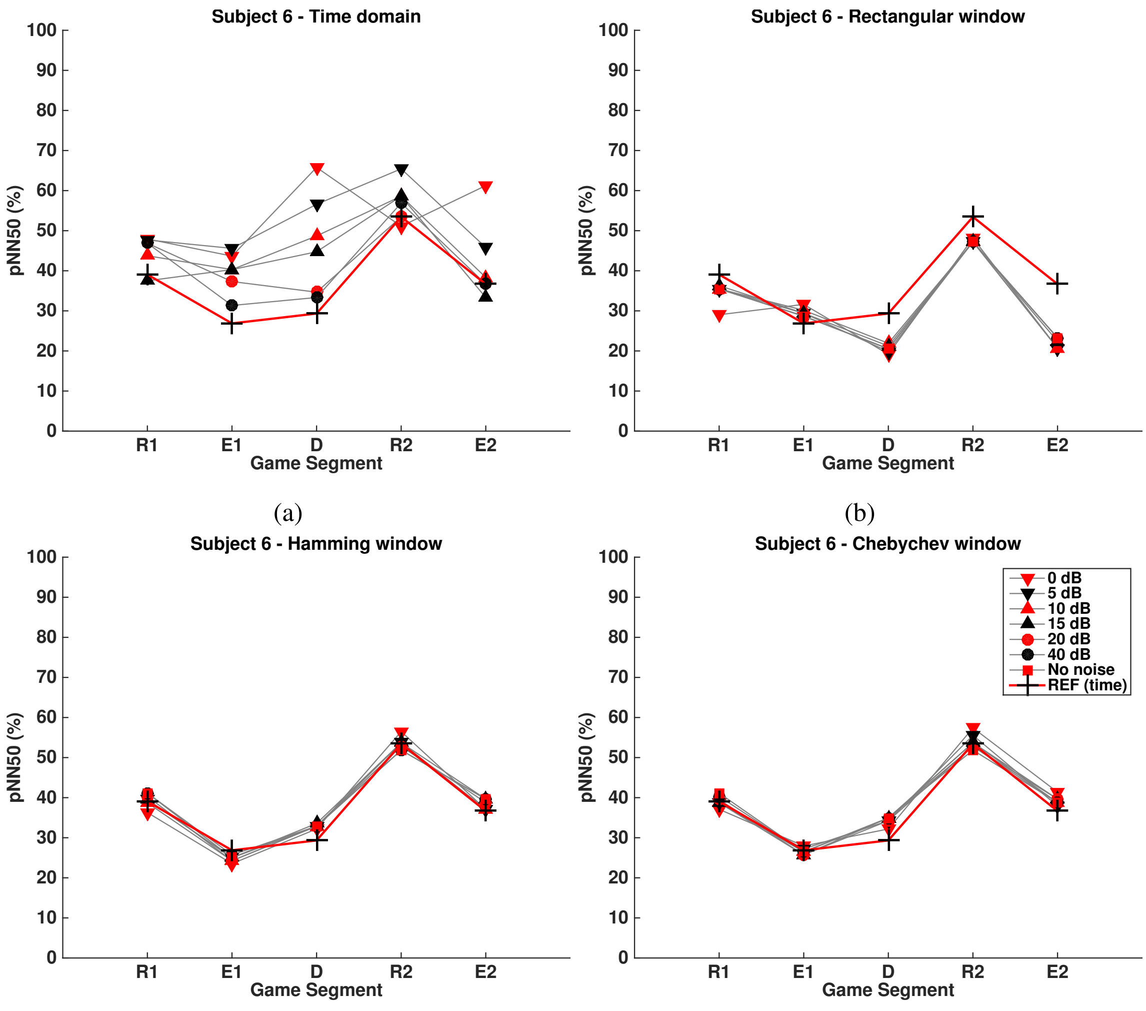
(d)

Figure 5.14: **pNN50 corresponding to all segments of the game, across different noise levels artificially added to the 2048 Hz signal for subject 5.**

Time domain (top-left), Rectangular window (top-right), Hamming window (bottom-left), and Chebychev window (bottom-right).

approach. (b) In time-frequency domain. Moreover, results obtained based on STFT suggest that implementing Hamming and Chebyshev windows on the experimental data results in “less” amount of error in pNN50 values, on average, which is consistent with what is obtained in [100] where the analytical signals Hamming and Chebyshev windowing produced superior results, compared with the corresponding values calculated by time domain approach.

The calculated pNN50 using the combined time-frequency approach does not match the



(a)

(b)

(c)

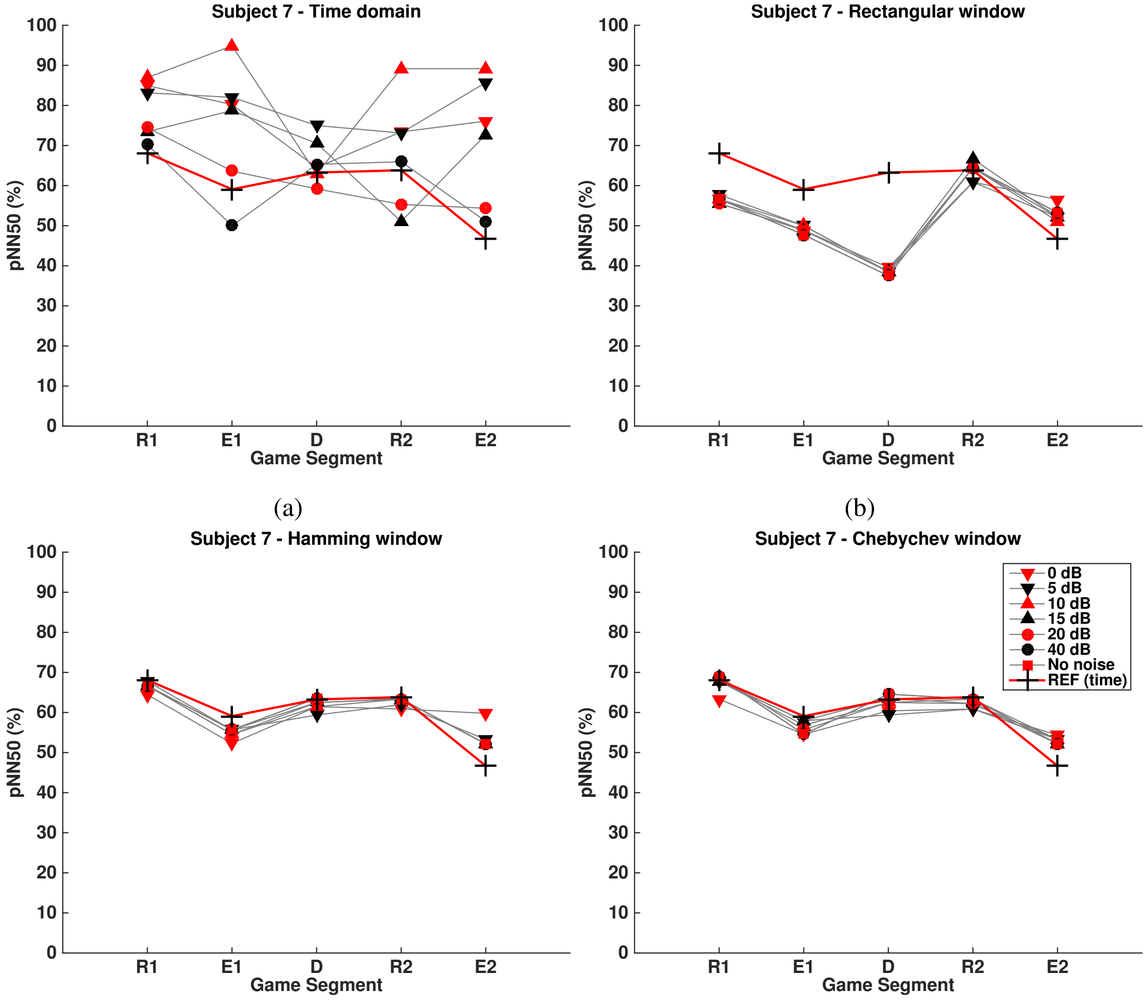
(d)

Figure 5.15: **pNN50 corresponding to all segments of the game, across different noise levels artificially added to the 2048 Hz signal for subject 6.**

Time domain (top-left), Rectangular window (top-right), Hamming window (bottom-left), and Chebychev window (bottom-right).

ground truth for all the 12 subjects and game segments. For instance,

For instance, in Figures 5.12, 5.18, and 5.21 (11, 18, 21), we present cases scenario with large amount of errors between the ground truth and the pNN50 values calculated using the proposed approach, see Table 5.4, Figure 5.22. Although such results warrant further efforts on improving the STFT algorithm, nevertheless, our proposed time-frequency calculation of the pNN50 outperforms the classical time domain method for most of the subjects and game segments, thus making our



(a)

(b)

(c)

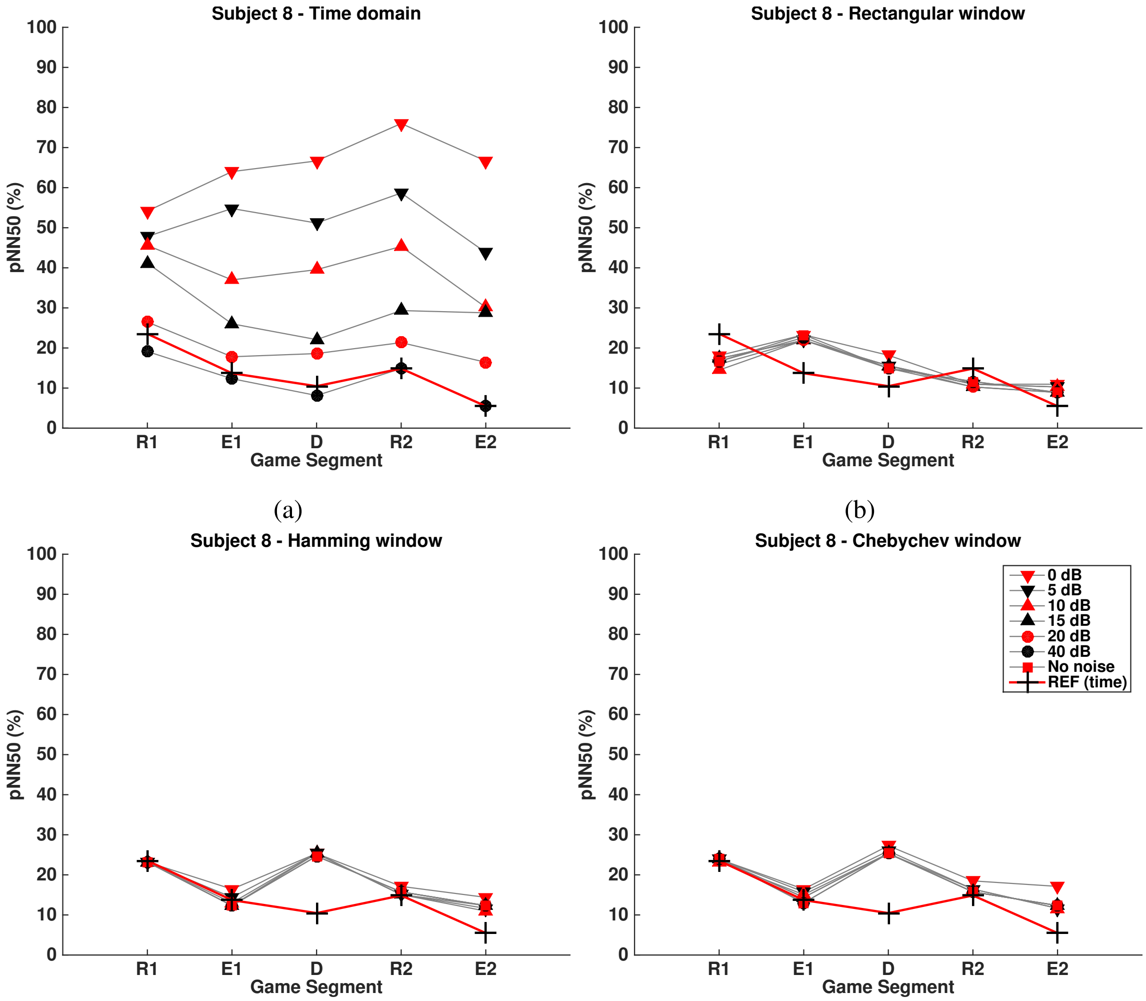
(d)

Figure 5.16: **pNN50 corresponding to all segments of the game, across different noise levels artificially added to the 2048 Hz signal for subject 7.**

Time domain (top-left), Rectangular window (top-right), Hamming window (bottom-left), and Chebychev window (bottom-right).

approach appealing for further investigation.

It should be noted that the lack of a ground truth in any experimental case study results in our inability to determine which method captures the pNN50 more accurately, when compared to an absolute truth. However, we clearly show that using the proposed time-frequency method for computing the pNN50 is more reliable, in general, with noisy BVP signals.

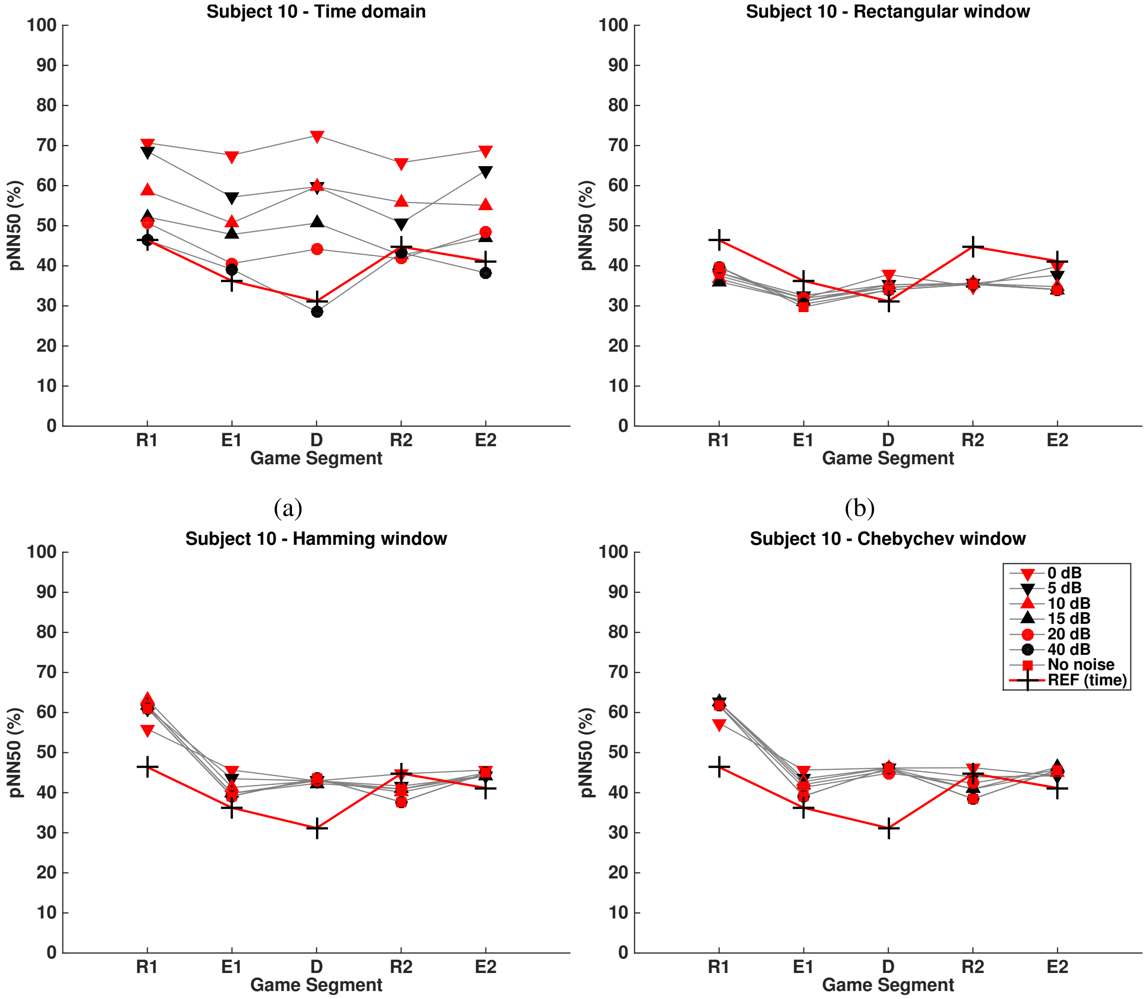


(a) (b)

(c) (d)

Figure 5.17: **pNN50 corresponding to all segments of the game, across different noise levels artificially added to the 2048 Hz signal for subject 8.**

Time domain (top-left), Rectangular window (top-right), Hamming window (bottom-left), and Chebychev window (bottom-right).



(a)

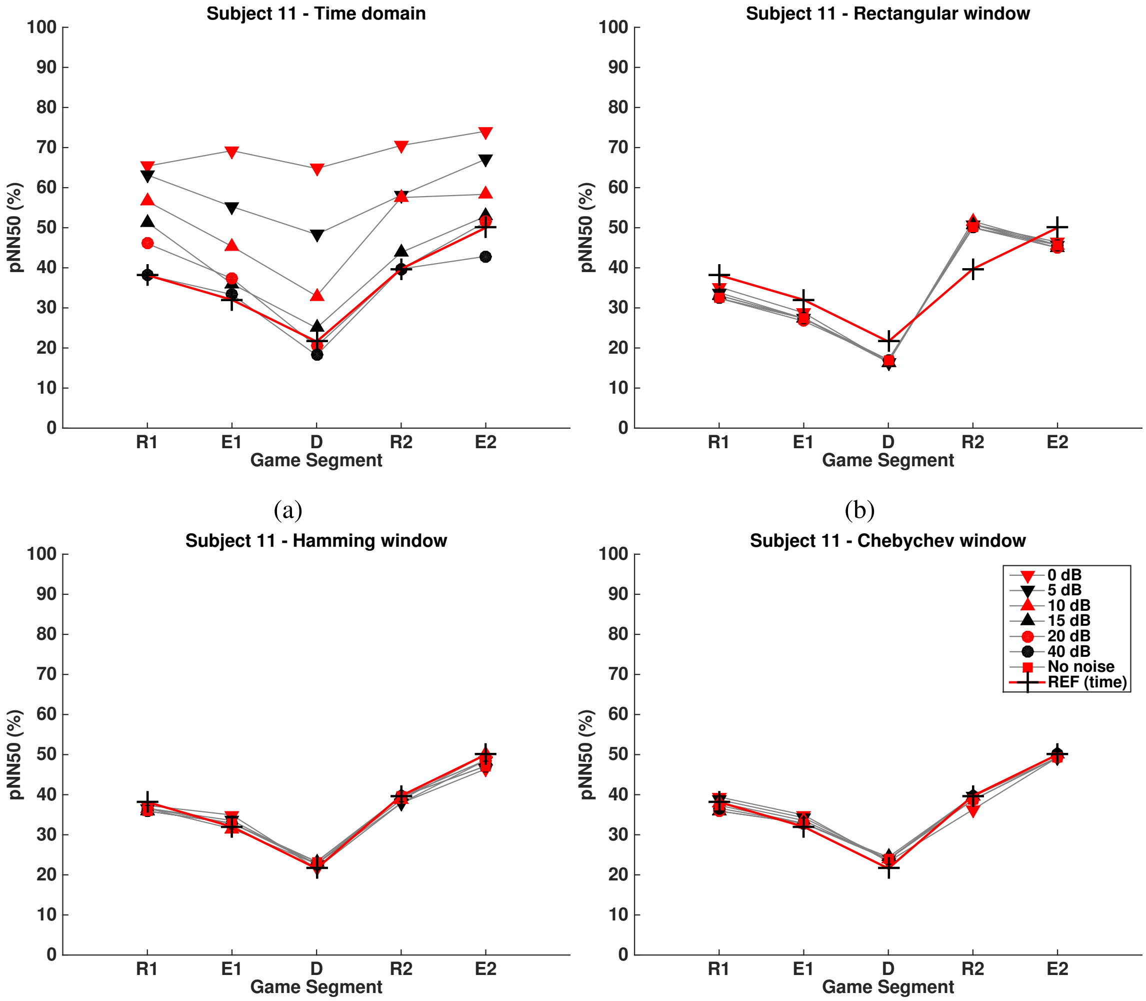
(b)

(c)

(d)

Figure 5.18: **pNN50 corresponding to all segments of the game, across different noise levels artificially added to the 2048 Hz signal for subject 10.**

Time domain (top-left), Rectangular window (top-right), Hamming window (bottom-left), and Chebychev window (bottom-right).



(a)

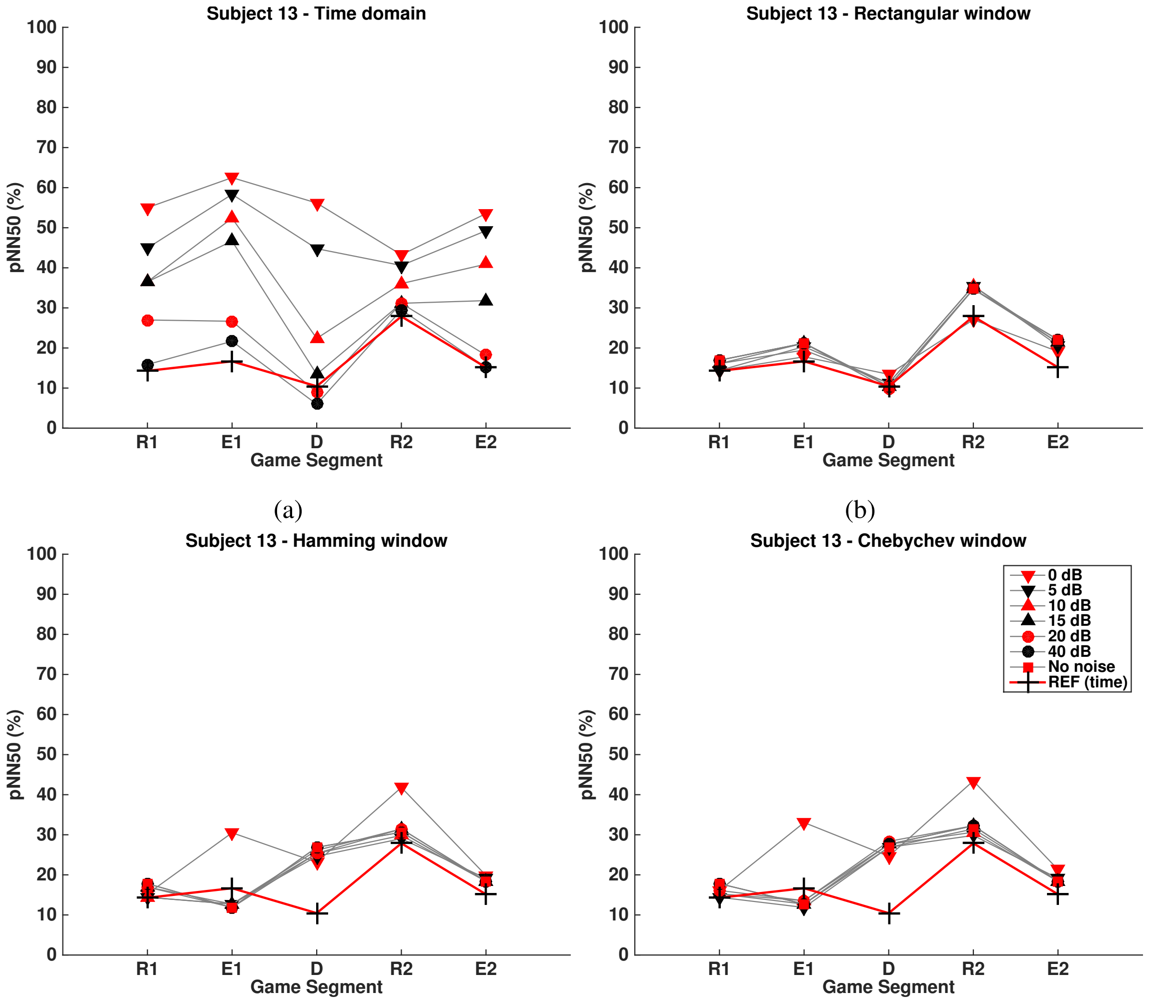
(b)

(c)

(d)

Figure 5.19: **pNN50 corresponding to all segments of the game, across different noise levels artificially added to the 2048 Hz signal for subject 11.**

Time domain (top-left), Rectangular window (top-right), Hamming window (bottom-left), and Chebychev window (bottom-right).



(a)

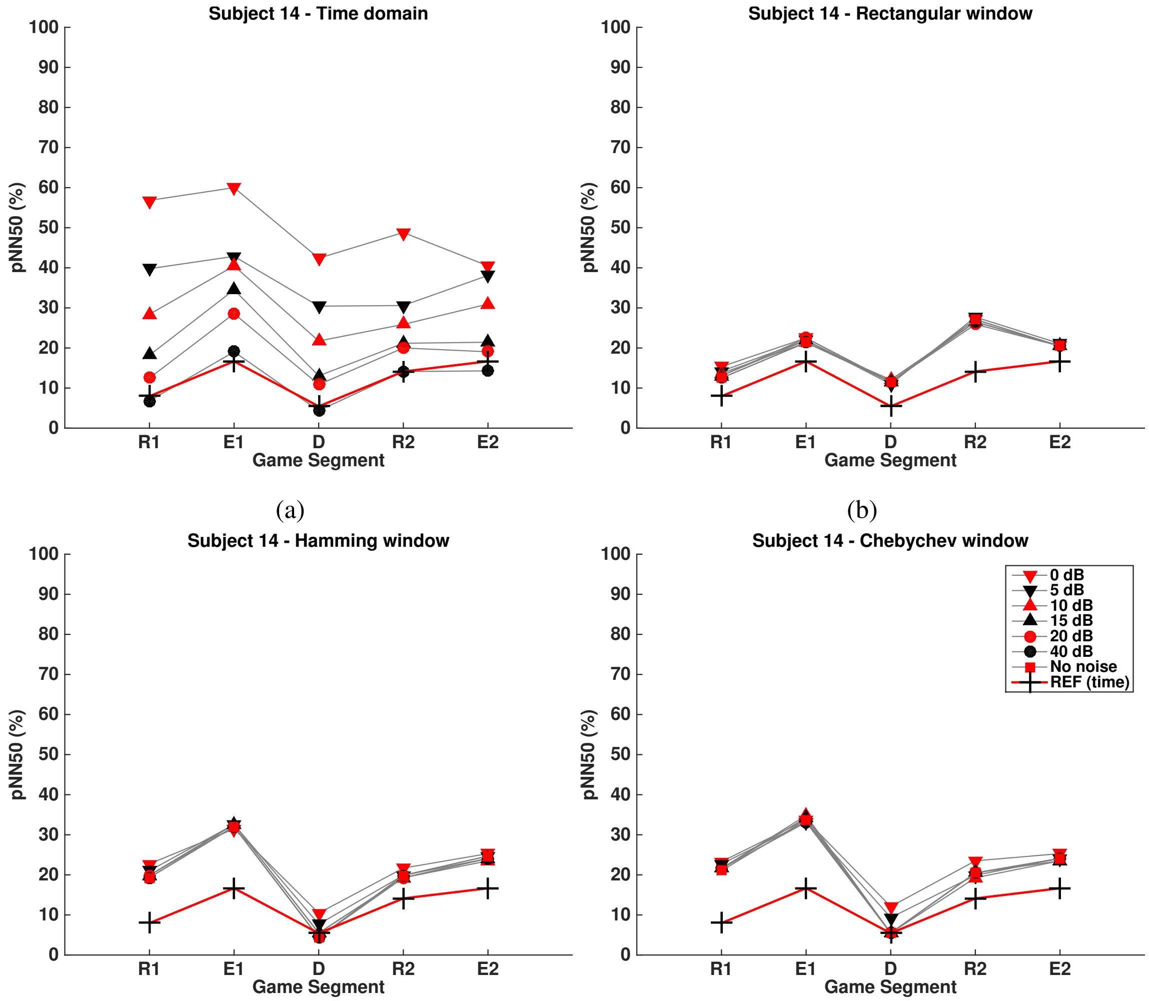
(b)

(c)

(d)

Figure 5.20: **pNN50 corresponding to all segments of the game, across different noise levels artificially added to the 2048 Hz signal for subject 13.**

Time domain (top-left), Rectangular window (top-right), Hamming window (bottom-left), and Chebychev window (bottom-right).



(a) (b)

(c) (d)

Figure 5.21: **pNN50 corresponding to all segments of the game, across different noise levels artificially added to the 2048 Hz signal for subject 14.**

Time domain (top-left), Rectangular window (top-right), Hamming window (bottom-left), and Chebychev window (bottom-right).

subject	time-domain	time-frequency domain			
		Rectangular	Hamming	Chebychev	
1	14.52 ± 12.53	5.64 ± 3.21	3.75 ± 2.17	4.33 ± 2.63	
2	23.75 ± 16.61	8.27 ± 5.76	13.54 ± 6.88	13.45 ± 6.65	
3	23.49 ± 14.51	8.49 ± 6.51	4.42 ± 3.08	3.70 ± 2.84	
4	25.44 ± 16.50	10.28 ± 5.44	6.43 ± 1.90	7.18 ± 2.35	
5	11.54 ± 9.44	3.49 ± 2.60	2.91 ± 2.46	3.53 ± 2.65	
6	9.71 ± 8.78	7.31 ± 4.45	1.98 ± 1.24	2.13 ± 1.85	
7	14.05 ± 11.74	10.91 ± 7.98	2.99 ± 2.70	2.73 ± 2.29	
8	23.32 ± 18.95	5.71 ± 2.05	4.74 ± 5.46	5.09 ± 5.70	
9	13.77 ± 10.94	6.45 ± 2.48	7.81 ± 4.86	8.62 ± 5.48	
10	13.25 ± 12.01	5.91 ± 2.59	1.36 ± 0.88	1.26 ± 0.96	
11	18.72 ± 15.41	3.67 ± 2.64	6.10 ± 5.34	6.62 ± 6.02	
12	15.99 ± 13.36	6.67 ± 3.17	8.56 ± 5.09	9.30 ± 5.77	
	17.30 ± 5.43	**6.90 ± 2.30**	**5.38 ± 3.41**	**5.66 ± 3.53**	average

Table 5.4: **pnn50 comparisons: STFT and time domain.**

The average of pairwise comparisons (absolute difference) between the pNN50 values corresponding to each segments of the game in time domain, and the ones that are calculated in time-frequency domain, with the ground truth

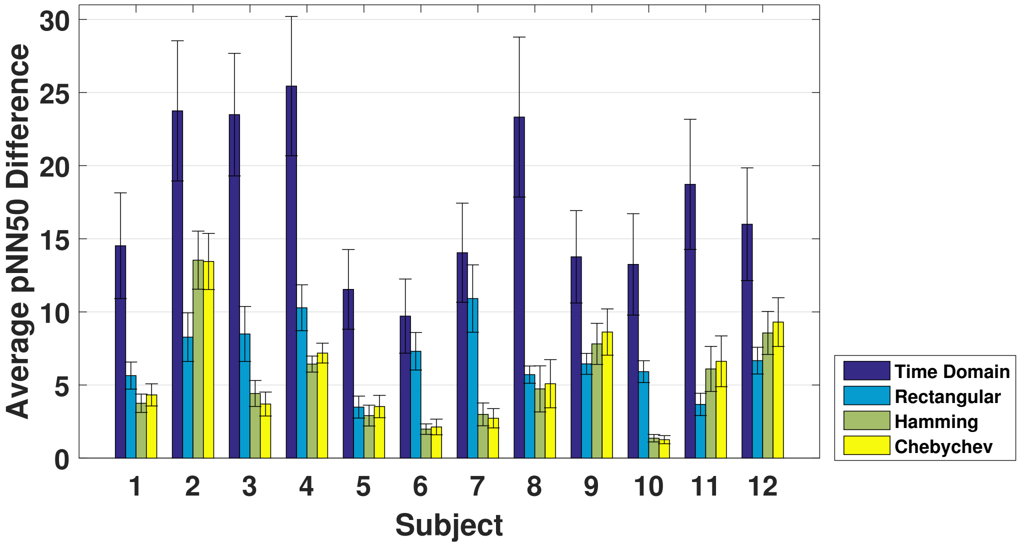


Figure 5.22: **The mean absolute difference error across all noise levels and game segments for each subjects.**

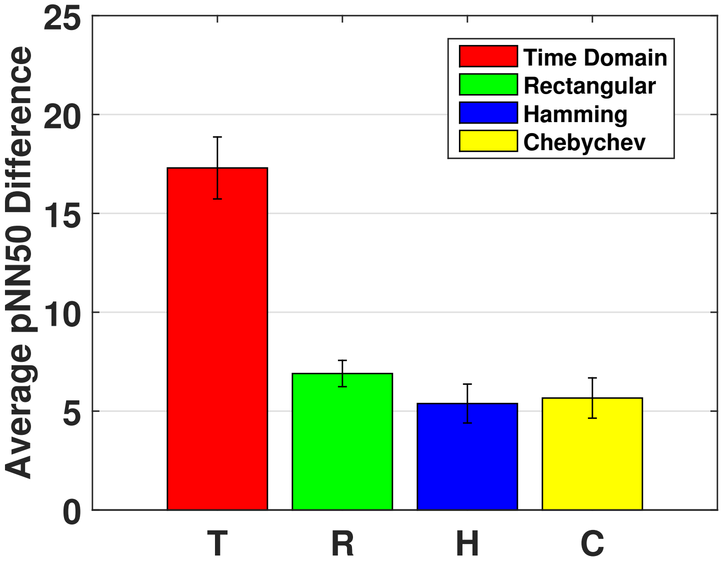


Figure 5.23: **Average of the mean absolute difference error across all subjects.** The average is calculate using the results presented in Table 5.4 across all 12 subjects participated in the study.

Chapter 6

MENTAL WORKLOAD EVALUATION VIA SUBJECTS' BEHAVIORAL METRICS

6.1 Introduction

Under surprise, panic, and mental overload, as well as due to lack of experience, humans may temporarily lose their capacity to rapidly make correct decisions [89, 28]. This may cause catastrophes and casualties in many scenarios [70], such as when a human is engaged with a task s/he is not familiar with, lacks proper and up-to-date training in a task encountered, and is required to quickly evaluate multiple and possibly competing parameters in a task.

In human-machine systems, negative effects of lack of training can be mitigated if the machine that the subject is interacting with could intelligently sense subjects' inexperience and provide assistance [134, 107, 47]. Such assistance would be extremely valuable in car driving, operating an aircraft and/or heavy machinery, and making critical decisions to carefully schedule landing/take-off of aircrafts at airport control towers. For this to be successfully implemented, one needs to address how the machine could know when the subject is indeed mentally overloaded. Affective computing can provide opportunities in this direction [104] through the study of human physiological responses, such as skin conductance [29], heart rate [89], EEG [134], EMG [81], and EOG [132].

Application of affective computing tools is ubiquitous, encompassing psychophysiology research [97], gaming [32], car driving [89], operating an aircraft simulator [116], and performing multi-tasks involving memory games and arithmetic operations [116]. Other studies include studying heart rate in real traffic pattern flights [135], associating heart rate irregularities to certain diseases [88, 4], utilizing EEG signals to infer intention to command robots with mind, and keeping track of eye blinks as an indicator of stress [116].

It was also argued in the literature that some technical issues may arise when using physiological sensors [12, 100, 123, 110, 41]. Moreover, latency in sensing could limit bandwidth. For example, in skin conductance sensors [29], the time difference between the onset of stimulus and sensory changes can be around 2–3 seconds, and analysis of HR in real time would require recording of multiple cycles before an inference can be made. Further, it is also necessary to make inferences more reliably for inexperienced subjects, whose physiological measurements may not well correlate with training algorithms and real-time detection schemes, see, e.g., [134]. One justification of this is because behavioral patterns of inexperienced subjects during task execution are much more unpredictable and variable, while those of experts are more strategic and systematic, matching well with computational or mathematical models.

Accordingly, we believe that one can infer mental workload changes thereby subjects' inexperience much faster by studying subjects' behavioral patterns. We hypothesize that such patterns must be affected when subjects are engaged with tasks that demand different levels of mental effort. These patterns, we believe, are correlated with the strategies subjects develop (or are unable to develop) while trying to cope with task difficulty, and they can be as simple as swiping the finger with different acceleration levels on a touchscreen display and performing different hand gestures as well as arm movements. From this viewpoint, it is critical to emphasize that these behavioral patterns are different from conventional performance metrics. The reason for this is that these patterns are end results of mental workload changes as dictated by subjects' inexperience and/or task difficulty, and are hence only loosely coupled with the specifics of a task being performed. On the other hand, performance metrics are directly and tightly associated with specific task outcomes as the very nature of performance definition. Undoubtedly, behavioral patterns and performance metrics are related, yet these patterns provide higher-level information more directly related to mental workload and strategy development, and less dependent on the specifics of a task. This makes the analysis of such patterns attractive, as features of these patterns can be used directly to study and compare subjects even across different tasks.

We argue that, if we could relate subjects' behavioral patterns to subjects' level of coping with a task, and there is some evidence supporting this in different contexts [41, 33], this would be a valuable “sensor” as such detections can be made much faster with higher bandwidth and in real time, through high-rate measurements such as via touchscreen mouse clicks and light-weight accelerometers attached to human body. Some studies already investigated behavioral patterns, especially hand gestures in touch-screen applications with iPads using subjects’ decision making times [41] and keystroke dynamics [33] as metrics with device usability as the research focus; others assessed decision making times with respect to number of choices needed to be evaluated by the subjects, see the well-known Hick-Hyman Law [55, 62] and Fitts Law [38]; and others implemented experimental studies in which subjects are trained in the tasks to be performed [115, 134, 116].

To the best of our knowledge, the above described research problem from the view point of making inferences to assess subjects’ inexperience, and using this inference toward an intelligent machine assistance scheme has so far not been studied. We recently started studying this research question from the perspective of how behavioral metrics relate to performance and heart-rate variability metrics [102]. In the cited study however balanced experiments and what the effects of game order are on subjects were not reported. The main focus of this article is to report these results. Specifically, here, we study and compare three classes of metrics obtained from balanced experiments in which subjects’ lack of training is explicitly considered: (a) a physiological metric called pNN50 associated with heart rate, which very well correlates with mental workload changes as we confirmed [101, 100], (b) metrics directly related to subjects’ task performance, and (c) behavioral metrics as described above. The main objective is to investigate whether or not behavioral metrics could offer any insight into drawing inferences regarding subjects’ lack of training or inexperience, and how in this regard they relate to the metrics in (a)–(b) in balanced experiments.

Here, twelve subjects play a touch-screen air traffic (AT) game, see Section 3.1 for details. In brief, this game, which also resembles a logistics/task-allocation game, has two difficulty levels (easy and difficult). The subjects learned the rules of both game levels, and they have sufficient training with the easy game, but not the difficult game level. While the subjects are taking the experiments, a BVP sensor is used to measure subjects’ heart rate activities, and subjects’ finger kinematics as well as decision making times are recorded through their interactions with the touch-screen game. Following this, experiments are balanced specifically by changing the order of the difficult game and easy game, and another group of twelve subjects take the same experiments. Next, the metrics (a)-(c) described above are calculated based on collected data for both experimental settings, to investigate the relationships amongst the metrics and with respect to game difficulty.

The chapter is organized as follows. We first present our air traffic (AT) game in details and introduce the main tasks/goal in the game, followed by the experimental protocol, and definition of metrics, including the formulation of behavioral metrics based on participants' finger-stroke data obtained from the touch-screen display. The results of the statistical analysis of these metrics with discussions are provided next, and the article ends with conclusions and future research directions.

6.2 Data Collection

The data collection is detailed in Chapter 3.1 where is the experimental setup is demonstrated for this study. In following we briefly introduce the experimental procedure.

6.2.1 Affective Computing using Blood Volume Pulse Sensor

The autonomic nervous system (ANS), which regulates many major physiological activities in human body, has two parts: the sympathetic nervous system, which modulates the body's resources for action under stressful conditions, and the parasympathetic nervous system, which relaxes and stabilizes the body into steady state [79]. Heart rate (HR) and heart rate variability (HRV) are linked with the state of ANS, and can hence be used to study mental workload [53]. In general, with higher levels of mental workload, it is expected that HR increases and HRV decreases [130]. A Blood Volume Pulse (BVP) [3] sensor can be used to measure HR activity and to calculate HRV [83]. HRV is calculated here in time domain based on inter-beat interval (IBI) time series, a well established technique [19, 2], using time-periods of consecutive peaks in HR data.

6.2.2 Experiment

The data analysis is conducted on the two data sets, namely Data Set 1 and Data Set 2, that are recorded in two experiments. Twelve subjects participate in the first experiment (Data Set 2). Another twelve subjects participate in the second experiment (Data Set 3), see Section 3.3.

The experiments are performed using a Dell PC machine running a 32 bit Windows 7 operating system. A 21.5 Dell$^{\text{TM}}$ ST2220T multi-touch monitor with 1920 × 1080 resolution, and at 60 Hz frame-rate is used for displaying the game. The subjects sit comfortably in front of the touch monitor, which is positioned vertically. Prior to the experiment, the subjects in Data Set 2 and Data Set 3 play the easy level of the game for two minutes to familiarize themselves with the game environment, interaction with the touch monitor, and performing the tasks in the game, i.e., drawing trajectories. Moreover, all the subjects are also instructed and presented the rules and challenges of the difficult game, to familiarize them with this game level. Subjects however do *not* play the difficult game.

We should note here that all the subjects in Data Set 2 have past experience playing both game levels in a study with identical experimental protocol [101]. In that study, each subject had played two sessions of the easy game and one session of the difficult game. Six weeks later, on average, Data Set 2 was conducted, hence we assume that subjects in Data Set 2 have not retained much of their experiences with the difficult game. All the subjects in Data Set 3 have had no prior experience with any of the game levels.

In Data Set 2, subjects play the following games in this particular order: Rest 1 (R1), Easy 1 (E1), Difficult (D), Rest 2 (R2), and Easy 2 (E2), each lasting one minute, where E1 and E2 are identical easy levels of the game, and R1 and R2 are rest periods when there is no game displayed on the screen and subjects rest. In Data Set 3, we switch the order of the games D and E1. Subjects play the following games in this particular order: Rest 1 (R1), Difficult (D), Easy 1 (E1), Rest 2 (R2), and Easy 2 (E2).

The reasoning behind the order of game segments is as follows: (*i*) All participants relaxed during R1 to reset and stabilize their physiological states. (*ii*) In many real world scenarios, subjects will be managing their tasks comfortably until a challenging scenario is encountered. It is therefore of interest to understand the transition from an easy game to a difficult one, i.e., from E1 to D (Data Set 2), and further the transition from a challenging scenario to a relaxing one, i.e., from D to E1 (Data Set 3). (*iii*) It is of strong interest to compare subjects' behavioral patterns in both easy games, and investigate whether or not the difficult game has any left over impact on subsequent games; and lastly (*iv*) it is of interest to study whether or not specific behavioral patterns found in D game or in the E1 game are dependent on the game order. Below, we will focus on items (*ii*)–(*iv*).

6.3 Metrics for Data Analysis

6.3.1 Physiological Metric pNN50

BVP sensory data is used here to calculate the well-known physiological heart rate metric pNN50 [79] following standard data processing techniques. This metric is computed based on temporal changes of the normal-to-normal (NN) heartbeat intervals also known as inter-beat-intervals (IBI). It is defined as the probability that the variation between consecutive NN intervals is larger than 50 ms, formulated as $\text{pNN50} = \frac{\#\text{NN50}}{\#\text{NN}}$, where #NN50 is the number of IBI differences greater than 50 ms, and #NN is the total number of all IBI differences. When subjects are engaged with a task requiring increased mental workload, pNN50 is expected to decrease [24]. We already validated this correlation with statistical analysis using recorded heart rate data (from Data Set 1, which is not reported here) as well as subjective NASA-TLX surveys [101].

Using the HR data collected in the experiments Data Set 2 (Data Set 2) and Data Set 3 (Data Set 3), pNN50 is calculated here in all game segments (including the two rest periods) for all the subjects.

6.3.2 Performance Metric

Subject's goal is to successfully play the game. For this, the subject (*a*) selects an airplane by touching the touch screen display (i.e., by putting the tip of the finger in the vicinity of an airplane), (*b*) draws a trajectory on the monitor, and (*c*) lifts the finger from the monitor. In light of these, two performance metrics are proposed: the number of finger-strokes (*# Strokes*), and the number of successful airplane assignments to the airports.

6.3.3 Behavioral Metrics

Utilizing information about how subjects play the game, we can describe metrics related to their behavior in the game. For this, we study the following behavioral metrics: *effort*, *response time*, and *reaction time delays* in drawing trajectories in the AT game[1]. Effort is defined as the amount of energy per equivalent mass one expends in order to draw a trajectory. We define response time as the finger-stroke duration time; and the time delay as the duration from the time when the subject completes a stroke until the beginning of the subsequent finger-stroke. This time delay arises mainly for decision-making purposes and when initiating physical movements of the arms, consistent with Hick-Hyman Law [55, 62] and Fitt's Law [38]. The metrics are formulated as follows:

6.3.4 Effort Metric

The instantaneous finger speed v_{ij} is calculated in pixels/second for the distance between the points $j-1$ and j in the ith finger-stroke within a time interval Δt_{ij}. Inspired from kinetic energy

[1] These metrics were formulated and studied only on Data Set 2 data in [102]. Here we present them for completeness and comparison with Data Set 3.

formulation $E_{Kinetic} = \frac{1}{2}mv^2$, we calculate the normalized total energy (TE) of finger-strokes as follows:

$$\mathrm{TE}_{g,i} = \sum_{j=1}^{N_i-1} v_{ij}^2, \tag{6.1}$$

where N_i is the number of points in trajectory i in a game-level $g = \mathrm{E1, D, E2}$. Notice that since the equivalent lumped mass m is different for each subject, and is unknown, we remove the term $\frac{1}{2}m$ from the energy calculations; in some way normalizing the energy expenditure with respect to mass. This allows us to compare TE metric across different subjects with possibly different actuation capacities commensurate with their physical capacities.

Using (6.1) next, the sum of the amount of normalized energy (SE) and mean of the normalized energy (ME) within each game-level, g, can be calculated respectively with the following formula:

$$\mathrm{SE}_g = \sum_{i=1}^{n_g} \mathrm{TE}_{g,i}, \quad \text{and} \quad \mathrm{ME}_g = \frac{\mathrm{SE}_g}{n_g}, \tag{6.2}$$

where n_g is the total number of strokes produced by the subject in a game-level g.

6.3.5 Finger-stroke Duration and Time Delay

Stroke Duration of the ith stroke is computed as the time difference between the moment a trajectory is started when the finger touches the screen and the moment the trajectory is completed. Stroke Delay Time is calculated as the duration when a subject's finger is not in touch with the screen between consecutive strokes. The mean of stroke durations (MSDur) and mean of stroke delay times (MSD) within a game-level for each subject are computed respectively with the following formula:

$$\mathrm{MSDur}_g = \frac{\sum_{i=1}^{n_g} \text{Stroke Duration}_{g,i}}{n_g}, \tag{6.3}$$

and

$$\mathrm{MSD}_g = \frac{\sum_{i=1}^{n_g-1} \text{Stroke Delay Time}_{g,i}}{n_g - 1}. \tag{6.4}$$

Notice, the formulation above consistent with Hick-Hyman Law [55, 62] and Fitt's Law [38] that we use in this section. In the next section, the above defined metrics will be calculated, and their statistical significance will be evaluated in relation to game difficulty in balanced experiments.

6.4 Results and Discussion

Among all participants in Data Set 2 and Data Set 3, within subjects statistical comparisons using a repeated measures general linear model (GLM) procedure are conducted on the physiological metric: pNN50; performance metrics: number of strokes and number of successful assignments; and behavioral metrics: effort, sum of normalized energy (SE), mean of normalized energy (ME), mean stroke duration (MSDur), and mean stroke delay (MSD), across the three game-difficulty groups (E1, D, and E2) with E1 and D game levels balanced. A value of $\alpha = 0.05$ was used to define statistical significance. Greenhouse-Geisser adjustments are examined for the above metrics distributions; the adjusted degrees of freedom are reported for those that violate the assumption of sphericity, otherwise degrees of freedom are reported as whole values. Whenever significant main effects appeared, post hoc comparisons of paired means were carried out using a least significant difference test (LSD). In this study, statistical analysis is conducted using IBM ® SPSS® version 20, following the procedures reported in [89, 113].

6.4.1 Physiological Metric: pNN50

In both Data Set 2 and Data Set 3, the values of pNN50 are significantly affected by the changes in game difficulty, $F(2, 10) = 10.667, p < 0.05$ and $F(2.886, 31.527) = 8.238, p < 0.05$ respectively, with observed powers slightly larger than 0.98, see Tables 6.1–6.2. In both Data Set 2 and Data Set 3, pair-wise comparisons reveal that pNN50 in game D is significantly lower (Figure 6.1a-6.1b) than that of the same metric for E1 and E2 ($p < 0.05$). Moreover, the metric value in E1 does not significantly differ from the same metric values for E2 ($p = 0.771$), see Tables 6.1–6.2.

Statistical analysis point out three key messages:

(A) Since pNN50 for D is smaller than that for E1 and E2, this is consistent with the fact that subjects found game D much more difficult than game E1 and E2, which is also consistent with what was observed in [101]. If we consider *only* the first two levels of the game, D and E1, we can conclude that subjects' perception of game difficulty is captured by the pNN50 metric, where the metric value is the lowest when subjects play the difficult level. Further, after the relaxed state R1, the metric pNN50 can successfully differentiate the difficult game level D from the easy game level E1, *regardless of the game order*.

(B) When all levels of the experiment are considered, we reveal that the game order has a nonlinear effect on pNN50. For example, the relaxation from D to R2 in Data Set 2, or to E1 in Data Set 3 is very pronounced, indicating it is less critical what follows D. This relaxation seems to also

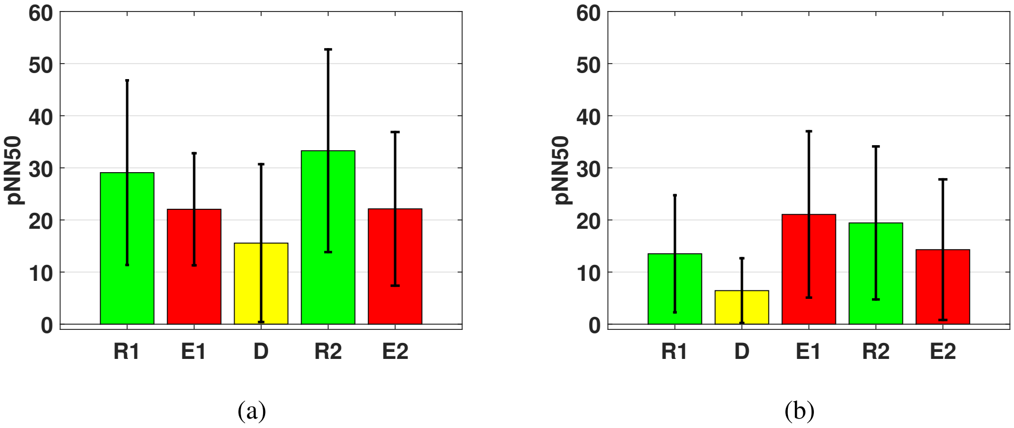


Figure 6.1: **The average values of pNN50 for each game-level across all subjects who participated in Data Set 2 (left) and in Data Set 3 (right).**
The error-bar represents the standard deviation.

present a type of overshoot, as observed from D to E1 in Data Set 3, which then settles in R2 and E2 (Figures 6.1b). This overshoot is visible in Data Set 2 as well but may not be clearly noticeable since this time D is followed by R2 (Figure 6.1a). This is valuable insight into how pNN50 interplays with game order, which, to the best of our knowledge, has not been studied in subsequent game levels. It indicates that one can use pNN50 metric to detect subject' leaving the difficult game, regardless of what the following game level is, whether an easy game or a rest period.

(C) If E1 and R2 periods in Data Set 3 were longer, we think that pNN50 metric values would be similar to those in Data Set 2, but this needs to be confirmed in future experiments.

In conclusion, as statistical analysis indicate, pNN50 shows some degree of dependency on the game order, yet it *could* be adequately used to differentiate certain game levels encountered by the subjects. Specifically, a transition from D to E1, or from E1 to D is detectable, as long as the subjects have been in a sufficiently long rest period R1 prior to playing the first game.

We note also that the above findings show consistency with respect to published work as discussed in the Introduction. However, the main reason for analyzing pNN50 here is to treat it as a baseline metric together with the performance metrics, and to study how those metrics relate to subjects' behavioral metrics.

6.4.2 Performance Metrics: Number of Strokes and Number of Assignments

The average of # of strokes (Figure 6.2a) and # of assignments (Figure 6.3) are significantly affected by changes in game-difficulty in Data Set 2 and Data Set 3 (with observed powers around 0.97), see Tables 6.1–6.2. In both sets of experiments, pair-wise comparisons reveal that # of strokes (Figure 6.2a-6.2b) and # of assignments in D (Figure 6.3a-6.3b), on average, are significantly lower than that of the same metric values for E1 and E2 ($p < 0.05$). Also the average of these two metrics for E1 do not significantly differ from the same metrics calculated for E2, see Tables 6.1–6.2.

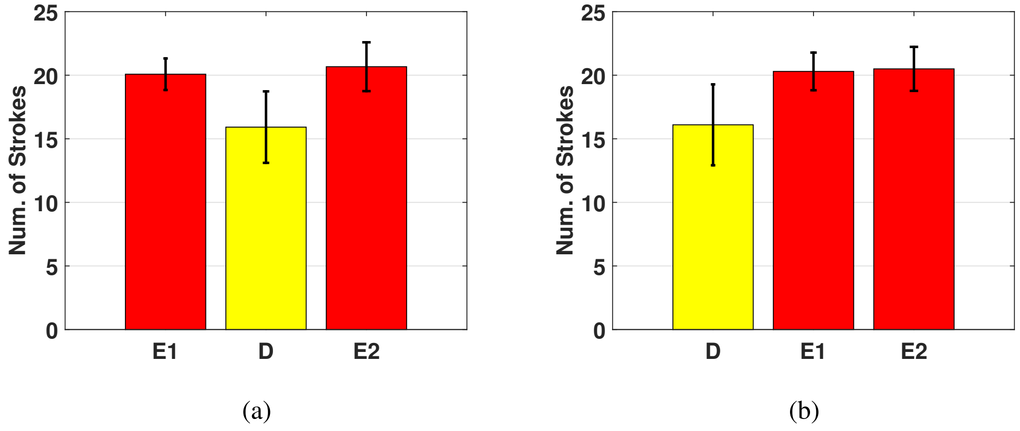


Figure 6.2: **The average values of Number of Finger-strokes for each game-level across all subjects who participated in Data Set 2 (left) and in Data Set 3 (right).**
The error-bar represents the standard deviation. Subjects rest during R1 and R2 consistent with the previous subsections. Since subjects do not interact with the touchscreen, R1 and R2 are not shown in the plots.

Consistent statistical results in Data Set 2 and Data Set 3 on the two performance metrics indicate that the subjects are challenged by the complexity of the game D, and subjects perform *consistent* playing in identical E1 and E2 games, even the two easy games are separated by either a rest period or a D game. These results have strong correlation with pNN50 metric, and even improve few of the inconsistencies detected in the pNN50 metric.

We remark here that it may sound as if the D game does not provide sufficiently many opportunities to the subjects, for assigning the airplanes to the airports, compared with E1 and E2. This is however not the case. In the E1 and E2 games, total of twenty airplanes merge into the screen

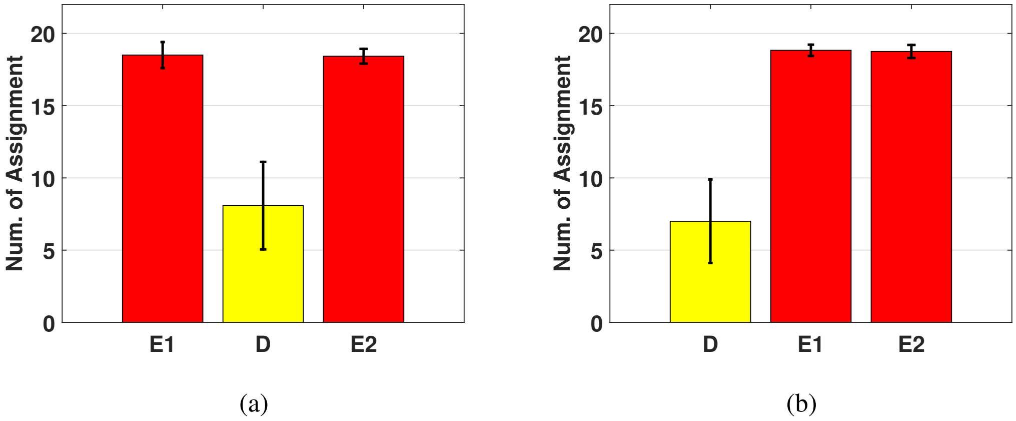


Figure 6.3: **The average values of Number of Trajectory Assignments for each game-level across all subjects who participated in Data Set 2 (left) and in Data Set 3 (right).** The error-bar represents the standard deviation. Subjects rest during R1 and R2 consistent with the previous subsections. Since subjects do not interact with the touchscreen, R1 and R2 are not shown in the plots.

in 60 sec game duration with a new airplane entering the screen every three seconds. The same arrival rate is used in the D game as well. Moreover, in the D game, whenever the subject makes a correct airplane assignment, then the color indicators switch immediately otherwise the subject has five seconds to make the subsequent correct assignment before the indicators switch colors again. Based on this setting, so long as the subject can make a correct decision every three seconds, the subject can have as many opportunities as in the E1, E2 games. Further, the AT game is designed such that in the D game there is always at least one airplane on the screen that can be assigned to the correct airport while at the same time respecting the game rules. Therefore, there is no idle time that could drop subjects' performance. Under these conditions, a very well trained player could handle correct assignments of twenty airplanes in both D and E1, E2 games. Nevertheless, the subjects are not very well trained, and due to increased cognitive load partly determined by their lack of training and inexperience in the D game, they fail to achieve this high performance as measured by the above described performance metrics.

6.4.3 Behavioral Metrics

6.4.3.1 Mean Energy (ME) Metric

In both Data Set 2 and Data Set 3, subjects' mean energy expenditure measured by ME (Figure 6.4) is significantly affected by changes in game-difficulty (with observed powers slightly larger than 0.96). Moreover, ME for the difficult game is significantly higher than that of the same metric for E1 and E2 ($p < 0.05$) on average; and the average of ME for E1 does not significantly differ from the same metric calculated for E2 in both Data Set 2 and Data Set 3, showing again consistency, see Tables 6.1–6.2.

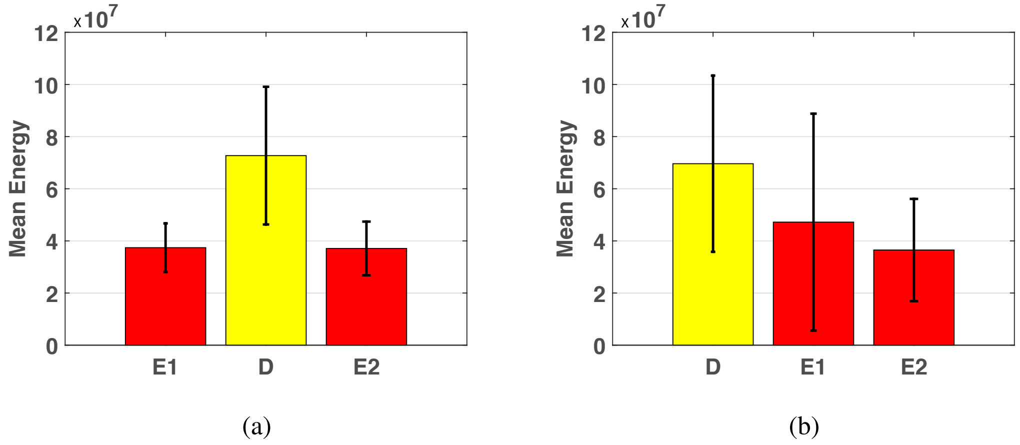
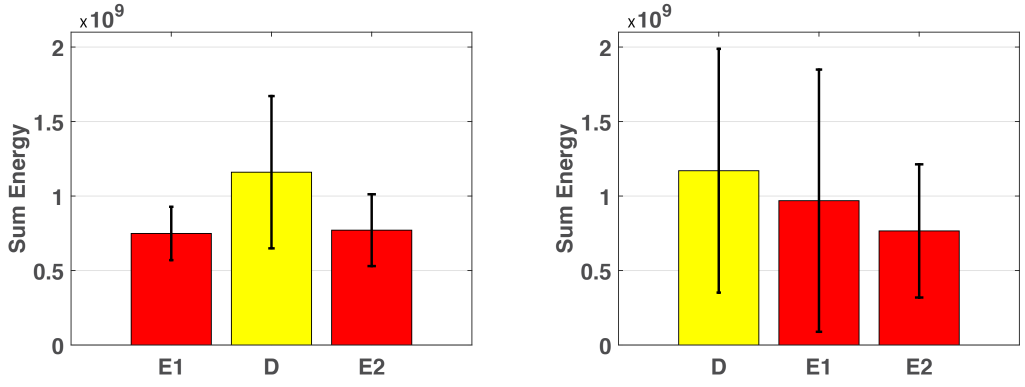


Figure 6.4: **The average values of Mean of Normalized Stroke Energy (ME) for each game-level across all subjects who participated in Data Set 2 (left) and in Data Set 3 (right).** The error-bar represents the standard deviation. Subjects rest during R1 and R2 consistent with the previous subsections. Since subjects do not interact with the touchscreen, R1 and R2 are not shown in the plots.

Comparing the results on # of finger-strokes and ME, we find out that the subjects, on average, draw *fewer* feasible trajectories but *faster* and *longer* in D than those in E1 and E2. That is, the number of trajectories in both E1 and E2 are significantly higher, but are drawn in a slower pace when compared with the difficult game. This is very likely due to the fact that the difficult level induces frustration and higher task load on the subjects, and that the subjects are less efficient in producing feasible trajectories due to the added game challenges in the difficult game.

6.4.3.2 Sum of Energy (SE) Metric

Total energy expenditures normalized by inertia (SE) are studied in Data Set 2 and Data Set 3. In contrast with ME, this metric does not render any statistically significant differences, see Tables 6.1–6.2. It is likely that subjects on "average" spent similar amount of cumulative effort/energy in drawing trajectories within different game levels, presenting some similarity in total effort put in different game levels. Nevertheless, limited statistical power on SE ($\beta = 0.439$ for Data Set 2 and $\beta = 0.627$ for Data Set 3) prevents drawing strong conclusions. Further experiments are required in future work to strengthen these discussions.



(a)

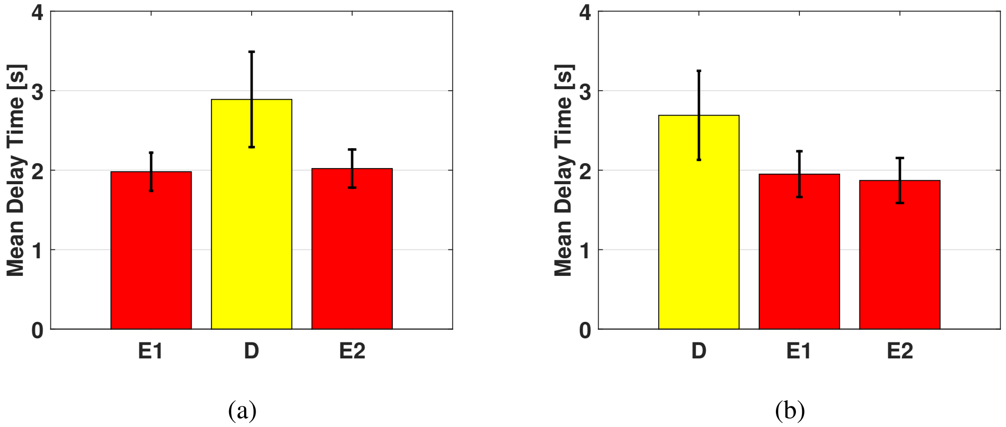
(b)

Figure 6.5: **The average values of Sum of Normalized Stroke Energy (SE) for each game-level across all subjects who participated in Data Set 2 (left) and in Data Set 3 (right).** The error-bar represents the standard deviation. Subjects rest during R1 and R2 consistent with the previous subsections. Since subjects do not interact with the touchscreen, R1 and R2 are not shown in the plots.

Based on the MSD metric however, we observe that it matters how a subject's effort is allocated temporally, and its distribution over time, as discussed next:

6.4.3.3 Mean Stroke Delay (MSD)

We find out that average values of MSD (Figure 6.6) in Data Set 2 and Data Set 3 are significantly affected by changes in game-difficulty (with observed powers $0.98 - 0.99$). Pair-wise comparisons reveal that the average of MSD in D is significantly higher (Figure 6.6) than that of the same metric in both E1 and E2 ($p < 0.05$). Moreover, the average of MSD in E1 does not significantly differ from the same metric calculated for E2 in both Data Set 2 and Data Set 3, see Tables 6.1–6.2.



(a)

(b)

Figure 6.6: **The average values of Mean Stroke Delay (MSD) for each game-level across all subjects who participated in Data Set 2 (left) and in Data Set 3 (right).**
The error-bar represents the standard deviation. Subjects rest during R1 and R2 consistent with the previous subsections. Since subjects do not interact with the touchscreen, R1 and R2 are not shown in the plots.

Based on these consistent results, we can argue that subjects' response times as a group in drawing a new trajectory significantly differ on average, across three game-levels, possibly because subjects need to be more strategic regarding how to assign the airplanes to the airports. This eventually demands more cognitive load and hence more decision making time. These results are also consistent with Hick-Hayman Law and Fitt's Law yet it is noteworthy that such a pattern is captured from subjects group behavior, and in balanced experiments.

6.4.3.4 Mean Stroke Duration (MSDur)

Statistical results on the average values of mean stroke duration (MSDur) are not consistent between Data Set 2 and Data Set 3. In Data Set 2, the average values of MSDur (Figure 6.7a) are significantly affected by changes in game-difficulty, $p < 0.05$, see Tables 6.1–6.2:Data Set 2. On the other hand in Data Set 3, MSDur (Figure 6.5b) does not differ significantly among the three game-level groups, $p = 0.157$, see Tables 6.1–6.2:Data Set 3. However, similar to SE, we observe limited statistical power on MSDur ($\beta = 0.500$ for Data Set 2 and $\beta = 0.688$ for Data Set 3), which limits the significance of the statistical comparisons conducted.

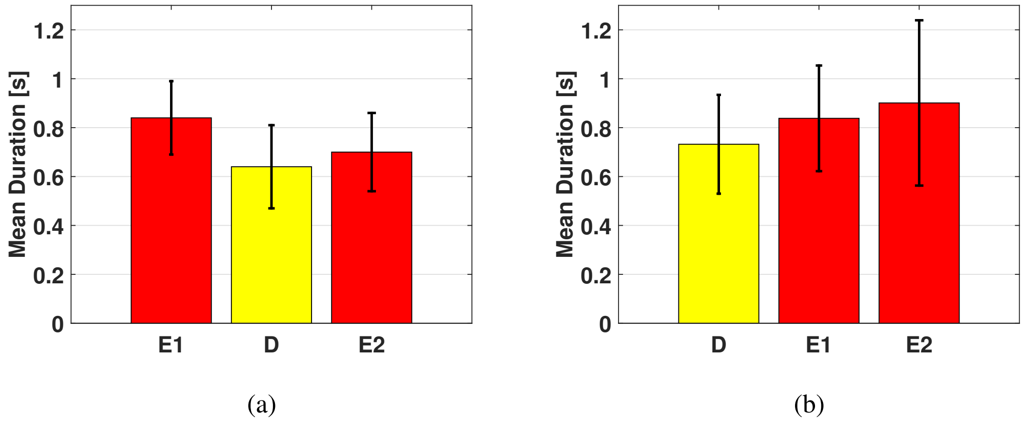


Figure 6.7: **The average values of Mean Stroke Duration (MSDur) for each game-level across all subjects who participated in Data Set 2 (left) and in Data Set 3 (right).**
The error-bar represents the standard deviation. Subjects rest during R1 and R2 consistent with the previous subsections. Since subjects do not interact with the touchscreen, R1 and R2 are not shown in the plots.

		E1			Difficult			E2		
	Metric	Mean	SD	SEM	Mean	SD	SEM	Mean	SD	SEM
Data Set 2	# Strokes	20.08	1.24	0.36	15.92	2.81	0.81	20.67	1.92	0.56
	# Assignment	18.50	0.90	0.26	8.08	3.03	0.87	18.42	0.51	0.15
	SE*	7.49	1.79	0.52	11.63	5.11	1.48	7.71	2.40	0.69
	ME*	0.37	0.09	0.02	0.73	0.26	0.07	0.37	0.10	0.03
	MSD	1.98	0.24	0.07	2.89	0.60	0.17	2.02	0.24	0.07
	MSDur	0.84	0.15	0.04	0.64	0.17	0.05	0.70	0.16	0.05
	pNN50	22.05	10.75	3.10	15.56	15.14	4.37	22.13	14.74	4.26
Data Set 3	# Strokes	20.30	1.48	0.43	16.10	3.18	0.92	20.50	1.73	0.50
	# Assignment	18.83	0.39	0.11	7.00	2.89	0.83	18.75	0.45	0.13
	SE*	9.69	8.80	2.54	11.7	8.18	2.36	7.66	4.47	1.29
	ME*	0.47	0.42	0.12	0.70	0.34	0.97	0.37	0.20	0.06
	MSD	1.95	0.29	0.08	2.69	0.56	0.16	1.87	0.28	0.08
	MSDur	0.84	0.22	0.06	0.73	0.20	0.06	0.90	0.34	0.10
	pNN50	21.05	15.96	4.60	6.44	6.21	1.79	14.3	13.49	3.89

Table 6.1: **Mean, standard deviation (SD), and standard error of the mean (SEM) for the physiological and behavioral metrics in different game-levels.**

*The values are $\times 1E08$.

	Metric	All Games			Paired Samples Test					
					D vs. E1		D vs. E2		E1 vs. E2	
		F	sig.	Observed Power	Mean Difference	sig.	Mean Difference	sig.	Mean Difference	sig.
Data Set 2	# Strokes†	$F(1.334, 14.671) = 13.811$	$p < 0.05^*$	0.968	-4.167	$p < 0.05^{**}$	-4.750	$p < 0.05^{**}$	-0.583	$p = 0.306$
	# Assignment†	$F(1.110, 12.208) = 115.288$	$p < 0.05^*$	0.999	-10.417	$p < 0.05^{**}$	-10.333	$p < 0.05^{**}$	0.083	$p = 0.754$
	SE†	$F(1.60, 11.19) = 3.791$	$p = 0.069$	0.561	–	–	–	–	–	–
	ME†	$F(1.146, 12.609) = 15.462$	$p < 0.05^*$	0.966	3.530	$p < 0.05^{**}$	3.560	$p < 0.05^{**}$	0.297	$p = 0.922$
	MSD†	$F(1.224, 13.459) = 16.080$	$p < 0.05^*$	0.978	0.908	$p < 0.05^{**}$	0.870	$p < 0.05^{**}$	-0.038	$p = 0.649$
	MSDur†	$F(1.266, 13.931) = 7.979$	$p < 0.05^*$	0.500	-0.207	$p < 0.05^{**}$	-0.066	$p = 0.345$	0.141	$p < 0.05^{**}$
	pNN50	$F(2, 10) = 10.667$	$p < 0.05^*$	0.982	-6.492	$p = 0.037^{**}$	-6.575	$p = 0.015^{**}$	-0.083	$p = 0.773$
Data Set 3	# Strokes†	$F(1.278, 14.057) = 22.276$	$p < 0.05^*$	0.997	-4.167	$p < 0.05^{**}$	-4.17	$p < 0.05^{**}$	-0.250	$p = 0.515$
	# Assignment†	$F(1.013, 11.144) = 179.213$	$p < 0.05^*$	0.999	-11.833	$p < 0.05^{**}$	-11.750	$p < 0.05^{**}$	0.083	$p = 0.339$
	SE	$F(1.310, 14.405) = 2.029$	$p = 0.175$	0.373	–	–	–	–	–	–
	ME	$F(2, 10) = 11.444$	$p < 0.05^*$	0.962	2.234	$p < 0.05^{**}$	3.302	$p < 0.05^{**}$	1.067	$p = 0.374$
	MSD†	$F(1.099, 12.090) = 21.846$	$p < 0.05^*$	0.993	0.742	$p < 0.05^{**}$	0.821	$p < 0.05^{**}$	0.079	$p = 0.105$
	MSDur†	$F(1.272, 13.995) = 2.214$	$p = 0.157$	0.312	–	–	–	–	–	–
	pNN50†	$F(2.866, 31.527) = 8.238$	$p < 0.05^*$	0.981	-14.617	$p < 0.05^{**}$	-7.858	$p = 0.019^{**}$	6.758	$p = 0.076$

Table 6.2: **Statistical differentiation.**

Results of the general linear model repeated measures analysis, and post hoc comparisons of paired means using a least significant difference are presented.

The $\alpha = 0.05$ level is chosen for statistical analysis on physiological and behavioral metrics.

† indicates that the assumption of sphericity is violated. Adjusted Greenhouse-Geisser degrees of freedom is thereby reported.

* indicates that significant difference is found among all three groups feature values.

** indicates that the difference between averages is significant in pair-wise analysis.

6.5 Summary

This chapter is on the study of a series of behavioral metrics pertaining to players' touch behavior and decision making times in a touch-screen game and in relation with baseline metrics, namely, task performance metrics and heart rate variability metric pNN50. Specifically, we answer (*i*) how well subjects play the game based on successful completion of tasks, (*ii*) how much effort/energy subjects put in drawing trajectories on the touch screen where the game is presented, and (*iii*) how a subject's response time/reaction time delays are affected by game difficulty.

Two sets of experiments are conducted to investigate these metrics. Based on experimental data obtained from 12 subjects in the first set, the average values of pNN50 computed using the recorded heart rate signal, and further the average values of the performance metrics, namely, # strokes, and # of assignments; and behavioral metrics, namely, ME, and MSD showed statistically significant variations across easy and difficult levels of the game, with no differences between identical easy games. In the second set, another 12 subjects participated; this time in balanced experiments between easy and difficult game levels. The results were consistent with what was obtained in the first set of experiments, concluding that the metrics ME and MSD well correlate with performance metrics (# strokes, # of assignments) as well as with pNN50. Therefore, regardless of the order of the game difficulty, ME and MSD metrics could be reliably used to infer task load changes, which can arise by a number of reasons including lack of training.

Results of the study here provide evidence that behavioral metrics especially ME and MSD can be utilized in future studies to differentiate between difficult and easy tasks, and to detect mental workload changes, and thereby subjects' inexperience within the difficult game; possibly via real time statistical inferences, and even across different tasks. Finally, additional experiments are needed to strengthen the statistics on SE and MSDur metrics in order to draw more reliable conclusions.

Chapter 7

RELATING SUBJECTS' BEHAVIORAL PATTERN TO GAME INEXPERIENCE USING CLASSIFICATION ALGORITHMS

7.1 Introduction

In Chapter 6, a series of behavioral metrics are studied. These metrics pertain to players' touch behaviors and decision making times in the game, and are in close relation with baseline metrics, namely, task performance metrics and heart rate variability metric pNN50.

Based on experimental data obtained from Data Sets 2–3, the results of statistical analysis provide evidence that behavioral metrics Finger-stroke Energy (ME) and Stroke Delay Time (MSD) can be utilized to differentiate between difficult and easy tasks, and to detect vulnerability, possibly via real time statistical inferences, and even across different tasks. Since strong statistical results were not obtained on Stroke Duration metric (MSDure), we could not conclude whether or not this metric can be used to infer task load changes.

Having showed the reliability of the behavioral metrics in differentiating the effect of mental workload increase on subjects' behavioral patterns as manifested by inexperience, in this chapter, we use the the finger-stroke metrics from Data Sets 2–3 in order to create a "person-independent model" aiming to distinguish subject's mental workload increases in real-time. Further we would like to investigate the validity of this fitted model on a set of new subjects (Data Set 4) to find out how well this fitted model could reach similar results in differentiating between difficult and easy tasks, and to potentially detect game inexperience.

This chapter is organized as follows: (*i*) First, we introduce the (behavioral metrics) variables used to train a model. (*ii*) This is followed by listing the learning methods we use to train the finger-stroke data. (*iii*) The analysis which lead us to finding the best fitted model comes next. (*iv*) Further, we put the model to test in real-time using Data Set 4 to investigate the validity of the fitted model. (*v*) We improce the accuracy of the model taking advantage of the variability in the classifier.

7.2 Behavioral Metrics Classification

The finger-stroke data extracted from Data Sets 2–3, are selected for the model fitting (training data). The data includes three variables (features): X_1 = Finger-stroke Energy, X_2 = Stroke Delay Time, X_3 = Stroke Duration, measured during the whole game segment (60 secs), see Chapter 6. Due to lack of statistical power in statistical analysis, we could not conclude whether or not the Stroke Duration can differentiate between difficult and easy tasks along with Finger-stroke Energy and Stroke Delay Time. Here, we would like to investigate the effect of Stroke Duration feature in training a model when compared with the case where this feature is excluded and only Finger-stroke Energy and Stroke Delay Time data are used.

The training data consists of 1289 rows, which is divided into 2 groups: (*a*) 360 rows of data that are extracted from the difficult game, labeled *Difficult*. (*b*) 929 rows of data extracted from both E1 and E2 games, labeled *Easy*[1]. Prior to the analysis, the X_1 feature is normalized where all the values in X_1 is divided by the $1E09$ to have all the X_1 values lie between 0 and 1. For behavioral pattern inference, MATLAB 2014b is used to train classifiers (model fitting) using various learning methods, namely, discriminant analysis, naive Bayes classifiers, decision trees, and support vector machine (SVM), which are briefly explained next:

Discriminant Analysis Discriminant analysis assumes that different classes generate data based on different Gaussian distributions. To train (create) a classifier, the fitting function estimates the parameters of a Gaussian distribution for each class. To predict the classes of new data, the trained

[1] The finger-stroke data extracted from E1 and E2 showed they are not statistically significant by different. Hence, we combined the two data sets.

classifier finds the class with the smallest misclassification cost [37]. We used four different types of discriminant analysis, namely, **Linear Discriminant Analysis (LDA)**, **Quadratic Discriminant Analysis (QDA)**, **Diagonal Linear Discriminant Analysis (DLDA)**, and **Diagonal Quadratic Discriminant Analysis (DQDA)**. DLDA and DQDA are similar to LDA and QDA, but with diagonal covariance matrix estimates. These diagonal choices are specific examples of a naive Bayes classifier [86], as described next, because they assume the variables are conditionally independent given the class label.

Naive Bayes classifiers are among the most popular classifiers. While the assumption of class-conditional independence between variables is not true in general, naive Bayes classifiers have been found to work well in practice on many data sets [96].

The naive Bayes classifies data in two steps: (*a*) training step in which by using the training data, the method estimates the parameters of a probability distribution, assuming predictors are conditionally independent given the class. (*b*) in prediction step, for any unseen test data, the method computes the posterior probability of that sample belonging to each class. The method then classifies the test data according the largest posterior probability.

The training step in naive Bayes classification is based on estimating $P(X|Y)$, the probability or probability density of predictors X given class Y. The naive Bayes classification model provides support for various distributions. In this study we used the following distributions to classify the finger-stroke data: (*i*) **Normal (Gaussian) Distribution (NB)** is appropriate for predictors that have normal distributions in each class. For each predictor we model with a normal distribution, the naive Bayes classifier estimates a separate normal distribution for each class by computing the mean and standard deviation of the training data in that class. (*ii*) **Kernel Distribution (KD)** is appropriate for predictors that have a continuous distribution. It does not require a strong assumption such as a normal distribution and can be used in cases where the distribution of a predictor may be skewed or have multiple peaks or modes. It requires more computing time and more memory than the normal distribution. For each predictor we model with a kernel distribution, the naive Bayes classifier computes a separate kernel density estimate for each class based on the training data for that class. By default the kernel is the normal kernel, and the classifier selects a width automatically for each class and predictor. The software supports specifying different kernels for each predictor, and different widths for each predictor or class [64].

Decision Tree (DT) is a commonly used machine learning technique that uses a divide-and-conquer approach to classify testing data [15]. In other words, a decision tree is a set of simple rules. Decision trees are also nonparametric because they do not require any assumptions about the distribution of the variables in each class. During the learning stage, the tree structure is constructed with internal nodes and leaves. Internal nodes represent the test conditions while the leaves represent the classification results. When the Decision Tree is being constructed, the most informative feature (with a higher information gain) will be used near the root [27].

Support Vector Machine (SVM) is a classifier that performs classification by constructing a high-dimensional hyper-plane [17]. The constructed high-dimensional hyper-plane is optimized to separate the testing data into two classes. SVM also allows different types of kernel functions to transform testing data points into a higher dimensional space and make the transformed data easier to be classified. Since SVM has recently become a popular machine learning technique for classification, we are interested in investigating how well it classifies our training data, and then if selected as the best fitted model, what its performance will be in our testing data set (Data Set 4).

7.2.1 Model Fitting Analysis on Data Sets 2–3

The classification analysis procedure followed to train a person-independent model is explained next:

(*a*) The finger-stroke features from Data Sets 2–3 are first classified using all learning algorithms summarized above, and then for each algorithm the resubstitution error on the training set is computed. Resubstitution error is the misclassification error which is the proportion of misclassified observations to the total number of observations.

(*b*) For each learning algorithm, the confusion matrix on the training set is also constructed. A confusion matrix contains information about known class labels and predicted class labels. Generally speaking, the (i,j) element in the confusion matrix is the number of samples whose known class label is class i and whose predicted class is j. The diagonal elements represent correctly classified observations.

(*c*) For each classification analysis, the true test error (generalization error), which is the expected prediction error on an independent set is calculated[2]. This testing is very important

[2] Notice, the re-substitution error will likely underestimate the test error.

as it shows the generalization capability of the stroke features fitted model. To this purpose, a well-known method "stratified 10-fold cross-validation" is applied to estimate the test error on classification algorithms. It randomly divides the training set into 10 disjoint subsets. Each subset has roughly equal size and roughly the same class proportions as in the training set. One subset is removed, the classification model is trained using the other nine subsets and used to classify the removed subset.

(*d*) Finally, the fitted model with the lowest generalization error is selected to be tested on Data Set 4.

The classification accuracy of each of the classifiers described in Section 7.2 is evaluated using two types of feature combinations: (*i*) the first type of feature combination includes Finger-stroke Energy, and Stroke Delay Time (X_1, and X_2); and with the goal of investigating the effect of Stroke Duration feature in training a model (*ii*) the 2nd type of feature combinations includes this feature along with Finger-stroke Energy, and Stroke Delay Time (X_1, X_2, and X_3).

Table 7.1 lists the re-substitution error for each algorithm on the training set (Data Sets 2–3) computed using two types of feature combinations. We found that the two lowest re-substitution errors are computed using the SVM and DT classifiers (0.0613, and 0.0628 respectively) with the all feature combinations (X_1, X_2, and X_3).

Features	**LDA**	**QDA**	**DLDA**	**DQDA**	**NB**	**KD**	**DT**	**SVM**
X_1, X_2	0.2196	0.2048	0.2064	0.2141	0.2063	0.1722	0.0791	0.0613
X_1, X_2, X_3	0.2149	0.1846	0.2118	0.1994	0.1994	0.1512	0.0628	0.0613

Table 7.1: **Re-substitution error for all classifiers and different feature combinations.** Re-substitution error is the misclassification error which is the proportion of misclassified observations to the total number of observations, with X_1 = Finger-stroke Energy, X_2 = Stroke Delay Time, and X_3 = Stroke Duration.

On Table 7.2, further information is provided about how well the fitted model using a given learning algorithm correctly classifies total observations (360 rows of data that are labeled *Difficult*, and 929 rows of data are labeled *Easy*). For example, of the 1289 training observations, 6.12% or 79 observations are misclassified by the SVM when all the features, X_1, X_2, and X_3, are included.

		Classification Algorithm															
		LDA		QDA		DLDA		DQDA		NB		KD		DT		SVM	
features		D	E	D	E	D	E	D	E	D	E	D	E	D	E	D	E
X_1, X_2	D	99	261	144	216	110	250	143	217	143	217	178	182	298	62	282	78
	E	22	907	48	881	26	903	49	880	49	880	40	889	40	889	1	928
X_1, X_2, X_3	D	113	247	168	192	115	245	147	213	147	213	197	163	322	38	286	74
	E	30	899	46	883	28	901	44	885	44	885	32	897	43	886	5	924

Table 7.2: **Confusion matrix for the combination of all classifiers and different feature combinations.**

The grey cells highlight correctly recognized instances (true-positives). $X_1 =$ Finger-stroke Energy, $X_2 =$ Stroke Delay Time, and $X_3 =$ Stroke Duration.

Since the re-substitution error will likely overestimate the accuracy of a given algorithm, to select the best fitted model, we are more interested in the test error (generalization error) of the learning algorithm, which is the expected prediction error on an independent set. In this case, since we do not have another labeled data set where all the finger-stroke data are known and labeled, a stratified 10-fold cross-validation is chosen for estimating the test error on classification algorithms. Table 7.3 demonstrates the setup to to generate 10 disjoint stratified subsets for 10-fold cross-validation analysis.

Num Observations	1289									
Num Test Sets	10									
Train Size	1161	1160	1160	1160	1160	1160	1160	1160	1160	1160
Test Size	128	129	129	129	129	129	129	129	129	129

Table 7.3: **A stratified 10-fold cross-validation setup.**

The setting is used to estimate the test error on classification algorithms.

Figure 7.1 plots the results of using 10-fold cross validation. For all classifiers, excluding Stroke Duration data, X_3, reduces in the classification accuracy. This provides evidence that Stroke Duration, along with Finger-stroke Energy and Stroke Delay Time, help the classification accuracy in differentiating between difficult and easy tasks. We also found that the best classification accuracy (83.90%), is obtained from using the KD classifier with the all feature combination (X_1, X_2, and X_3). The 2nd and 3rd best classification accuracies are for QDA and the DT classifiers (81.07%, and 80.76% respectively), see Table 7.4. Notice that the structure of the decision tree uses the energy

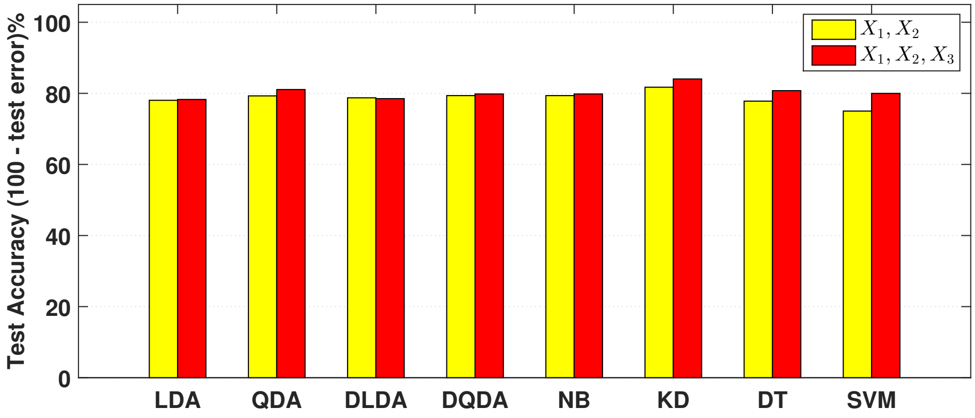


Figure 7.1: **Accuracy of the 10-fold cross validation of all classifiers using different feature combinations.**
$X_1 =$ Finger-stroke Energy, $X_2 =$ Stroke Delay Time, and $X_3 =$ Stroke Duration.

of the x-axis from the Stroke Delay Time data, X_2, as the root test condition, see Figure 7.2. It proves that *touch delay* information provides higher information gain in the decision tree learning stage. Table 7.2 shows more details of the cross validation results on 1200 samples. The grey cells highlight correctly recognized instances (true-positives).

Features	**LDA**	**QDA**	**DLDA**	**DQDA**	**NB**	**KD**	**DT**	**SVM**
X_1, X_2	78.05	79.29	78.74	79.36	79.36	81.82	77.81	75.02
X_1, X_2, X_3	78.28	81.07	78.51	79.83	79.83	83.90	80.76	79.98

Table 7.4: **Accuracy of all classifiers using different feature combinations.**
The accuracy is simply calculated using (1− Cross Validation Error) ×100.

7.3 Re-construction: Validating the Fitted Models on New Experimental Data

Having computed the generalization error of all learning algorithms using all combinations of the three finger-stroke features, KD is found to be the best fitted model on Data Sets 2–3, with the least amount of error in classifying the touch behavior data, see Section 7.2.1. In this Sec-

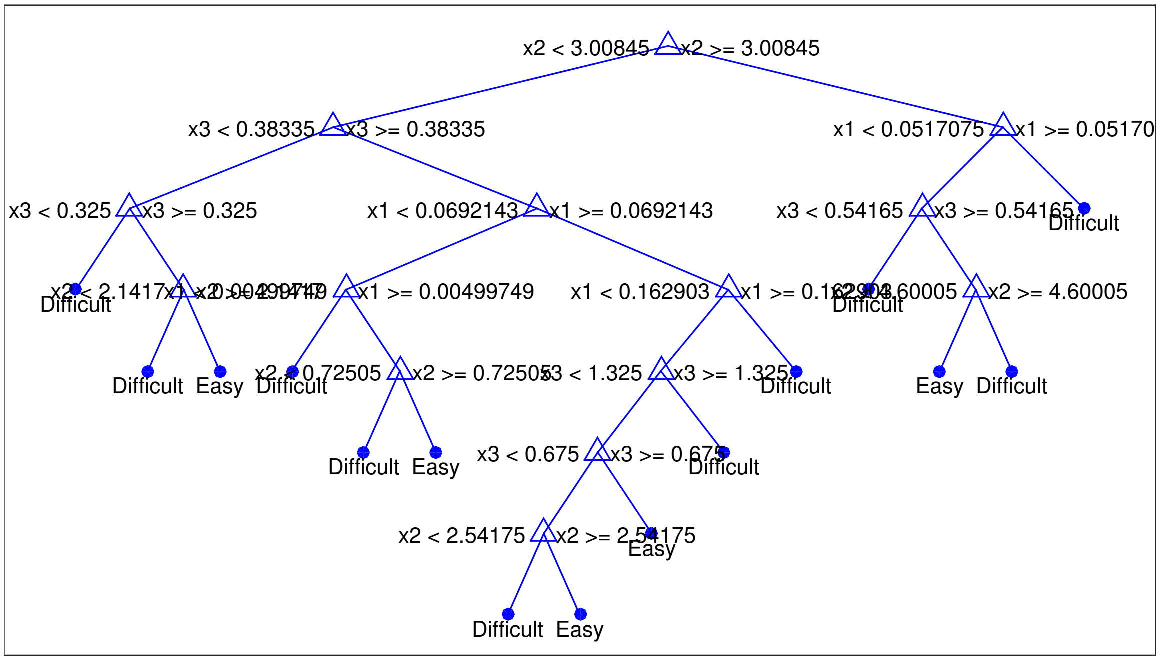


Figure 7.2: **The structure of the decision tree classifier using all three features.** X_1 = Finger-stroke Energy, X_2 = Stroke Delay Time, and X_3 = Stroke Duration. The decision tree (DT) classifier the way it is constructed, uses the energy of the x-axis from the Stroke Delay Time data, X_2, as the root test condition.

tion, the results of the prediction analysis using the KD classifier is presented, in order to answer ”how well the this classifier, as a person-independent model, could distinguish the different levels of the game with different subjects?”

7.3.1 Adaptive Game: Implementing the KD Classifier in the Game

The trained KD classifier, calibrated with Data Sets 2–3 by using the three features, X_1, X_2, and X_3, is put to test in Data Set 4, see Sections 3.3–3.4, in order to study its validity. With this purpose, the KD classifier is implemented in the game which is used to conduct a new set of experiments (Data Sets 4) and perform real-time classification as subjects play the game. That is, this new version of the game is capable of *predicting* whether the subject plays the Difficult game or the Easy one, using the touch behavior metrics. Specifically, after touch behavior metrics of each finger-stroke are calculated first, and then these metrics are passed into the KD classifier in order to predict what game level the subject is playing. Notice, since one of the metrics for the classification is the *delay time* in between two consecutive finger-strokes, X_2, the learning algorithm is triggered after the *second* finger-stroke is made by the player. All the KD classifier predictions along with their corresponding touch behavior metrics, and the expected/desired known labels of each (easy or difficult) are recorded for further analysis.

7.3.2 Experimental Protocol

11 Subjects participated in our study (Data Set 4), see the experimental protocol details in Sections 3.3–3.4. Notice that none of the subjects have had any prior experience with any of the games. Subjects play the game for ten trials. In one trial, the subjects play one easy, and one difficult level of the game. The order of the games are randomly selected, and the subjects are not made aware of what the upcoming game level will be. Since in the experimental setup there is no resting time considered, the subjects are therefore instructed to rest in between consecutive trials in case they need to relax.

All the KD classifier predictions, and the expected/desired known labels are recorded for further analysis. The labels are simply the game level in which finger-strokes are made. In the end, we have a two column matrix, in which one column lists the predictions (E or D), and the other column holds the expected prediction of the game level (E or D).

7.3.3 Analysis of Experimental Results when KD is Implemented in the Game

Having recorded all the predictions made by the KD classifier, we next (*a*) investigate the validity of the KD classifier tested on Data Set 4, and further, (*b*) calculate how well this fitted model using Data Sets 2–3, could reach similar results in differentiating between difficult and easy tasks, and to detect subjects' inexperience in real-time. Specifically, we are interested to find out how accurate the predictions are. In other words, does the prediction match with the game label or not?

Figures 7.3–7.4 demonstrate the KD prediction accuracy for the two sample subjects across all trials played our game. The prediction accuracy clearly differs for the two subjects. We find out that for subject 1 (Figure 7.3) the overall prediction accuracy is 82.18%, whereas for subject 4 (Figure 7.4) the KD classifier could correctly distinguish the levels of the game by 58.99%. Overall, the performance of the implemented KD classifier for all the easy games is 94.27% and 98.33%, respectively for Subjects 1 and 4. However, in all difficult games, the prediction accuracy is not as good: it reduces to 67.30% for Subject 1, and further it becomes worse for Subject 4 when it could correctly predict 18.75% of all the finger-strokes made across all the difficult game.

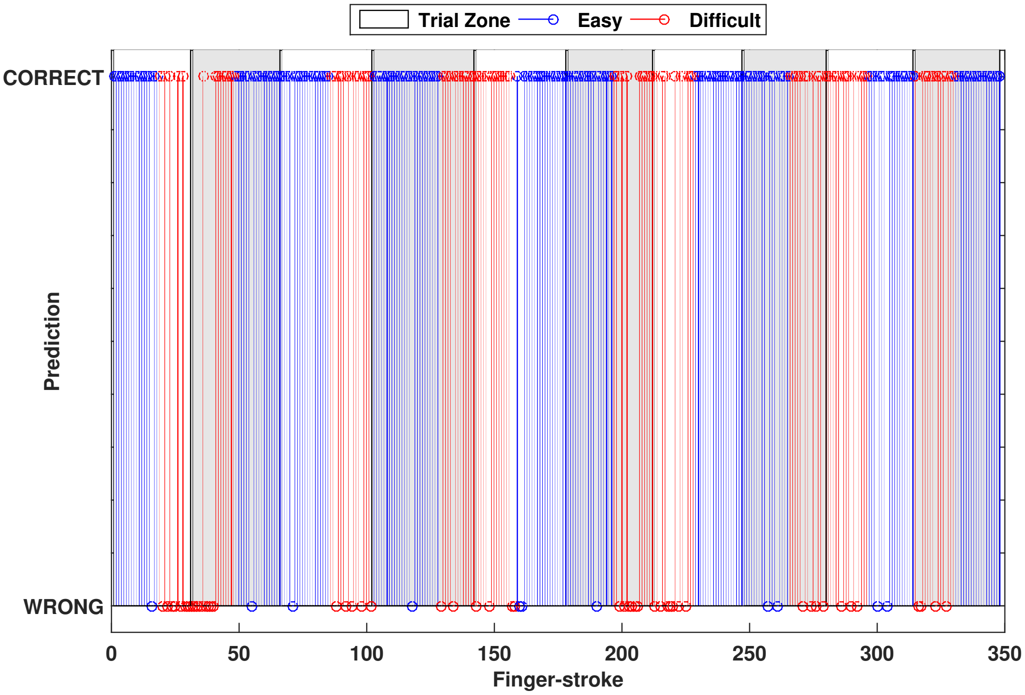


Figure 7.3: **The performance of KD Classifier for subject 1 in Data Set 4.**
CORRECT: when the prediction result *do* matches with the game level. WRONG (incorrect): when the prediction result does not matches with the game level.

Figure 7.5 depicts the overall performance of the KD classifier in Data Set 4. As depicted in the figure, the performance of the fitted model drops significantly in the difficult games when compared with the easy game. On average, considering both game levels, KD classifier could correctly classify the finger-stroke data by 65.67% ± 6.23%, Figure 7.6. The performance of KD algorithm in Data Set 4 is lower than its accuracy level, 83.90% obtained when calibrated using Data Sets 2–3, see Section 7.2.1. Moreover, for all the easy games, the fitted model, KD, correctly distinguishes the finger-stroke data of Data Set 4 by 95.94% ± 2.38%, on average, Figure 7.6. The performance degrades significantly to 32.62%±12.93%, on average, in all the difficult game, Figure 7.6.

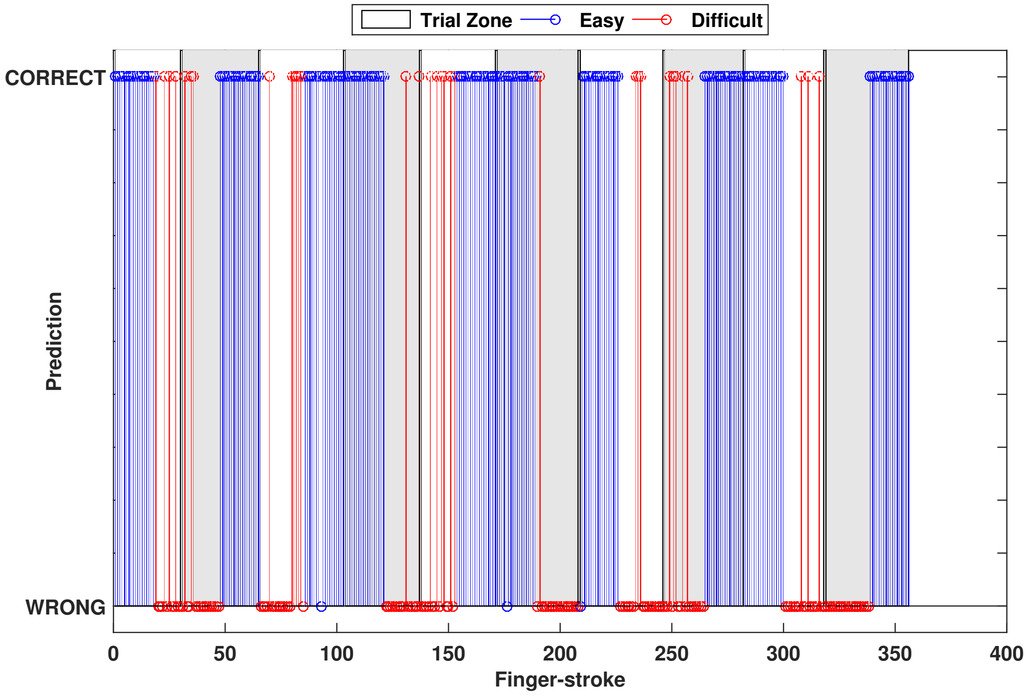


Figure 7.4: **The performance of KD Classifier for subject 4 in Data Set 4.** CORRECT: when the prediction result *do* matches with the game level. WRONG (incorrect): when the prediction result does not matches with the game level.

7.3.4 Re-construction of Data Set 4 Using all Classifiers Calibrated on Data Sets 2–3

As the KD classier did a poor performance in re-construction of the finger-stroke data for the difficult level in Data Set 4, we decided to test the rest of the classifiers trained and calibrated using the Data Sets 2-3.

Since we have recorded all the features corresponding to the stroke data in Data Set 4, we could simply feed those features offline into our classifier to predict whether the trajectory is drawn in the easy game or the difficult level. This process, in fact, is similar to implementation of the KD classier in the game: the finger-stroke data features, $X_1, X_2, X3$, is being fed to the classifies one after each other and then classifier predicts the output.

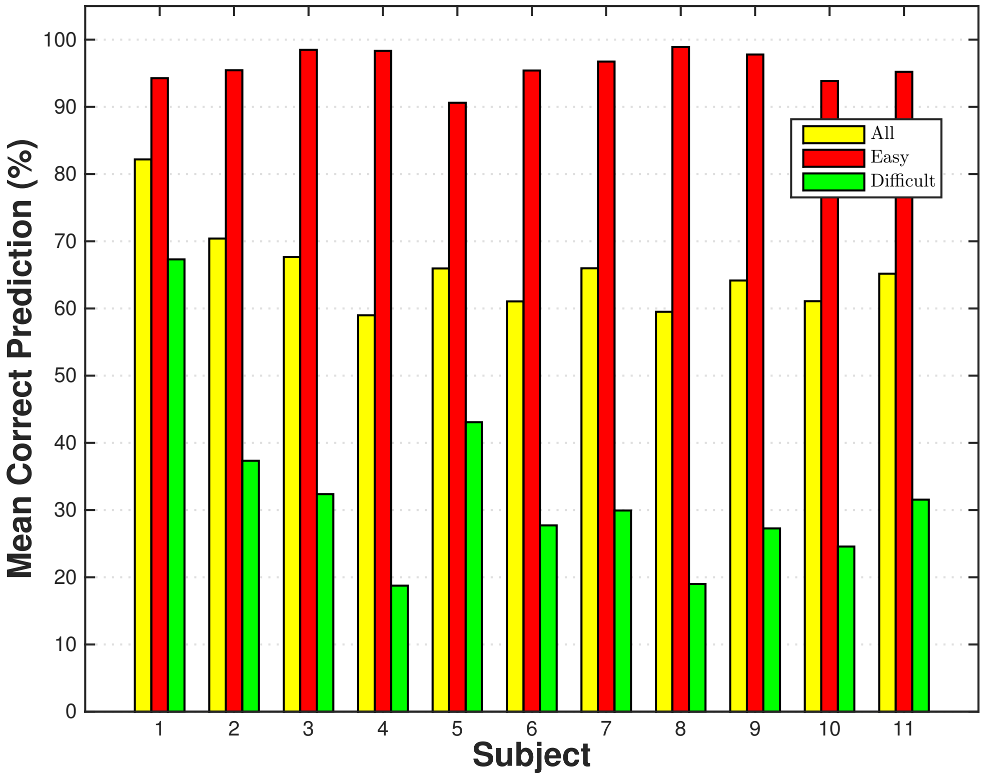


Figure 7.5: **The overall performance of KD Classifier in Data Set 4 across all subjects.** ALL: Average of the KD classification accuracy across all easy and difficult games. EASY: Average of the KD classification accuracy across all easy games. DIFFICULT: Average of the KD classification accuracy across all difficult games.

7.3.5 Analysis of Experimental Results Comparing all the Classifiers on Data Set 4

All the predictions of each classifier are recorded for further analysis. Figure 7.7 depicts the overall performance of all classifiers, on average, in Data Set 4. As depicted in the figure, the sensitivity of all the classifier models drops significantly in the difficult game when compared with the easy level. On average, the SVM could correctly classify the finger-stroke data with an accuracy of $69.75\% \pm 4.18\%$. Next are, DT, NB (and DQDA), and KD with $67.07\% \pm 4.04\%$, $66.34\% \pm 4.29\%$, and $65.67\% \pm 6.23\%$ accuracy, respectively, see Table 7.5. The sensitivity of all the classifiers, similar to KD, are not as well as their performance in the easy games. In Data Set 4, DT did the best in correctly classifying the finger-stroke data from the difficult game with

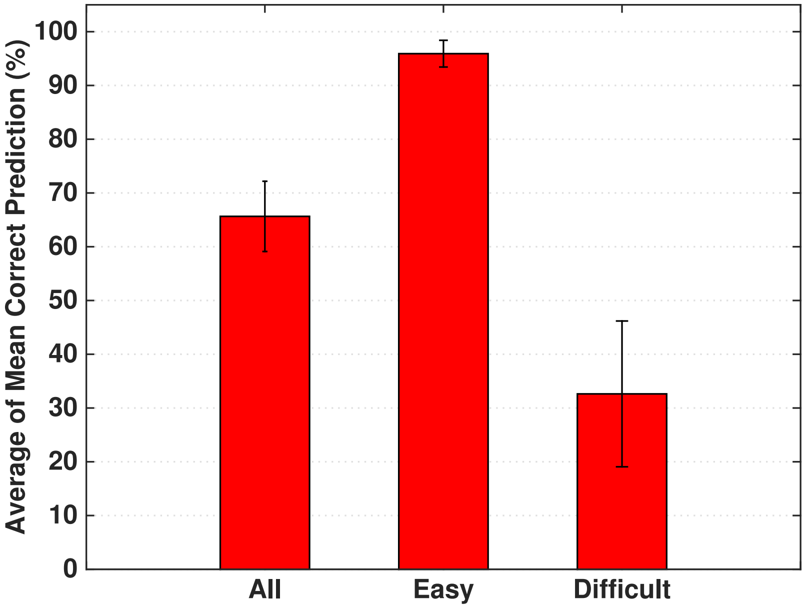


Figure 7.6: **The overall performance of KD Classifier in Data Set 4, on average, across all subjects.**

ALL: Average of the KD classification accuracy across all easy and difficult games. EASY: Average of the KD classification accuracy across all easy games. DIFFICULT: Average of the KD classification accuracy in all difficult games.

an accuracy of $46.97\% \pm 8.83\%$. Next, are SVM, NB (and DQDA), and KD with 44.85 ± 8.63, 34.17 ± 9.13, and 32.62 ± 12.93 accuracy, respectively, see Table 7.5.

The experimental results on the sensitivity of all the classifier in Data Set 4 reveals that all calibrated models can reliably classify the finger-strokes data in the easy game. However, in the difficult game, the sensitivity of the classifiers in distinguishing the game level is not satisfactory. That is, further analysis is required in order improve the fitted model accuracy level when tested in the difficult game. In the next section we address this in more details.

7.3.6 Re-configuration: Improvements based on Classifiers Variability

In order to improve the performance of the fitted models, especially in the difficult game, we need to study the prediction data in further details. With this aim, we study the finger-stroke classification results in two different ways:

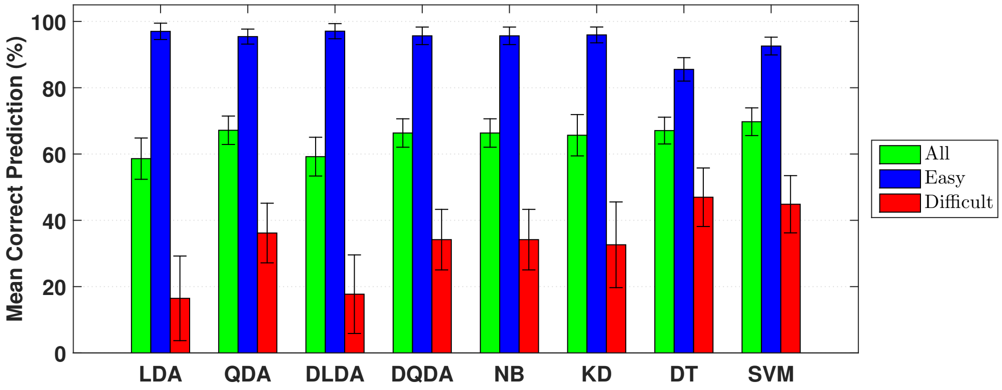


Figure 7.7: **The overall performance of all Classifier in Data Set 4, on average, across all subjects.**

ALL: Average of the classification accuracy across all easy and difficult games. EASY: Average of the classification accuracy across all easy games. DIFFICULT: Average of the classification accuracy in all difficult games.

Fitted Model	All	Easy	Difficult
LDA	58.61 ± 6.27	97.01 ± 2.46	16.46 ± 12.76
QDA	67.17 ± 4.28	95.42 ± 2.26	36.17 ± 8.99
DLDA	59.21 ± 5.85	97.06 ± 2.28	17.71 ± 11.85
DQDA	66.34 ± 4.29	95.66 ± 2.64	34.17 ± 9.13
NB	66.34 ± 4.29	95.66 ± 64	34.17 ± 9.13
KD	65.67 ± 6.23	95.94 ± 2.38	32.62 ± 12.93
DT	67.07 ± 4.04	85.53 ± 3.54	46.97 ± 8.83
SVM	69.75 ± 4.18	92.57 ± 2.68	44.85 ± 8.63

Table 7.5: **Accuracy of all classifiers using three features,** $X_1, X_2, X3$, **in Data Set 4.**

The accuracy is simply the ratio of the number of predictions match with the desired output to the total predictions. ALL: Average of the classification accuracy across all easy and difficult games. EASY: Average of the classification accuracy across all easy games. DIFFICULT: Average of the classification accuracy in all difficult games.

Case 1: We *know* the desired/expected label for each prediction made by the classifier. Therefore we have a clear understating whether or not a classification result (prediction) is correct or not. With this knowledge, we investigate "how often an incorrect prediction is made within a game?" In other words, we aim to find the frequency of the occurrence of incorrect predictions.

In this case, we are aware of the desired predictions, which are the game levels. Therefore, we can simply construct a list, namely *prediction accuracy*, in which 0 represents an *incorrect* prediction, and 1 represents a *correct* one. To answer the above question, we first construct the prediction accuracy lists for all the games across all the subjects. Next, the number of incorrect predictions are counted. Then, we record how many times this many incorrect predictions happen across all games, regardless of the order of incorrect predictions.

We should note that in this analysis, we only consider the first 5 predictions (the first 6 finger-strokes) within a game. That is, the number of incorrect predictions within each game range from 0, meaning all predictions are correct and match with the game level; to 5, where all the predictions results are incorrect.

Case 2: We *do not know* the desired/expected label for each prediction made by the classifier. Here, we are investigating *how consistent* the prediction results are within a game level. In other words, we would like to answer: "how often the prediction results switch from easy to difficult, or from difficult to easy within a game?" In this analysis we consider (*2a*) all the finger-strokes in a game, and (*2b*) the first 5 predictions (the first 6 finger-strokes) within a game.

For the both cases above, Case 2a and Case 2b, we count the number of times the consecutive prediction results are switched within a game. Next, the occurrence of this many prediction switches are recorded across all specific game levels.

Notice that for the Case 2b analysis, as we only consider the first 5 predictions (the first 6 finger-strokes) within a game, the number of prediction switches range from 0, meaning all predictions are consistent; to 4, where every other prediction results are switching.

The overarching goal of all the study cases mentioned above, i.e., *Case 1*, *Case 2a*, and *Case 2b*, is to understand whether the classification results on Data Set 4 are following known distributions or not, and whether the predictions results in the easy games are different than those of in the difficult game. That is, we have more information in hand to check how likely each prediction is correct.

All the analysis are conducted using Statistics Toolbox (V9.1) in MATLAB R2014b. The goodness of fit in distribution fitting is determined using the Bayesian information criterion (BIC)[3],

[3] or Schwarz criterion (also SBC, SBIC). It is based, in part, on the likelihood function and it is closely related to the Akaike information criterion (AIC).

where the model with the lowest BIC is preferred [117].

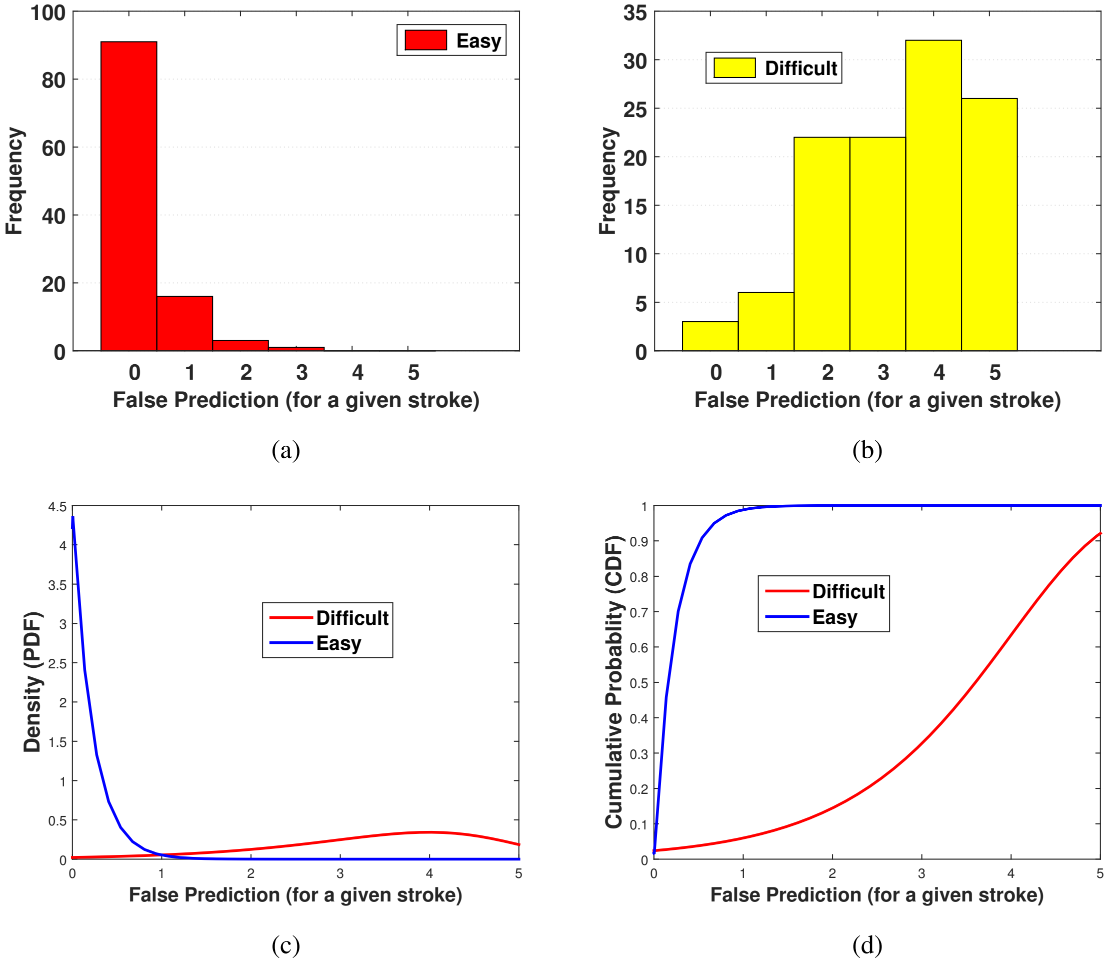
7.3.7 Improving the Re-configuration Analysis: Results and Discussion

Here, we present the results of the analysis mentioned in the previous section, when implemented on SVM, DT, KD, and DQDA classifiers when calibrated in Data Sets 2–3 using X_1, X_2, X_3 features.

Case 1: The frequency of the occurrence of incorrect predictions in the first five prediction results across all easy games, using KD, DQDA, DT, and SVM (Figures 7.8a, 7.9a, 7.10a, and 7.11a, respectively) are different than in all the difficult game levels (Figures 7.8b, 7.9b, 7.10b, and 7.11b, respectively).

For example, using KD algorithm, across all easy game levels in Data Set 4, we observe 91 times that all the first 5 prediction results are correct, Figure 7.8a. This number degrades significantly across all difficult games where for only 3 times all the first five KD classification results predicted the difficult game level correctly, Figure 7.8b. Moreover, when DQDA model is used, for 22 times, one incorrect classification result are observed in all easy games, Figure 7.9a whereas in all difficult games there are only 9 times that one incorrect prediction exists among the first five classification results, Figure 7.9b. Finally, when DT is used, in all easy games, we observe three incorrect prediction results only three times within the first five predictions, Figure 7.10a. This number increased dramatically across all difficult games where we observe the occurrence of three incorrect predictions 32 times, Figure 7.10b.

The occurrence of the incorrect classification results across all easy games follow an exponential distribution ($\mu = 0.23$, $\mu = 0.22$, $\mu = 0.7$, and $\mu = 0.34$, respectively for KD, DQDA, DT, and SVM), whereas the occurrence of the incorrect prediction results across all difficult games when KD and SVM are used, follow the Extreme Value distribution ($\mu = 3.99, \sigma = 1.07$; and $\mu = 3.36, \sigma = 1.15$) respectively for KD (Figure 7.8c) and SVM (Figure 7.11c); when DQDA is used (Figure 7.9c) it follows the Generalized Extreme Value distribution ($k = -0.45, \sigma = 1.25, \mu = 2.70$), and finally it follows the Normal distribution ($\sigma = 1.36, \mu = 2.83$) when DT is used (Figure 7.10c). Table 7.9 lists the evaluation results of the above fitted distributions. Based on the evaluation results for the cumulative density functions we clearly understand how likely it is possible to observe specific number of incorrect predictions results in the first five classifications in the easy or the difficult game levels. For example, the probability of observing *one or more* incorrect prediction



(a)

(b)

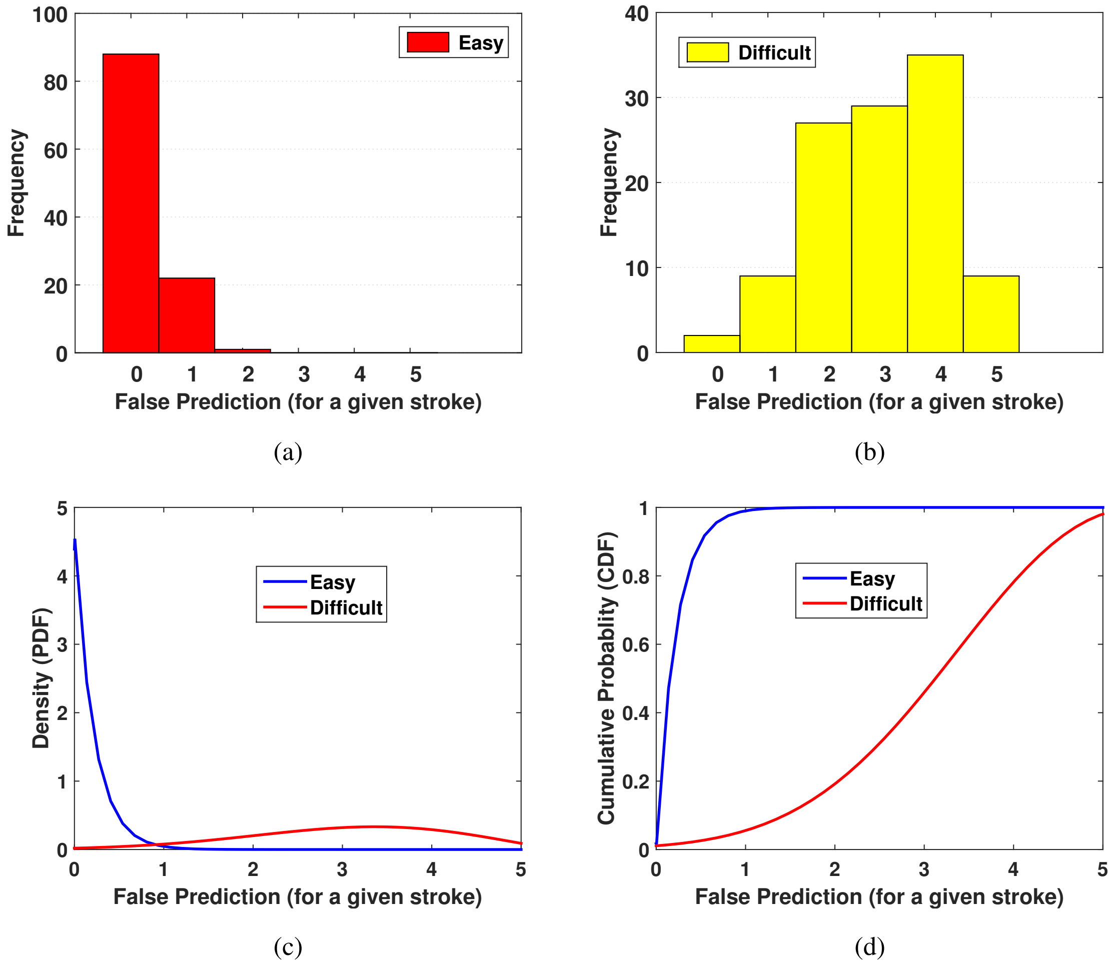
(c)

(d)

Figure 7.8: **Occurrence of incorrect predictions within a game for the first five predictions using KD in Data Set 4.**

Top-left: frequency of incorrect prediction results across all easy games. Top-right: frequency of incorrect prediction results across all difficult games. Bottom-left: the best fitted Probability Density Function (PDF) on frequency of incorrect prediction results across all easy and difficult games. Bottom-right: the best fitted Cumulative Density Function (CDF) on frequency of incorrect prediction results across all easy and difficult games.

results among the first 5 classifications is 1%, 1.2%, 34%, and 5.4% when DQDA, KD, DT, and SVM are used respectively, in the easy game. However, there is 94%, 94%, 91%, and 88% chance to find *one or more* incorrect predictions in the difficult game among the first five predictions, when DQDA, KD, DT, and SVM are used respectively, see Table 7.9.



(a)

(b)

(c)

(d)

Figure 7.9: **Occurrence of incorrect predictions within a game for the first five predictions using DQDA in Data Set 4.**

Top-left: frequency of incorrect prediction results across all easy games. Top-right: frequency of incorrect prediction results across all difficult games. Bottom-left: the best fitted Probability Density Function (PDF) on frequency of incorrect prediction results across all easy and difficult games. Bottom-right: the best fitted Cumulative Density Function (CDF) on frequency of incorrect prediction results across all easy and difficult games.

Case 2a: We observed that the classification results are more consistent in the easy games, when KD and DQDA are used, Figures 7.12a, 7.13a, whereas for all the models in the difficult game levels, we observe significant inconsistency in consecutive predictions results, Figures 7.12b, 7.13b, 7.14b, 7.15b, respectively for KD, DQDA, DT, and SVM. For example, when KD is used, across

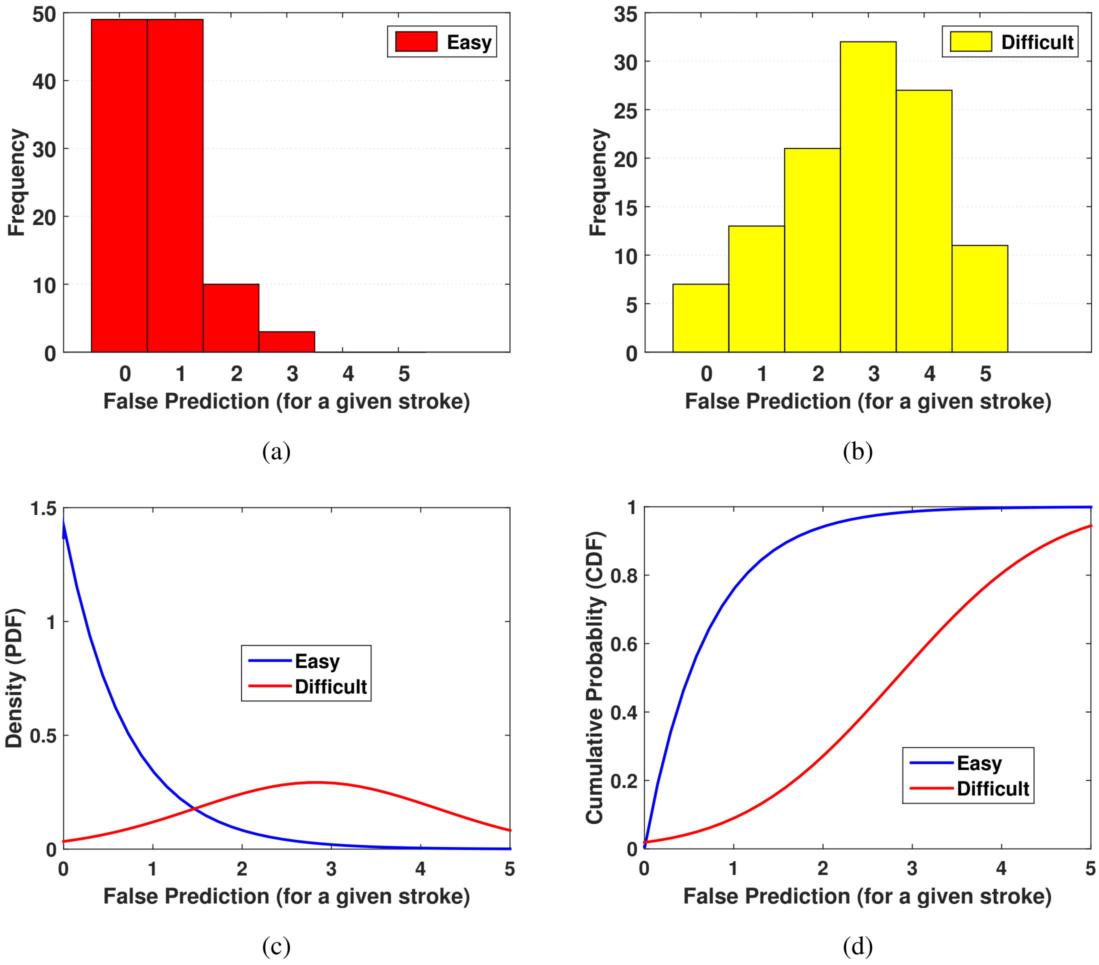
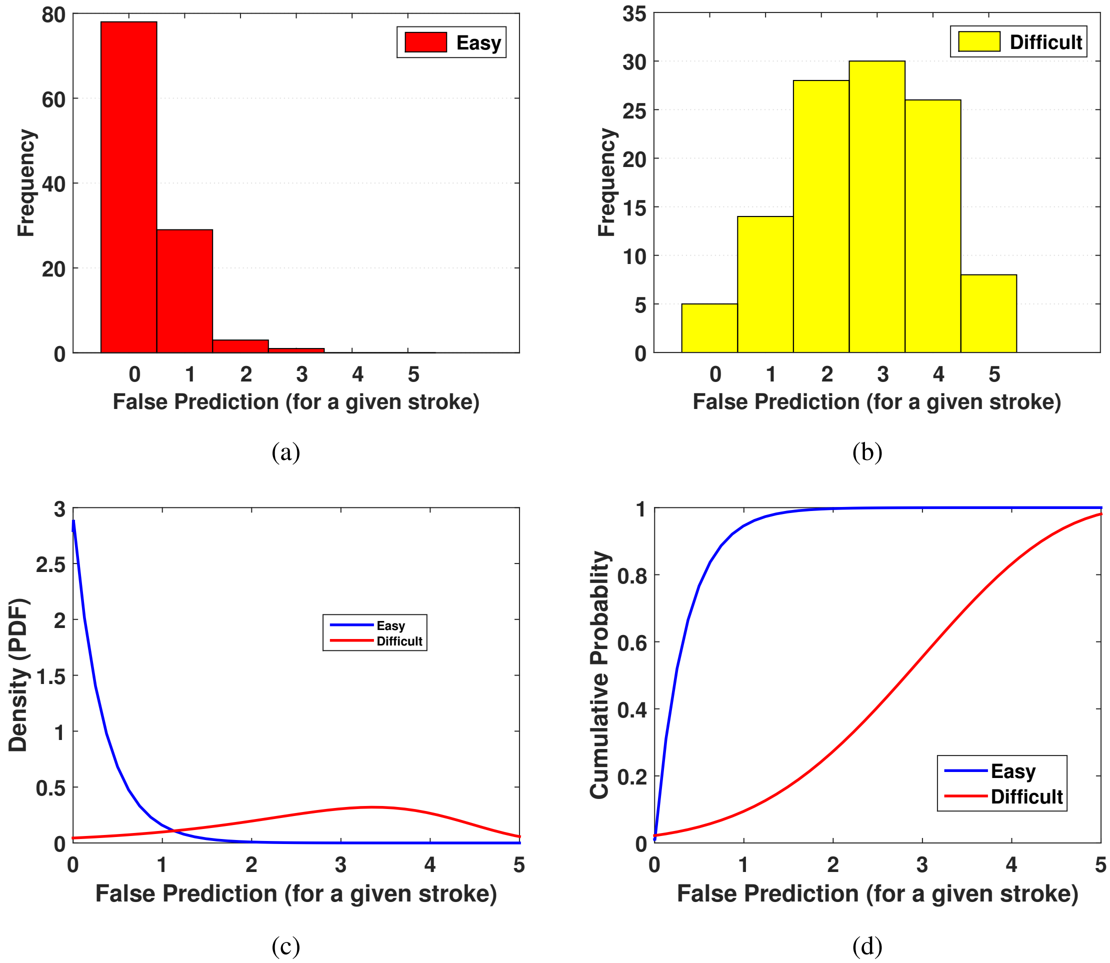


Figure 7.10: **Occurrence of incorrect predictions within a game for the first five predictions using DT in Data Set 4.**

Top-left: frequency of incorrect prediction results across all easy games. Top-right: frequency of incorrect prediction results across all difficult games. Bottom-left: the best fitted Probability Density Function (PDF) on frequency of incorrect prediction results across all easy and difficult games. Bottom-right: the best fitted Cumulative Density Function (CDF) on frequency of incorrect prediction results across all easy and difficult games.

all easy games in Data Set 4, the classification results predict easy game level consistently for 54 times, Figure 7.12a. However, in all the difficult game levels, we obtain only *one* case where all the prediction results are correct, Figure 7.12b. Notice, the 0 index in the figures represents the case when all the prediction results are consistent. Moreover, when SVM is used, across all easy



(a)

(b)

(c)

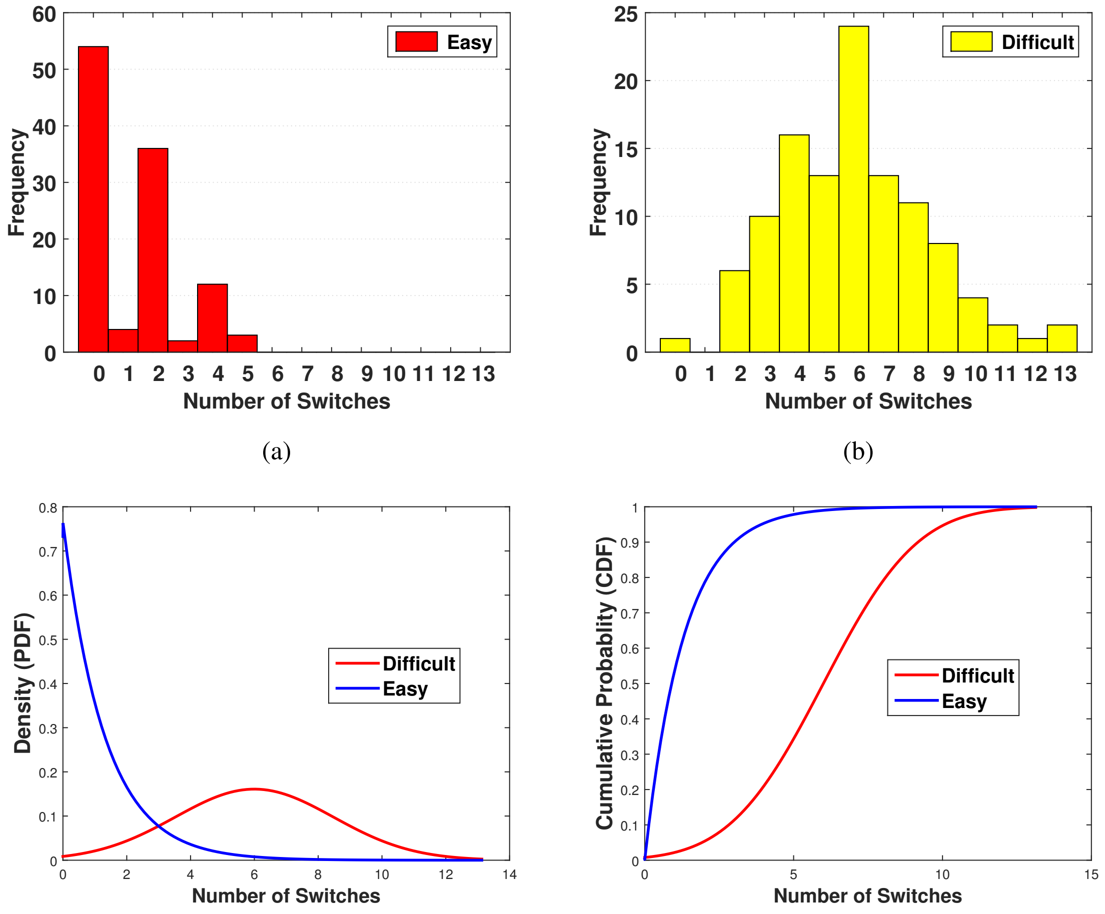
(d)

Figure 7.11: **Occurrence of incorrect predictions within a game for the first five predictions using SVM in Data Set 4.**

Top-left: frequency of incorrect prediction results across all easy games. Top-right: frequency of incorrect prediction results across all difficult games. Bottom-left: the best fitted Probability Density Function (PDF) on frequency of incorrect prediction results across all easy and difficult games. Bottom-right: the best fitted Cumulative Density Function (CDF) on frequency of incorrect prediction results across all easy and difficult games.

games, we find out that for 32 times the prediction results switch twice within a game, Figure 7.15a. This amount degrades significantly in the difficult games, where there is only 1 case when the classification results switch twice, Figure 7.15b.

The frequency of the changes in consecutive prediction results across all easy game levels



(a)

(b)

(c)

(d)

Figure 7.12: **Frequency of changes in consecutive predictions using KD in Data Set 4.** Top-left: frequency of changes in consecutive predictions across all easy games. Top-right: frequency of changes in consecutive predictions across all difficult games. Bottom-left: the best fitted Probability Density Function (PDF) on frequency of changes in consecutive predictions across all easy and difficult games. Bottom-right: the best fitted Cumulative Density Function (CDF) on frequency of changes in consecutive predictions across all easy and difficult games.

follows the exponential distribution ($\mu = 1.30$, $\mu = 1.46$, and $\mu = 2.39$), respectively when KD (Figure 7.12c), DQDA (Figure 7.13c), and SVM (Figure 7.15c) are used. Moreover, when DT is used, the frequency of the changes in consecutive prediction results across all easy game levels follows the Generalize Extreme Value distribution ($k = -0.15, \sigma = 2.31, \mu = 3.33$), Figure 7.14c. The frequency of the changes in consecutive predictions across all difficult games follow the Normal

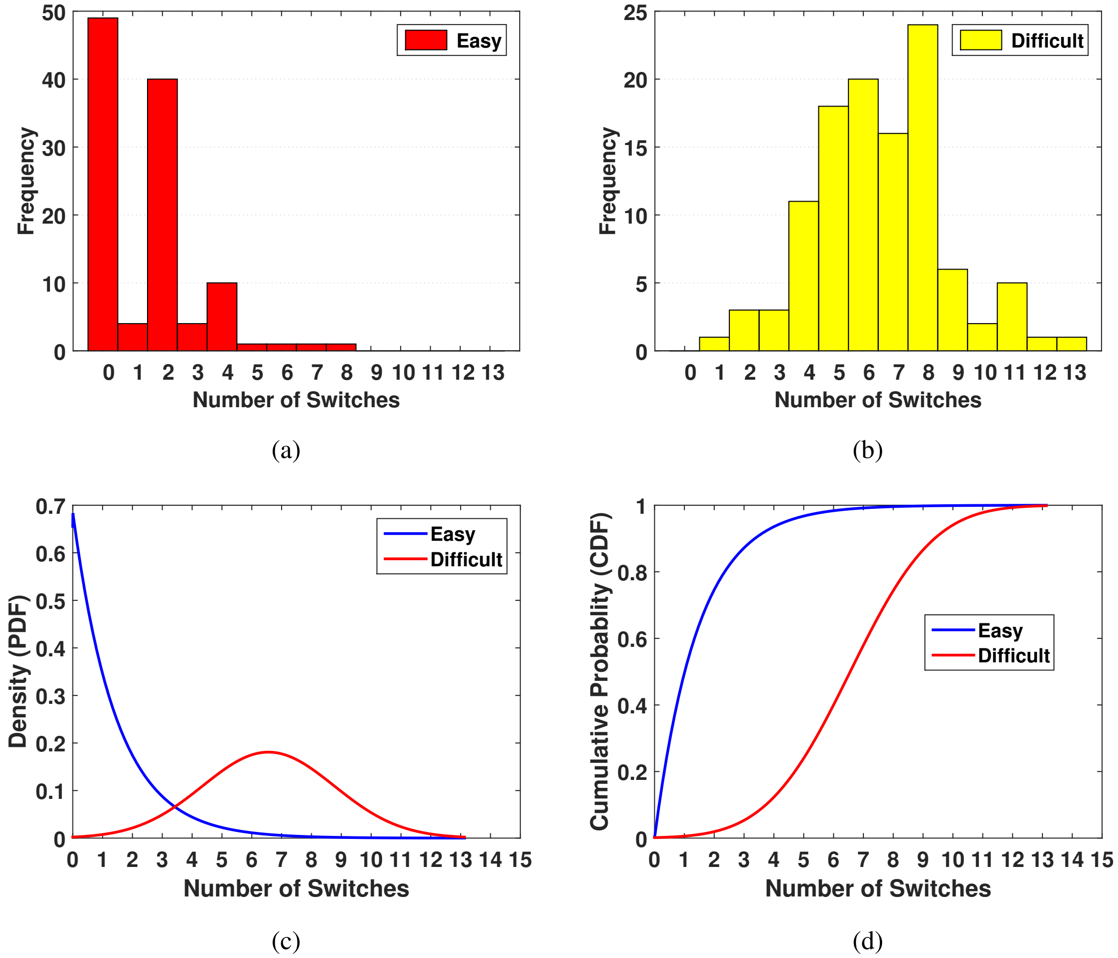
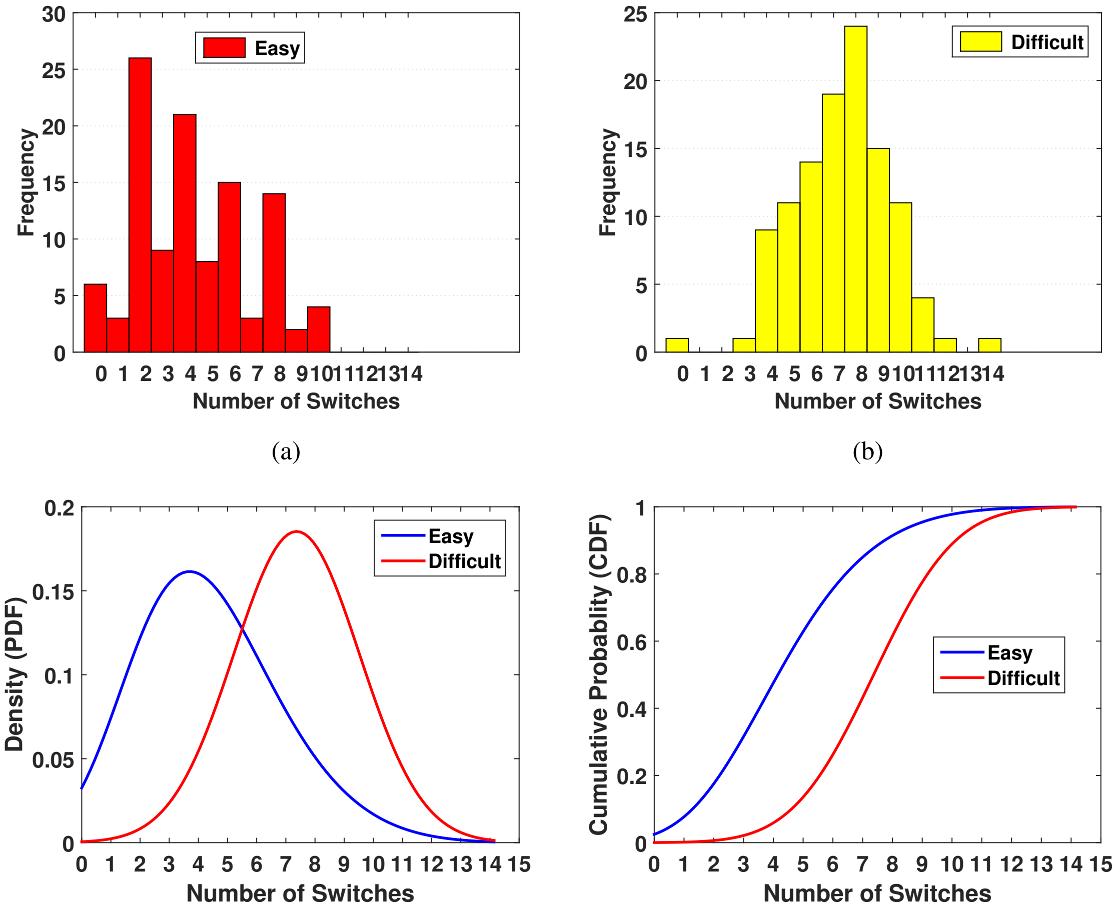


Figure 7.13: **Frequency of changes in consecutive predictions using DQDA in Data Set 4.** Top-left: frequency of changes in consecutive predictions across all easy games. Top-right: frequency of changes in consecutive predictions across all difficult games. Bottom-left: the best fitted Probability Density Function (PDF) on frequency of changes in consecutive predictions across all easy and difficult games. Bottom-right: the best fitted Cumulative Density Function (CDF) on frequency of changes in consecutive predictions across all easy and difficult games.

distribution ($\mu = 6, \sigma = 2.48; \mu = 6.56, \sigma = 2.21; \mu = 7.37, \sigma = 2.15$; and $\mu = 7.23, \sigma = 2.17$) respectively when KD (Figure 7.12c), DQDA (Figure 7.13c), DT (Figure 7.14c), and SVM (Figure 7.15c) are used. Table 7.10 lists the evaluation results of the above fitted distributions. Based on the evaluation results for the cumulative density functions we gain knowledge on how likely it is possible to find inconsistency on the classification results in the easy or the difficult game levels.



(a)

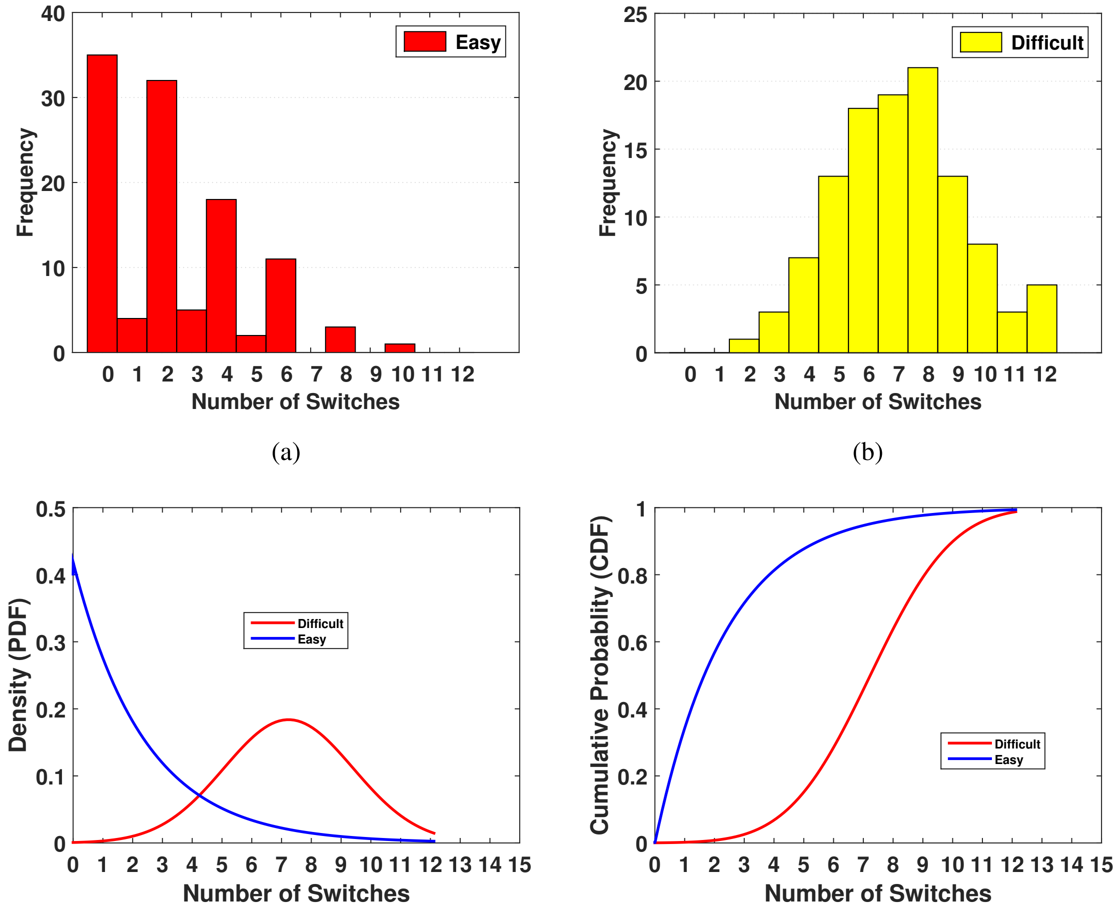
(b)

(c)

(d)

Figure 7.14: **Frequency of changes in consecutive predictions using DT in Data Set 4.** Top-left: frequency of changes in consecutive predictions across all easy games. Top-right: frequency of changes in consecutive predictions across all difficult games. Bottom-left: the best fitted Probability Density Function (PDF) on frequency of changes in consecutive predictions across all easy and difficult games. Bottom-right: the best fitted Cumulative Density Function (CDF) on frequency of changes in consecutive predictions across all easy and difficult games.

For example, when the KD model is used in the easy game, the probability of observing *two or more* inconsistent results among all consecutive predictions is 22%, however there is 95% chance to find two or more inconsistent results in the difficult game among all consecutive predictions, see Table 7.9.



(a)

(b)

(c)

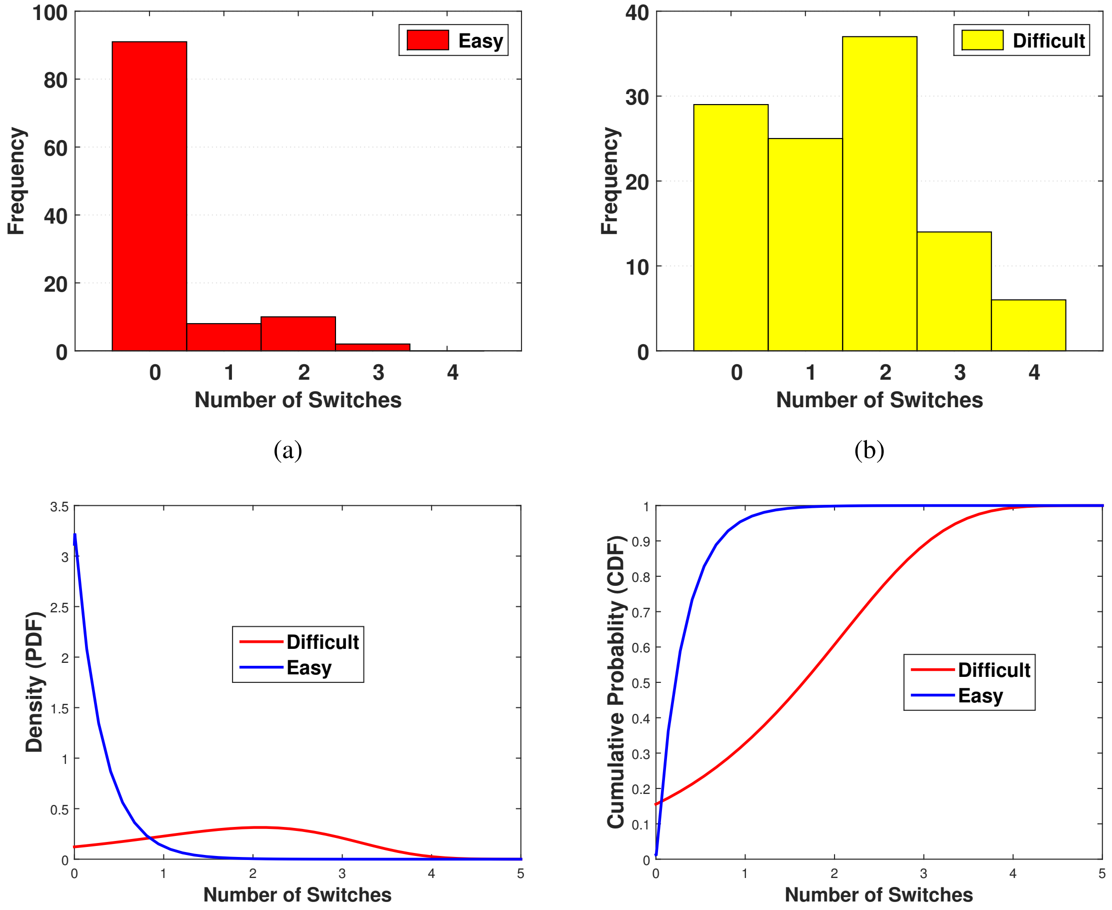
(d)

Figure 7.15: **Frequency of changes in consecutive predictions using SVM in Data Set 4.** Top-left: frequency of changes in consecutive predictions across all easy games. Top-right: frequency of changes in consecutive predictions across all difficult games. Bottom-left: the best fitted Probability Density Function (PDF) on frequency of changes in consecutive predictions across all easy and difficult games. Bottom-right: the best fitted Cumulative Density Function (CDF) on frequency of changes in consecutive predictions across all easy and difficult games.

Case 2b: Similar to the Case 2a study, when KD (Figure 7.16a), DQDA (Figure 7.17), and SVM (Figure 7.19) are used, we observed that the classification results are more consistent in the easy games, whereas in the difficult game levels, we witness significant inconsistency in consecutive predictions results (Figures 7.16b, 7.17b, 7.18b, and 7.19b, respectively for KD, DQDA, DT, and SVM). For example, across all easy games in Data Set 4, when KD is used, the first five classification results are predicting easy game level consistently for 91 times, Figure 7.16a. However, in all the difficult game levels, for only 29 cases the first five classifications results are consistent, Figure 7.16b. Similar to Case 2a, the 0 index in the figures, represents the case when all the first five prediction results are consistent. Moreover, when DQDA is used, across all easy games, we find out that for 13 times the prediction results switch twice among the first five classifications, Figure 7.17a. This amount increases significantly in all the difficult games, Figure 7.17b, where there are 45 cases when the classification results among the first five predictions switch twice.

Similar to Case 2a, but considering only the first 5 predictions, the frequency of the inconsistency in consecutive prediction results across all easy games follows the exponential distribution ($\mu = 0.31, \mu = 0.34, \mu = 1.01, \text{and}\mu = 0.49$), respectively when KD (Figure 7.16c), DQDA (Figure 7.17c), DT (Figure 7.18c), and SVM (Figure 7.19c) are used. Moreover, across all difficult games, the frequency of the inconsistency in consecutive classifications results follows Extreme Value distribution ($\mu = 2.08, \sigma = 1.17$; $\mu = 2.28, \sigma = 1.05$), respectively when KD (Figure 7.16c), and DT (Figure 7.18c) are used. Moreover, when DQDA (Figure 7.17c), and SVM (Figure 7.19c) are used, the frequency of the inconsistency in consecutive classifications results follows Normal distribution ($\mu = 1.79, \sigma = 0.98$; $\mu = 1.75, \sigma = 1.02$). Table 7.11 lists the evaluation results of the above fitted distributions. Based on the evaluation results for the cumulative density functions we understand how likely we observe inconsistency among the first five classification results within the easy or the difficult game level. For example, in the easy game, the probability of observing *one or more* inconsistent results among the first five consecutive predictions is 3%, however there is 77% chance to observe one or more inconsistent results in the difficult game among the first five consecutive predictions, see Table 7.9.

The results of the Case 1, Case 2a, and Case 2b studies can now be implemented in the game in order to reconfigure the KD classifier. These results can hence be used along with the KD classifier to gain insight on “the accuracy of prediction result within a specific game,” which in turn can successfully improve the classification performance especially in the difficult game. Moreover, results of Case 1 analysis can be used to understand the probability of the occurrence of *the specific number of incorrect predictions* within a game among the first five predictions. In other words, if we have the first five prediction results in hand, the results of the Case 1 analysis can determine how likely these classification results are made in the easy game or the difficult one. For example, given the first five prediction results as $\{c, w, w, c, c\}$ where c represents the correct prediction, and w is the incorrect classification, the results of the Case 1 analysis can determine these five predictions result are made in the difficult game by 85% chance; but by less than 1% chance in the easy game,



(a)

(b)

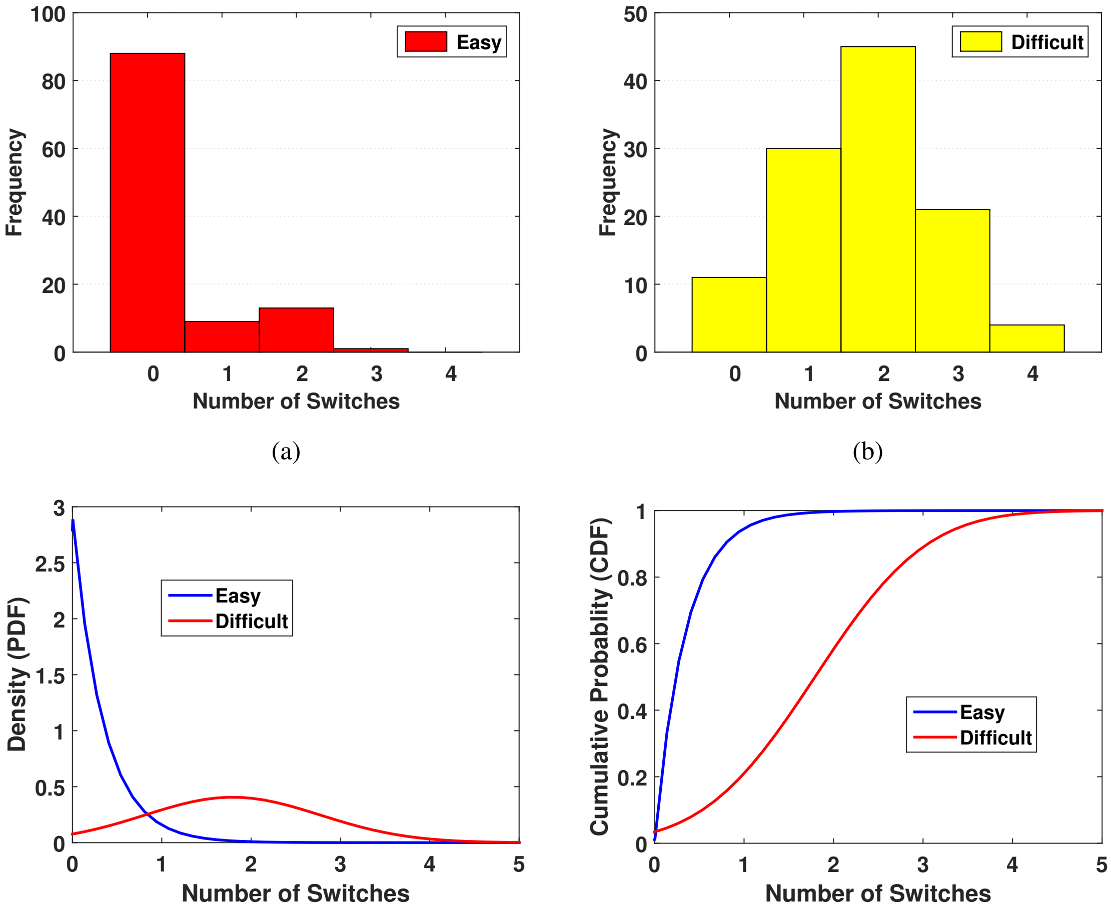
(c)

(d)

Figure 7.16: **Frequency of changes in consecutive predictions for the first 5 predictions using KD in Data Set 4.**

Top-left: frequency of changes in consecutive predictions for the first 6 finger-strokes across all easy games. Top-right: frequency of changes in consecutive predictions for the first 6 finger-strokes across all difficult games. Bottom-left: the best fitted Probability Density Function (PDF) on frequency of changes in consecutive predictions for the first 6 finger-strokes across all easy and difficult games. Bottom-right: the best fitted Cumulative Density Function (CDF) on frequency of changes in consecutive predictions for the first 6 finger-strokes across all easy and difficult games.

see Table 7.9. Recall also that Case 1 analysis results do not depend on the *order* of the incorrect predictions. In other words, for the following classifications results: $\{w, c, w, c, c\}$, $\{w, c, c, c, w\}$, $\{c, c, c, w, w\}$, and $\{c, c, w, c, w\}$, we obtain the same likelihood mentioned above. That is, all of



(a)

(b)

(c)

(d)

Figure 7.17: **Frequency of changes in consecutive predictions for the first 5 predictions using DQDA in Data Set 4.**

Top-left: frequency of changes in consecutive predictions for the first 6 finger-strokes across all easy games. Top-right: frequency of changes in consecutive predictions for the first 6 finger-strokes across all difficult games. Bottom-left: the best fitted Probability Density Function (PDF) on frequency of changes in consecutive predictions for the first 6 finger-strokes across all easy and difficult games. Bottom-right: the best fitted Cumulative Density Function (CDF) on frequency of changes in consecutive predictions for the first 6 finger-strokes across all easy and difficult games.

these five prediction results are made in the difficult game by 85% chance; but by less than 1% these are the first five classification results in the easy game, see Table 7.9.

The results of the Case 2a and Case 2b analysis can be used to understand the proba-

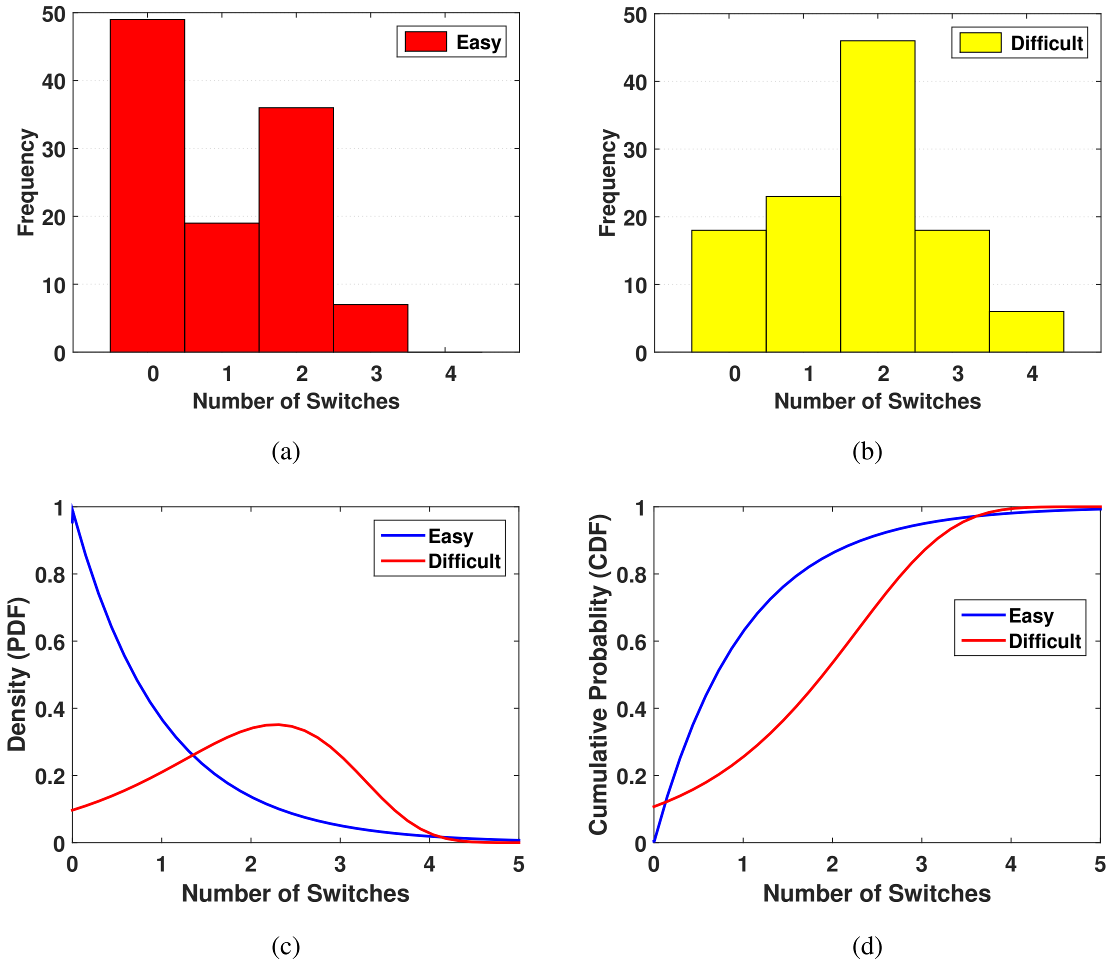
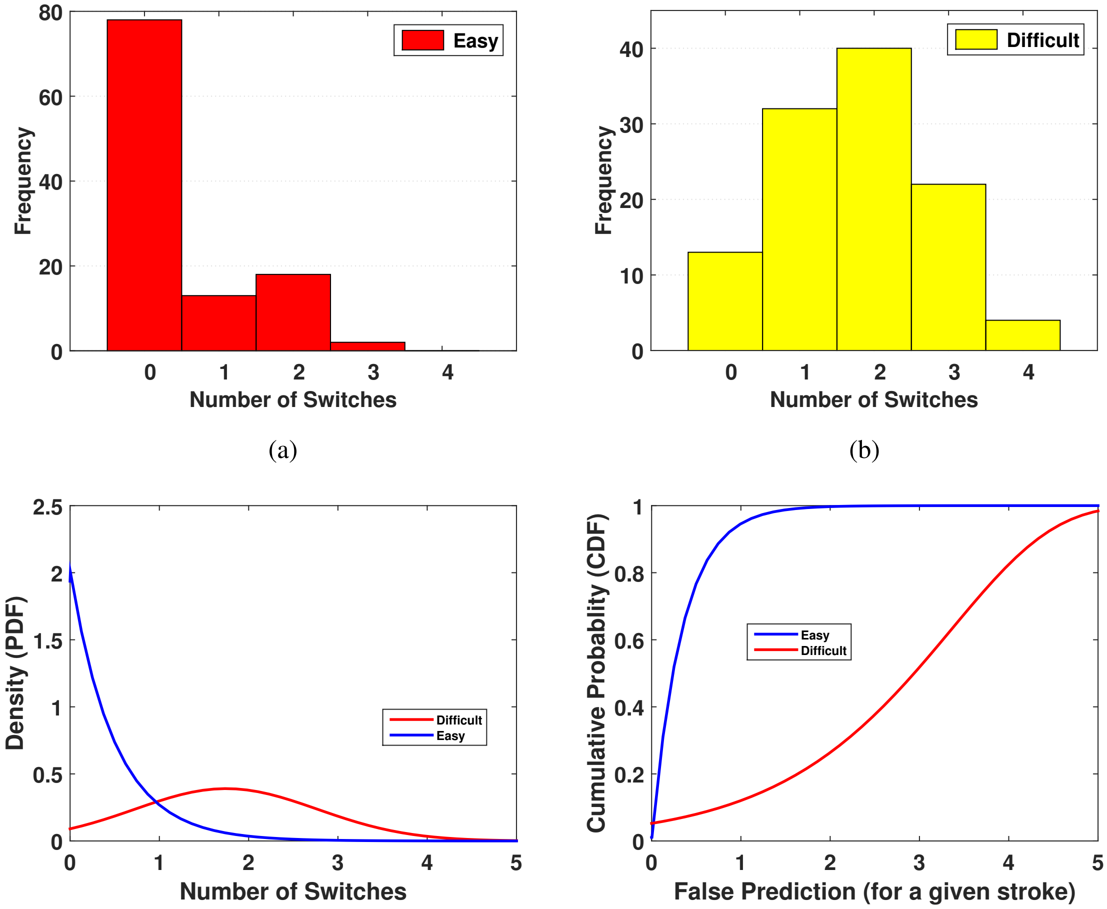


Figure 7.18: **Frequency of changes in consecutive predictions for the first 5 predictions using DT in Data Set 4.**

Top-left: frequency of changes in consecutive predictions for the first 6 finger-strokes across all easy games. Top-right: frequency of changes in consecutive predictions for the first 6 finger-strokes across all difficult games. Bottom-left: the best fitted Probability Density Function (PDF) on frequency of changes in consecutive predictions for the first 6 finger-strokes across all easy and difficult games. Bottom-right: the best fitted Cumulative De-eps-converted-to.pdfnsity Function (CDF) on frequency of changes in consecutive predictions for the first 6 finger-strokes across all easy and difficult games.

bility of the occurrence of *the specific number of inconsistent classification results in consecutive predictions* among all (Case 2a) or the first five (Case 2b) predictions made in a game level. That



(a) (b)

(c) (d)

Figure 7.19: **Frequency of changes in consecutive predictions for the first 5 predictions using SVM in Data Set 4.**

Top-left: frequency of changes in consecutive predictions for the first 6 finger-strokes across all easy games. Top-right: frequency of changes in consecutive predictions for the first 6 finger-strokes across all difficult games. Bottom-left: the best fitted Probability Density Function (PDF) on frequency of changes in consecutive predictions for the first 6 finger-strokes across all easy and difficult games. Bottom-right: the best fitted Cumulative Density Function (CDF) on frequency of changes in consecutive predictions for the first 6 finger-strokes across all easy and difficult games.

is, having all or the first five prediction results at hand, results of Case 2a and Case 2b can help determine how likely these classification results are made in the easy or the difficult level. For example, given the first five prediction results as $\{E, D, E, E, E\}$ where E represents the EASY, and

D is DIFFICULT, we are observing that the prediction results switch twice: E to D for the first and second predictions, and D to E for the second and the third predictions. Therefore results of the Case 2b analysis can determine these five predictions are made in the difficult game by 39% chance; but by less than 1% these are the first five classification results in the easy game, see Table 7.11.

The results of the Case 2a and Case 2b analysis do not depend on *where* the inconsistent results are made among all or among the first five classification results. In other words, for the following classification results: $\{E, D, D, D, E\}$, $\{D, D, E, D, D\}$, $\{D, D, D, E, D\}$, and $\{E, E, E, D, E\}$, Case 2b method evaluates the same likelihood as mentioned above. That is, all of these five prediction results are made in the difficult game by 39% chance; but by less than 1% probability, these are the first five classification results in the easy game, see Table 7.11.

Results of Case 1 can be used as a post reconfiguration method since we have the knowledge of the desired prediction, and the results of Case 2a and Case 2b analysis can be directly implemented in the game to improve the classification results in real-time. We should note that Case 1 and Case 2b results are limited to use for the first five prediction results. That is, given the limited number of classification results at hand, which are not necessarily made in the beginning of a game level, further analysis are required in order to recognize the existing patterns in these predictions to infer the specific game level that these classification results are made within.

	The probability of the occurrence of the incorrect classification							
	Easy				Difficult			
# False Prediction	DQDA	KD	DT	SVM	DQDA	KD	DT	SVM
0	97.7778	96.8085	87.5	93.9759	2.2222	3.1915	12.5	6.0241
1	70.9677	72.7273	79.0323	67.4419	29.0323	27.2727	20.9677	32.5581
2	3.5714	12	32.2581	9.6774	96.4286	88	67.7419	90.3226
3	0	4.3478	8.5714	3.2258	100	95.6522	91.4286	96.7742
4	0	0	0	0	100	100	100	100
5	0	0	0	0	100	100	100	100

Table 7.6: **The evaluation of the occurrence of the incorrect classification results among the first five predictions using the histogram plots.**

For each index in Figures 7.9a–7.9b, 7.8a–7.8b, 7.10a–7.10b, and 7.11a–7.11b, the bin sizes in the Easy and Difficult games are combined and the probability is calculated.

	The probability of the inconsistency in consecutive classifications							
	Easy				Difficult			
# Switches	DQDA	KD	DT	SVM	DQDA	KD	DT	SVM
0	100	98.1818	85.7143	100	0	1.8182	14.2857	0
1	80	100	100	100	20	0	0	0
2	93.0233	85.7143	100	96.9697	6.9767	14.2857	0	3.0303
3	57.1429	16.6667	90	62.5	42.8571	83.3333	10	37.5
4	47.619	42.8571	70	72	52.381	57.1429	30	28
5	5.2632	18.75	42.1053	13.3333	94.7368	81.25	57.8947	86.6667
6	4.7619	0	51.7241	37.931	95.2381	100	48.2759	62.069
7	5.8824	0	13.6364	0	94.1176	100	86.3636	100
8	4	0	36.8421	12.5	96	100	63.1579	87.5
9	0	0	11.7647	0	100	100	88.2353	100
10	0	0	26.6667	11.1111	100	100	73.3333	88.8889
11	0	0	0	0	100	100	100	100
12	0	0	0	0	100	100	100	100

Table 7.7: **The evaluation of the frequency of the inconsistency in consecutive classification results using the histogram plots using the histogram plots.**

For each index in Figures 7.13a–7.13b, 7.12a–7.12b, 7.14a–7.14b, and 7.15a–7.15b, the bin sizes in the Easy and Difficult games are combined and the probability is calculated.

	The probability of the inconsistency in consecutive classifications							
	Easy				Difficult			
# Switches	DQDA	KD	DT	SVM	DQDA	KD	DT	SVM
0	88.8889	75.8333	73.1343	85.7143	11.1111	24.1667	26.8657	14.2857
1	23.0769	24.2424	45.2381	28.8889	76.9231	75.7576	54.7619	71.1111
2	22.4138	21.2766	43.9024	31.0345	77.5862	78.7234	56.0976	68.9655
3	4.5455	12.5	28	8.3333	95.4545	87.5	72	91.6667
4	0	0	0	0	100	100	100	100

Table 7.8: **The evaluation of the frequency of the inconsistency in consecutive classification results among the first five predictions using the histogram plots.**

For each index in Figures 7.17a–7.17b, 7.16a–7.16b, 7.18a–7.18b, and 7.19a–7.19b, the bin sizes in the Easy and Difficult games are combined and the probability is calculated.

	$F(x)$															
	PDF								CDF							
	Easy				Difficult				Easy				Difficult			
x	DQDA	KD	DT	SVM	DQDA	KD	DT	SVM	DQDA	KD	DT	SVM	DQDA	KD	DT	SVM
0	4.625	4.44	1.4231	2.9211	0.0199	0.0221	0.0338	0.0443	0	0	0	0	0.0109	0.024	0.0188	0.0524
1	0.0453	0.0524	0.3429	0.1574	0.0794	0.0539	0.1188	0.0981	0.9902	0.9882	0.759	0.9461	0.0555	0.0597	0.0895	0.1204
2	0.0004	0.0006	0.0826	0.0085	0.2019	0.1243	0.2435	0.1959	0.9999	0.9999	0.9419	0.9971	0.1918	0.1446	0.2713	0.2636
3	0	0	0.0199	0.0005	0.3205	0.2481	0.2908	0.3057	1	1	0.986	0.9998	0.4593	0.3271	0.55	0.518
4	0	0	0.0048	0	0.2908	0.3424	0.2024	0.2654	1	1	0.9966	1	0.7819	0.6339	0.8053	0.8246
5	0	0	0.0012	0	0.0905	0.1855	0.0821	0.0568	1	1	0.9992	1	0.981	0.9218	0.9447	0.9843

Table 7.9: **The evaluation of the occurrence of the incorrect classification results among the first five predictions.**

PDF: Probability Density Function. CDF: Cumulative Density Function.

	$F(x)$															
	PDF								CDF							
	Easy				Difficult				Easy				Difficult			
x	DQDA	KD	DT	SVM	DQDA	KD	DT	SVM	DQDA	KD	DT	SVM	DQDA	KD	DT	SVM
0	0.6852	0.7655	0.0326	0.4189	0.0022	0.0086	0.0005	0.0007	0	0	0.0247	0	0.0015	0.0078	0.0003	0.0004
1	0.3453	0.356	0.0744	0.2755	0.0076	0.0211	0.0023	0.003	0.496	0.5349	0.0769	0.3422	0.0059	0.0219	0.0015	0.002
2	0.174	0.1656	0.122	0.1812	0.0214	0.0438	0.0083	0.01	0.746	0.7837	0.1755	0.5673	0.0194	0.0533	0.0063	0.0079
3	0.0877	0.077	0.1546	0.1192	0.0492	0.0774	0.0236	0.0274	0.872	0.8994	0.3157	0.7154	0.0534	0.1131	0.0212	0.0255
4	0.0442	0.0358	0.1603	0.0784	0.0923	0.1162	0.0545	0.0605	0.9355	0.9532	0.4755	0.8128	0.1231	0.2099	0.0588	0.068
5	0.0223	0.0167	0.142	0.0516	0.1409	0.1484	0.1011	0.1082	0.9675	0.9782	0.6282	0.8768	0.2399	0.3433	0.1355	0.1516
6	0.0112	0.0077	0.1111	0.0339	0.1751	0.1609	0.1514	0.1564	0.9836	0.9899	0.7553	0.919	0.4001	0.5	0.2624	0.2847
7	0.0057	0.0036	0.0786	0.0223	0.1772	0.1484	0.1826	0.1828	0.9917	0.9953	0.8499	0.9467	0.5793	0.6567	0.4319	0.457
8	0.0029	0.0017	0.051	0.0147	0.1461	0.1162	0.1775	0.1728	0.9958	0.9978	0.9142	0.9649	0.7432	0.7901	0.6152	0.6379
9	0.0014	0.0008	0.0306	0.0097	0.098	0.0774	0.1391	0.132	0.9979	0.999	0.9543	0.9769	0.8658	0.8869	0.7756	0.7921
10	0.0007	0.0004	0.0169	0.0064	0.0536	0.0438	0.0878	0.0816	0.9989	0.9995	0.9775	0.9848	0.9406	0.9467	0.8891	0.8988
11	0.0004	0.0002	0.0086	0.0042	0.0238	0.0211	0.0447	0.0408	0.9995	0.9998	0.9899	0.99	0.978	0.9781	0.9541	0.9587
12	0.0002	0.0001	0.004	0.0027	0.0086	0.0086	0.0183	0.0165	0.9997	0.9999	0.996	0.9934	0.9932	0.9922	0.9843	0.986
13	0.0001	0	0.0016	0.0018	0.0025	0.003	0.0061	0.0054	0.9999	1	0.9986	0.9957	0.9982	0.9976	0.9955	0.9961
14	0	0	0.0005	0.0012	0.0006	0.0009	0.0016	0.0014	0.9999	1	0.9996	0.9972	0.9996	0.9994	0.999	0.9991
15	0	0	0.0001	0.0008	0.0001	0.0002	0.0003	0.0003	1	1	0.9999	0.9981	0.9999	0.9999	0.9998	0.9998

Table 7.10: **The evaluation of the frequency of the inconsistency in consecutive classification results among all predictions.**

PDF: Probability Density Function. CDF: Cumulative Density Function.

	$F(x)$															
	PDF								CDF							
	Easy				Difficult				Easy				Difficult			
x	DQDA	KD	DT	SVM	DQDA	KD	DT	SVM	DQDA	KD	DT	SVM	DQDA	KD	DT	SVM
0	2.9211	3.2647	0.9911	2.0182	0.0769	0.1219	0.0967	0.0905	0	0	0	0	0.0341	0.1554	0.1071	0.0437
1	0.1574	0.1247	0.3679	0.2682	0.2932	0.2281	0.2097	0.2987	0.9461	0.9618	0.6288	0.8671	0.2099	0.3277	0.2552	0.2322
2	0.0085	0.0048	0.1365	0.0356	0.397	0.3137	0.3403	0.3786	0.9971	0.9985	0.8622	0.9823	0.5835	0.6067	0.5352	0.5975
3	0.0005	0.0002	0.0507	0.0047	0.1909	0.2091	0.2597	0.1843	0.9998	0.9999	0.9489	0.9977	0.8904	0.8885	0.8636	0.8897
4	0	0	0.0188	0.0006	0.0326	0.0254	0.0279	0.0344	1	1	0.981	0.9997	0.9876	0.9942	0.9944	0.9862
5	0	0	0.007	0.0001	0.002	0.0001	0	0.0025	1	1	0.993	1	0.9994	1	1	0.9993
6	0	0	0.0026	0	0	0	0	0.0001	1	1	0.9974	1	1	1	1	1

Table 7.11: **The evaluation of the frequency of the inconsistency in consecutive classification results among the first five predictions.** PDF: Probability Density Function. CDF: Cumulative Density Function.

7.4 Summary

Following Chapter 6, in this chapter the behavioral metrics variables, $X_1 =$ Finger-stroke Energy, $X_2 =$ Stroke Delay Time, $X_3 =$ Stroke Duration, are selected from Data Sets 2–3 in order to create a ”person-independent model” aiming to distinguish subject’s mental workload increase in real-time arising due to their inexperience. Using the aforementioned behavioral features several classifier algorithms are tested and the Kernel Density (KD) classifiers is found to have the best performance in our experiments using 10-fold cross validation.

The fitted model with KD is implemented in the game to be used in another set of experiments, Data Set 4, with different subject. Analyzing the experimental results showed that KD classifier can correctly classify the finger-strokes made in all game levels by 65.67%, and in the easy game by 95.94% accuracy. However, the performance of KD model degrades significantly to 32.62% in the difficult game.

Observing the KD performance in Data Set 4, we decided to reconstruct the Data Set 4 finger-stroke data using all the rest of the fitted models. The reconstruction results reveal that the SVM model could distinguish the finger-stroke data across all game levels by 69.75%, and by 92.57% in the easy games alone. Moreover, DT performs better in the difficult game where its accuracy is 46.97% when compared with the rest of the fitted models.

Next we investigate ways to improve the performance of the classifiers, especially for the difficult game. For this, we take advantage of variability in classifier decisions. To this end, we find the best fitted distributions on collect sequence prediction results across all easy and difficult games. Evaluating the probability density functions (PDF) and the cumulative density functions (CDF), we implement such information in order to reconfigure the corresponding classifier. Knowing “the probability of the accuracy of a given prediction result within a specific game,” we show that accuracy of detecting the difficult game level can significantly improve. This shows that one can successfully improve the fitted model performance in real-time especially in the difficult game using classifier variability as a feature.

Chapter 8

CONCLUSIONS AND FUTURE DIRECTIONS

This dissertation has outlined the development of an affective sensing approach as well as analysis of touch behavioral patterns to detect in real-time the mental states as well as mental workload changes of subjects unfamiliar with certain tasks in a touch screen game.

For this, an open source air traffic (AT) game as well as a strategic experimental protocol which could elicit different levels of mental workload by probing the inexperience aspect of volunteers human subjects are developed. By conducting a series of human subjects experiments, the experimental game is verified to indeed induce different levels of mental workload to the subjects.

We showed that subjects' inexperience in an experimental task can manifest itself as obvious variations in physiological measurements, which can then be detected using affective computing tools by non-invasive monitoring of the BVP and SC signals. Moreover, we find out that the affective sensing presents consistency on different experiments, benefiting from balanced game order, and hence, subjects' inexperience in a challenging task via the ensuing mental workload changes is detectable.

More importantly, we proposed a new practical way to fuse electrodermal activity and heart related measurements together in order to better evaluate human operator mental states without any standard training procedure. This lead to a new metric called combined metric score (CMS) which was calibrated based on metrics from our first pool of subjects, and then verified and tested on different sets of subjects. Hence, CMS offers the potential to be used in future studies as a single scalar quantity that could be used to make predictions on subjects' inexperience and/or what difficulty levels the subjects are encountering while playing various game levels.

Observing limitations in using affective computing tools specifically in real-time motivated us to study human subjects touch behavioral patterns as indicators of mental workload changes. Here, new sets of measurements mainly based on subjects' finger-strokes data are proposed, mainly related to the amount of effort invested in order to infer inexperienced subjects' mental workload changes. To test the validity of proposed effort-related metrics, the analysis results are compared with affective computing results and also with known subjects' performance and NASA-TLX questionnaire results. Strong correlation between subject's performance in familiar/unfamiliar situations, i.e., changes in mental workload are observed with subjects' physiological and touch behavioral measurements.

The findings suggest that different levels of subjects' overall performance are directly correlated with the physiological measurements. In addition, we find out that lack of experience in the presence of high mental workload produces remarkably different physiological responses, which were also associated with performance.

Having showed the reliability of the behavioral metrics in differentiating the effects of mental workload increase on subjects' behavioral patterns, finger-stroke metrics are used to create a fitted model aiming to built an *adaptive* environment. The touch based behavioral measures are used to train a model (calibration) to infer subject's inexperience in real-time. More importantly, as opposed to many studies where the model is trained and tested on the same pool of subjects, here, the sensitivity of our fitted model based on subjects' behavioral data is tested in new sets of experiments with different subjects (re-construction).

The proposed real-time detection of subjects' mental states via subjects behavioral pattern provided a successful starting point for further improving this approach. More importantly, we provided a series of analysis to re-evaluate the trained model performance in real-time mainly based on the variability of the classifiers in order to improve the performance of the fitted model.

Results obtained in this dissertation point out many future opportunities in synergistic human-machine systems, and pave the way toward real-time adaptive machines that can perform inferences to evaluate the probability of a human error in critical tasks, and can in turn provide a set of assistance modalities to the humans, with the aim to minimize such errors.

The development of this dissertation research can further be enhanced by exploring additional research avenues in the future. These directions are summarized next:

Although the power of the statistical analyses used throughout this dissertation seems to be quite favorable supporting the findings, it also calls for future studies in expanded populations as the number of subjects in this study is at the lower limit of running statistical analysis which might add some uncertainty or variability in statistical analysis results. Specifically, having larger pools of subjects can help improve the training of the touch behavioral based fitted model, and thereby improve the efficiency of real-time inference.

The fitted model based on subjects' behavioral analysis was trained only with very few finger-stroke features. We can certainly improve the performance of our model by adding more touch behavior based features. For instance, having a touch screen technology equipped with pressure sensors could provide valuable information. Verifying the methods to better identify human subjects' mental states with certain statistical and classification reliability offers opportunities toward building the machine to provide reliable assistance to the subjects.

Finally, we should note that the analysis presented in this dissertation shows strong promise in further investigating the reverse problem of "how inexperience could possibly be inferred through affective computing tools and discerned from other dimensions of mental workload." This scientific question remains as an open problem to be studied in the future, which can benefit from the results obtained in this dissertation.

Bibliography

[1] Jans Aasman, Gijsbertus Mulder, and Lambertus JM Mulder. Operator effort and the measurement of heart-rate variability. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 29(2):161–170, 1987.

[2] U Rajendra Acharya, K Paul Joseph, Natarajan Kannathal, Choo Min Lim, and Jasjit S Suri. Heart rate variability: a review. *Medical and biological engineering and computing*, 44(12):1031–1051, 2006.

[3] John Allen. Photoplethysmography and its application in clinical physiological measurement. *Physiological measurement*, 28(3):R1, 2007.

[4] Ivana Antelmi, Rogério Silva De Paula, Alexandre R Shinzato, Clóvis Araújo Peres, Alfredo José Mansur, and Cesar José Grupi. Influence of age, gender, body mass index, and functional capacity on heart rate variability in a cohort of subjects without heart disease. *The American Journal of Cardiology*, 93(3):381–385, 2004.

[5] André E Aubert, Dirk Ramaekers, Frank Beckers, Rik Breem, Carl Denef, Frans Van de Werf, and Hugo Ector. The analysis of heart rate variability in unrestrained rats. validation of method and results. *Computer methods and programs in biomedicine*, 60(3):197–213, 1999.

[6] A Barreto and JING Zhai. Physiologic instrumentation for real-time monitoring of affective state of computer users. *WSEAS Transactions on Circuits and Systems*, 3(3):496–501, 2003.

[7] Armando Barreto, Jing Zhai, and Malek Adjouadi. Non-intrusive Physiological Monitoring for Automated Stress Detection in Human-Computer Interaction. In *Human–Computer Interaction*, pages 29–38. Springer, 2007.

[8] Patricia Benner. From novice to expert. *Menlo Park*, 1984.

[9] Gary G Berntson, Karen S Quigley, Jaye F Jang, and Sarah T Boysen. An approach to artifact identification: Application to heart period data. *Psychophysiology*, 27(5):586–598, 1990.

[10] Gary G Berntson and Jeffrey R Stowell. Ecg artifacts and heart period variability: don't miss a beat! *Psychophysiology*, 35(1):127–132, 1998.

[11] Gianluca Borghini, Laura Astolfi, Giovanni Vecchiato, Donatella Mattia, and Fabio Babiloni. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, 44:58–75, 2014.

[12] Gianluca Borghini, Laura Astolfi, Giovanni Vecchiato, Donatella Mattia, and Fabio Babiloni. Measuring neurophysiological signals in aircraft pilots and car drivers for the assessment of mental workload, fatigue and drowsiness. *Neuroscience & Biobehavioral Reviews*, 44:58–75, 2014.

[13] Wolfram Boucsein. *Electrodermal activity*. Springer, 2012.

[14] David H Brainard. The Psychophysics Toolbox. *Spatial Vision*, 10(4):433–436, 1997.

[15] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and regression trees*. CRC press, 1984.

[16] Jeffrey B Brookings, Glenn F Wilson, and Carolyne R Swain. Psychophysiological responses to changes in workload during simulated air traffic control. *Biological psychology*, 42(3):361–377, 1996.

[17] Christopher JC Burges. A tutorial on support vector machines for pattern recognition. *Data mining and knowledge discovery*, 2(2):121–167, 1998.

[18] T Burkholder, Mary Chambers, Kurt Hotmire, Robert D Wurster, Stephanie Moody, and Walter C Randall. Gross and microscopic anatomy of the vagal innervation of the rat heart. *The Anatomical Record*, 232(3):444–452, 1992.

[19] A John Camm, Marek Malik, JT Bigger, G Breithardt, S Cerutti, RJ Cohen, P Coumel, EL Fallen, HL Kennedy, RE Kleiger, et al. Heart rate variability: standards of measurement, physiological interpretation and clinical use. task force of the european society of cardiology and the north american society of pacing and electrophysiology. *Circulation*, 93(5):1043–1065, 1996.

[20] Alex Cao, Keshav K Chintamani, Abhilash K Pandya, and R Darin Ellis. NASA TLX: Software for assessing subjective mental workload. *Behavior research methods*, 41(1):113–117, 2009.

[21] Hsiao-Lung Chan, Hui-Hsun Huang, and Jiunn-Lee Lin. Time-frequency analysis of heart rate variability during transient segments. *Annals of Biomedical Engineering*, 29(11):983–996, 2001.

[22] Guillaume Chanel, Cyril Rebetez, Mireille Bétrancourt, and Thierry Pun. Boredom, engagement and anxiety as indicators for adaptation to difficulty in games. In *Proceedings of the 12th international conference on Entertainment and media in the ubiquitous era*, pages 13–17. ACM, 2008.

[23] V.C. Chen and H. Ling. *Time-Frequency Transforms for Radar Imaging and Signal Analysis*. Technology and Engineering. Artech House, 2001.

[24] Burcu Cinaz, Bert Arnrich, Roberto La Marca, and Gerhard Tröster. Monitoring of mental workload levels during an everyday life office-work scenario. *Personal and Ubiquitous Computing*, 17(2):229–239, 2013.

[25] Gari D Clifford, Francisco Azuaje, and Patrick McSharry. *Advanced methods and tools for ECG data analysis*. Artech House, Inc., 2006.

[26] Ned L Cooney, Mark D Litt, Priscilla A Morse, Lance O Bauer, and Larry Gaupp. Alcohol cue reactivity, negative-mood reactivity, and relapse in treated alcoholic men. *Journal of abnormal psychology*, 106(2):243, 1997.

[27] Don Coppersmith, Se June Hong, and Jonathan RM Hosking. Partitioning nominal attributes in decision trees. *Data Mining and Knowledge Discovery*, 3(2):197–217, 1999.

[28] James W Danaher. Human error in atc system operations. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 22(5):535–545, 1980.

[29] Michael E Dawson, Anne M Schell, and Diane L Filion. 7 the electrodermal system. *Handbook of Psychophysiology*, page 159, 2007.

[30] M Doppelmayr, T Finkenzeller, and P Sauseng. Frontal midline theta in the pre-shot phase of rifle shooting: Differences between experts and novices. *Neuropsychologia*, 46(5):1463–1467, 2008.

[31] Anders Drachen, Lennart E Nacke, Georgios Yannakakis, and Anja Lee Pedersen. Correlation between heart rate, electrodermal activity and player experience in first-person shooter games. In *Proceedings of the 5th ACM SIGGRAPH Symposium on Video Games*, pages 49–54. ACM, 2010.

[32] Magy Seif El-Nasr, Anders Drachen, and Alessandro Canossa. *Game analytics: Maximizing the value of player data*. Springer, 2013.

[33] Clayton Epp, Michael Lippold, and Regan L Mandryk. Identifying emotional states using keystroke dynamics. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 715–724. ACM, 2011.

[34] K Anders Ericsson and Andreas C Lehmann. Expert and exceptional performance: Evidence of maximal adaptation to task constraints. *Annual review of psychology*, 47(1):273–305, 1996.

[35] Damian Farrow and Bruce Abernethy. Do expertise and the degree of perceptionaction coupling affect natural anticipatory performance? *Perception*, 32(9):1127–1139, 2003.

[36] B. Figner and R. O. Murphy. Using Skin Conductance in Judgment and Decision Making Research. In M. Schulte-Mecklenbeck, A. Kuehberger, and R. Ranyard, editors, *A Handbook of Process Tracing Methods For Decision Research*, pages 163–184. New York, NY: Psychology Press., 2010.

[37] Ronald A Fisher. The use of multiple measurements in taxonomic problems. *Annals of eugenics*, 7(2):179–188, 1936.

[38] Paul M Fitts. The information capacity of the human motor system in controlling the amplitude of movement. *Journal of experimental psychology*, 47(6):381, 1954.

[39] Gary M Friesen, Thomas C Jannett, Manal Afify Jadallah, Stanford L Yates, Stephen R Quint, and H Troy Nagle. A comparison of the noise sensitivity of nine qrs detection algorithms. *Biomedical Engineering, IEEE Transactions on*, 37(1):85–98, 1990.

[40] Ying Gao. *A Digital Signal Processing Approach for Affective Sensing of a Computer User through Pupil Diameter Monitoring*. PhD thesis, FIU Electronic Thesis and Dissertations, 2009.

[41] Yuan Gao, Nadia Bianchi-Berthouze, and Hongying Meng. What does touch tell us about emotions in touchscreen-based gameplay? *ACM Transactions on Computer-Human Interaction (TOCHI)*, 19(4):31, 2012.

[42] WB Gartner and MR Murphy. Concepts of workload. *BO Hartman and RE McKenzie, Survey of methods to assess workload: AGARD-AG-246*, 1979.

[43] Alan Gevins and Michael E Smith. Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science*, 4(1-2):113–131, 2003.

[44] Dimitris Giakoumis, Athanasios Vogiannou, Illka Kosunen, Kostantinos Moustakas, Dimitrios Tzovaras, and George Hassapis. Identifying Psychophysiological Correlates of Boredom and Negative Mood Induced During HCI. In *1st International Workshop on Bio-inspired HumanMachine Interfaces and Healthcare Applications*, pages 3–12, 2010.

[45] A.C. Guyton and J.E. Hall. *Textbook of medical physiology*. Elsevier Saunders, 2006.

[46] R Haahr. Reflectance pulse oximetry sensor for the electronic patch. *Master of Science dissertation, Dept. of Micro and Nanotechnology, Technical University of Denmark*, 2006.

[47] Peter Hancock, Mark H Chignell, et al. Mental workload dynamics in adaptive interface design. *Systems, Man and Cybernetics, IEEE Transactions on*, 18(4):647–658, 1988.

[48] Peter A Hancock and James L Szalma. Operator stress and display design. *Ergonomics in Design: The Quarterly of Human Factors Applications*, 11(2):13–18, 2003.

[49] Sandra G Hart and Lowell E Staveland. Development of nasa-tlx (task load index): Results of empirical and theoretical research. *Advances in psychology*, 52:139–183, 1988.

[50] Jennifer A Healey and Rosalind W Picard. Detecting Stress During Real-World Driving Tasks Using Physiological Sensors. *IEEE Transactions on Intelligent Transportation Systems*, 6(2):156–166, 2005.

[51] Sylvie Hébert, Renée Béland, Odrée Dionne-Fournelle, Martine Crête, and Sonia J Lupien. Physiological stress response to video-game playing: The contribution of built-in music. *Life sciences*, 76(20):2371–2380, 2005.

[52] Carrie Heeter, Yu-Hao Lee, Ben Medler, and Brian Magerko. Chapter 32 - conceptually meaningful metrics: Inferring optimal challenge and mindset from gameplay. In Magy Seif El-Nasr, Anders Drachen, and Alessandro Canossa, editors, *Game Analytics: Maximizing the Value of Player Data*. Springer, 2013.

[53] Andreas Henelius, Kati Hirvonen, Anu Holm, Jussi Korpela, and Kiti Müller. Mental workload classification using heart rate metrics. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pages 1836–1839. IEEE, 2009.

[54] Matthew J Hertenstein, Rachel Holmes, Margaret McCullough, and Dacher Keltner. The communication of emotion via touch. *Emotion*, 9(4):566, 2009.

[55] William E Hick. On the rate of gain of information. *Quarterly Journal of Experimental Psychology*, 4(1):11–26, 1952.

[56] Justin G Hollands and Christopher D Wickens. Engineering psychology and human performance. *Journal of surgical oncology*, 1999.

[57] Adam Hoover, Anirud Singh, Stephanie Fishel-Brown, and Eric Muth. Real-time detection of workload changes using heart rate variability. *Biomedical Signal Processing and Control*, 7(4):333–341, 2012.

[58] VD Hopkin. Implications of automation on air traffic control. *Aviation psychology*, 1989.

[59] Norden E. Huang, Zheng Shen, Steven R. Long, Manli C. Wu, Hsing H. Shih, Quanan Zheng, Nai-Chyuan Yen, Chi Chao Tung, and Henry H. Liu. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis. *Proceedings of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, 454(1971):903–995, 1998.

[60] Eva Hudlicka. To feel or not to feel: The role of affect in human–computer interaction. *International Journal of Human-Computer Studies*, 59(1):1–32, 2003.

[61] Eva Hudlicka. Affective game engines: motivation and requirements. In *Proceedings of the 4th International Conference on Foundations of Digital Games*, pages 299–306. ACM, 2009.

[62] Ray Hyman. Stimulus information as a determinant of reaction time. *Journal of experimental psychology*, 45(3):188, 1953.

[63] Georg Jahn, Astrid Oehme, Josef F Krems, and Christhard Gelau. Peripheral detection as a workload measure in driving: Effects of traffic complexity and route guidance system use in a driving study. *Transportation Research Part F: Traffic Psychology and Behaviour*, 8(3):255–275, 2005.

[64] George H John and Pat Langley. Estimating continuous distributions in bayesian classifiers. In *Proceedings of the Eleventh conference on Uncertainty in artificial intelligence*, pages 338–345. Morgan Kaufmann Publishers Inc., 1995.

[65] Peter GAM Jorna. Spectral analysis of heart rate and psychological state: A review of its validity as a workload index. *Biological psychology*, 34(2):237–257, 1992.

[66] David B Kaber, Carlene M Perry, Noa Segall, and Mohamed A Sheik-Nainar. Workload State Classification With Automation During Simulated Air Traffic Control. *The International Journal of Aviation Psychology*, 17(4):371–390, 2007.

[67] JWH Kalsbeek and JH Ettema. Scored regularity of the heart rate pattern and the measurement of perceptual or mental load. *Ergonomics*, 6(3):306–307, 1963.

[68] J.F. Kenney and E.S. Keeping. *Mathematics of statistics*. Mathematics of Statistics. Van Nostrand, 1947.

[69] Kyung Hwan Kim, SW Bang, and SR Kim. Emotion recognition system using short-term monitoring of physiological signals. *Medical and biological engineering and computing*, 42(3):419–427, 2004.

[70] Stella Kim and Laura Smith-Spark. *CNN: Asiana says pilot error partly to blame for San Francisco plane crash*, 2014 (accessed March 5, 2015).

[71] J Matias Kivikangas, Guillaume Chanel, Ben Cowley, Inger Ekman, Mikko Salminen, Simo Järvelä, and Niklas Ravaja. A review of the use of psychophysiological methods in game research. *Journal of Gaming & Virtual Worlds*, 3(3):181–199, 2011.

[72] J Matias Kivikangas, Inger Ekman, Guillaume Chanel, Simo Järvelä, M Salminen, B Cowley, Pentti Henttonen, and Niklas Ravaja. Review on psychophysiological methods in game research. *Proc. of 1st Nordic DiGRA*, 2010.

[73] B-U Kohler, Carsten Hennig, and Reinhold Orglmeister. The principles of software qrs detection. *Engineering in Medicine and Biology Magazine, IEEE*, 21(1):42–57, 2002.

[74] Azadeh Kushki, Jillian Fairley, Satyam Merja, Gillian King, and Tom Chau. Comparison of blood volume pulse and skin conductance responses to mental and affective stimuli at different anatomical site. *Physiological Measurement*, 32(10):1529, 2011.

[75] Peter J Lang. The emotion probe: studies of motivation and attention. *American psychologist*, 50(5):372, 1995.

[76] MN Levy and MR Warner. Parasympathetic effects on cardiac function. *Neurocardiology*, pages 53–76, 1994.

[77] M Malik, T Cripps, T Farrell, and AJ Camm. Prognostic value of heart rate variability after myocardial infarction. a comparison of different data-processing methods. *Medical and Biological Engineering and Computing*, 27(6):603–611, 1989.

[78] Marek Malik. Heart rate variability. *Annals of Noninvasive Electrocardiology*, 1(2):151–181, 1996.

[79] Marek Malik, J Thomas Bigger, A John Camm, Robert E Kleiger, Alberto Malliani, Arthur J Moss, and Peter J Schwartz. Heart rate variability standards of measurement, physiological interpretation, and clinical use. *European Heart Journal*, 17(3):354–381, 1996.

[80] S. Mallat. *A Wavelet Tour of Signal Processing, The Sparse Way*. Academic Press, 2009.

[81] Regan L Mandryk. Physiological measures for game evaluation. *Game usability: Advice from the experts for advancing the player experience*, pages 207–235, 2008.

[82] Regan L Mandryk and M Stella Atkins. A fuzzy physiological approach for continuously modeling emotion during interaction with play technologies. *International Journal of Human-Computer Studies*, 65(4):329–347, 2007.

[83] Regan L Mandryk and Kori M Inkpen. Physiological indicators for the evaluation of co-located collaborative play. In *Proceedings of the 2004 ACM conference on Computer supported cooperative work*, pages 102–111. ACM, 2004.

[84] Regan L Mandryk, Kori M Inkpen, and Thomas W Calvert. Using psychophysiological techniques to measure user experience with entertainment technologies. *Behaviour & Information Technology*, 25(2):141–158, 2006.

[85] Regan Lee Mandryk. *Modeling user emotion in interactive play environments: A fuzzy physiological approach*. PhD thesis, School of Computing Science-Simon Fraser University, 2005.

[86] Christopher D Manning, Prabhakar Raghavan, Hinrich Schütze, et al. *Introduction to information retrieval*, volume 1. Cambridge university press Cambridge, 2008.

[87] David Marshall, Damien Coyle, Shane Wilson, and Michael Callaghan. Games, gameplay, and bci: The state of the art. *Computational Intelligence and AI in Games, IEEE Transactions on*, 5(2):82–99, 2013.

[88] Martial M. Massin, Benedicte Derkenne, and Goetz von Bernuth. Correlations Between Indices of Heart Rate Variability in Healthy Children and Children with Congenital Heart Disease. *Cardiology*, 91(2):109–113, 1999.

[89] Bruce Mehler, Bryan Reimer, and Joseph F Coughlin. Sensitivity of physiological measures for detecting systematic variations in cognitive demand from a working memory task an on-road study across three age groups. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 54(3):396–412, 2012.

[90] Bruce Mehler, Bryan Reimer, Joseph F Coughlin, and Jeffery A Dusek. Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers. *Transportation Research Record: Journal of the Transportation Research Board*, 2138(1):6–12, 2009.

[91] Bruce Mehler, Bryan Reimer, and Ying Wang. A comparison of heart rate and heart rate variability indices in distinguishing single-task driving and driving under secondary cognitive workload. In *Proceedings of the Sixth International Driving Symposium on Human Factors in Driver Assessment, Training, and Vehicle Design*, pages 590–597, 2011.

[92] N Meshkati and PA Hancock. *Human mental workload*. Elsevier, 2011.

[93] Ulla Metzger and Raja Parasuraman. Automation in future air traffic management: Effects of decision aid reliability on controller performance and mental workload. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 47(1):35–49, 2005.

[94] Ronald R Mourant and Thomas H Rockwell. Strategies of visual search by novice and experienced drivers. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 14(4):325–335, 1972.

[95] G Mulder. Sinusarrhythmia and mental work load. In *Mental Workload*, pages 327–343. Springer, 1979.

[96] Kevin P Murphy. Naive bayes classifiers. *University of British Columbia*, 2006.

[97] Nicholas P Murray and Carmen Russoniello. Acute physical activity on cognitive function: a heart rate variability examination. *Applied Psychophysiology and Biofeedback*, 37(4):219–227, 2012.

[98] Michael Nolan. *Fundamentals of air traffic control*. Cengage Learning, 2010.

[99] Reetta Orsila, Matti Virtanen, Tiina Luukkaala, Mika Tarvainen, Pasi Karjalainen, Jari Viik, Minna Savinainen, and Clas-Hakan Nygard. Perceived Mental Stress and Reactions in Heart Rate Variability–A Pilot Study Among Employees of an Electronics Company. *International Journal of Occupational Safety and Ergonomics (JOSE)*, 14(3):275–283, 2008.

[100] Payam Parsinejad, Yolanda Rodriguez-Vaqueiro, Jose Angel Martinez-Lorenzo, and Rifat Sipahi. Combined time-frequency calculation of pnn50 metric from noisy heart rate measurements. In *ASME 2014 Dynamic Systems and Control Conference*, pages V001T06A004–V001T06A004. American Society of Mechanical Engineers, 2014.

[101] Payam Parsinejad and Rifat Sipahi. A touchscreen game to induce mental workload on human subjects. In *40th Annual Northeast Bioengineering Conference (NEBEC)*, pages 1–2, Boston, MA, 2014. IEEE.

[102] Payam Parsinejad and Rifat Sipahi. Assessment of human vulnerability in a touch-screen game; metrics and analysis. In *ASME 2015 Dynamic Systems and Control Conference*, pages V001T09A004–V001T09A004. American Society of Mechanical Engineers, 2015.

[103] Denis G Pelli. The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial Vision*, 10(4):437–442, 1997.

[104] Rosalind W Picard. *Affective computing*. MIT press, 1997.

[105] Rosalind W Picard and Jennifer Healey. Affective wearables. *Personal Technologies*, 1(4):231–240, 1997.

[106] Rosalind W Picard and Jocelyn Scheirer. The galvactivator: A glove that senses and communicates skin conductivity. In *Proceedings 9th Int. Conf. on HCI*, 2001.

[107] Rosalind W. Picard, Elias Vyzas, and Jennifer Healey. Toward machine emotional intelligence: Analysis of affective physiological state. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23(10):1175–1191, 2001.

[108] Walter Piechulla, Christoph Mayser, Helmar Gehrke, and Winfried König. Reducing drivers? mental workload by means of an adaptive man–machine interface. *Transportation Research Part F: Traffic Psychology and Behaviour*, 6(4):233–248, 2003.

[109] Gillian Porter, Tom Troscianko, and Iain D Gilchrist. Effort during visual search and counting: Insights from pupillometry. *The Quarterly Journal of Experimental Psychology*, 60(2):211–229, 2007.

[110] Daniel S Quintana and James AJ Heathers. Considerations in the assessment of heart rate variability in biobehavioral research. *Frontiers in Psychology*, 5, 2014.

[111] John T Ramshur. *Design, Evaluation, and Application of Heart Rate Variability Analysis Software (HRVAS)*. PhD thesis, The University of Memphis, jul 2010.

[112] Jeromie Rand, Adam Hoover, Stephanie Fishel, Jason Moss, Jennifer Pappas, and Eric Muth. Real-time correction of heart interbeat intervals. *Biomedical Engineering, IEEE Transactions on*, 54(5):946–950, 2007.

[113] Bryan Reimer and Bruce Mehler. The impact of cognitive workload on physiological arousal in young adult drivers: a field study and simulation validation. *Ergonomics*, 54(10):932–942, 2011.

[114] Peter Richter, Thomas Wagner, Ralf Heger, and Gunther Weise. Psychophysiological analysis of mental load during driving on rural roads-a quasi-experimental field study. *Ergonomics*, 41(5):593–609, 1998.

[115] Dennis W Rowe, John Sibert, and Don Irwin. Heart rate variability: Indicator of user state as an aid to human-computer interaction. In *Proceedings of the SIGCHI conference on Human*

factors in computing systems, pages 480–487. ACM Press/Addison-Wesley Publishing Co., 1998.

[116] Kilseop Ryu and Rohae Myung. Evaluation of mental workload with a combined measure based on physiological indices during a dual task of tracking and mental arithmetic. *International Journal of Industrial Ergonomics*, 35(11):991–1009, 2005.

[117] Gideon Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 6(2):461–464, 1978.

[118] Cornelia Setz, Bert Arnrich, Johannes Schumm, Roberto La Marca, G Troster, and Ulrike Ehlert. Discriminating Stress From Cognitive Load Using a Wearable EDA Device. *IEEE Transactions on Information Technology in Biomedicine*, 14(2):410–417, 2010.

[119] Dilbag Singh, K Vinod, and SC Saxena. Sampling frequency of the rr interval time series for spectral analysis of heart rate variability. *Journal of medical engineering & technology*, 28(6):263–272, 2004.

[120] Robert Morris Stern, William J Ray, and Karen S Quigley. *Psychophysiological recording*. Oxford University Press, 2001.

[121] J Ridley Stroop. Interference in serial verbal reactions. *Journal of Experimental Psychology*, 18:643–661, 1935.

[122] Feng-Tso Sun, Cynthia Kuo, Heng-Tze Cheng, Senaka Buthpitiya, Patricia Collins, and Martin Griss. Activity-aware mental stress detection using physiological sensors. In *Mobile computing, applications, and services*, pages 211–230. Springer, 2012.

[123] Joachim Taelman, Steven Vandeput, Arthur Spaepen, and Sabine Van Huffel. Influence of Mental Stress on Heart Rate and Heart Rate Variability. In *4th European Conference of the International Federation for Medical and Biological Engineering*, pages 1366–1369, 2009.

[124] Richard J Tafalla. Gender Differences in Cardiovascular Reactivity and Game Performance Related to Sensory Modality in Violent Video Game Play. *Journal of Applied Social Psychology*, 37(9):2008–2023, 2007.

[125] RA Thuraisingham. Preprocessing RR interval time series for heart rate variability analysis and estimates of standard deviation of RR intervals. *Computer Methods and Programs in Biomedicine*, 83(1):78–82, 2006.

[126] Kim J Vicente, D Craig Thornton, and Neville Moray. Spectral analysis of sinus arrhythmia: A measure of mental effort. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 29(2):171–182, 1987.

[127] Mark Wiggins*, Catherine Stevens, Amanda Howard, Irene Henley, and David O'Hare. Expert, intermediate and novice performance during simulated pre-flight decision-making. *Australian Journal of Psychology*, 54(3):162–167, 2002.

[128] Gillian M Wilson and M Angela Sasse. Investigating the impact of audio degradations on users: Subjective vs. objective assessment methods. In *Proc. OZCHI 2000*, pages 135–142, 2000.

[129] Glenn F Wilson. Applied use of cardiac and respiration measures: Practical considerations and precautions. *Biological Psychology*, 34(2):163–178, 1992.

[130] Glenn F Wilson. An analysis of mental workload in pilots during flight using multiple psychophysiological measures. *The International Journal of Aviation Psychology*, 12(1):3–18, 2002.

[131] Glenn F Wilson, Jared D Lambert, and Chris A Russell. Performance enhancement with real-time physiologically controlled adaptive aiding. In *Proceedings of the human factors and ergonomics society annual meeting*, volume 44, pages 61–64. SAGE Publications, 2000.

[132] Glenn F Wilson and Christopher A Russell. Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 45(4):635–644, 2003.

[133] Glenn F Wilson and Christopher A Russell. Psychophysiologically determined adaptive aiding in a simulated ucav task. *Human performance, situation awareness, and automation: Current research and trends*, pages 200–204, 2004.

[134] Glenn F Wilson and Christopher A Russell. Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. *Human Fctors: The Journal of the Human Factors and Ergonomics Society*, 49(6):1005–1018, 2007.

[135] Yong-Jie Yao, Yao-Ming Chang, Xiao-Ping Xie, Xin-Sheng Cao, Xi-Qing Sun, and Yan-Hong Wu. Heart rate and respiration responses to real traffic pattern flight. *Applied psychophysiology and biofeedback*, 33(4):203–209, 2008.

[136] RD Yatess and DJ Goodman. Probability and stochastic processes. a friendly introduction for electrical and computer engineering. 1999.

[137] Jing Zhai, Armando B Barreto, Craig Chin, and Chao Li. Realization of stress detection using psychophysiological signals for improvement of human-computer interactions. In *Southeast-Con, 2005. Proceedings. IEEE*, pages 415–420. IEEE, 2005.