The Use of the CAfFEINE Framework in a Step-by-Step Assembly Guide

Devin Kyle Ketchum

Thesis submitted to the faculty of the Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Master of Science
In
Computer Engineering

Thomas L. Martin
Benjamin R. Knapp
Denis Gracanin

December 13, 2019
Blacksburg, VA

Keywords: Affective Computing, Intelligent Environment, Context-Aware

The Use of the CAfFEINE Framework in a Step-by-Step Assembly Guide

Devin Kyle Ketchum

ABSTRACT

Today's technology is becoming more interactive with voice assistants like Siri. However, interactive systems such as Siri make mistakes. The purpose of this thesis is to explore using affect as an implicit feedback channel so that such mistakes would be easily corrected in real time. The CAfFEINE Framework, which was created by Dr. Saha, is a context-aware affective feedback loop in an intelligent environment. For the research described in this thesis, the focus will be on analyzing a user's physiological response to the service provided by an intelligent environment. To test this feedback loop, an experiment was constructed using an on-screen, step-by-step assembly guide for a Tangram puzzle. To categorize the user's response to the experiment, baseline readings were gathered for a user's stressed and non-stressed state. The Paced Stroop Test and two other baseline tests were conducted to gather these two states. The data gathered in the baseline tests was then used to train a support vector machine to predict the user's response to the Tangram experiment.

During the data analysis phase of the research, the results for the predictions on the Tangram experiment were not as expected. Multiple trials of training data for the support vector machine were explored, but the data gathered throughout this research was not enough to draw proper conclusions. More focus was then given to analyzing the pre-processed data of the baseline tests in an attempt to find a factor or group of factors to determine if the user's physiological responses would be useful to train the Support Vector Machine. There were trends found when comparing the area under the curves of

the Paced Stroop Test phasic driver plots. It was found that these comparison factors might be a useful approach for differentiating users based upon their physiological responses during the Paced Stroop Test.

The Use of the CAfFEINE Framework in a Step-by-Step Assembly Guide

Devin Kyle Ketchum

General Audience Abstract

The purpose of this thesis was to use the CAfFEINE Framework, proposed by Dr. Saha, in a real-world environment. Dr. Saha's Framework utilizes a user's physical responses, i.e. heart rate, in a smart environment to give information to the smart devices. For example, if Siri were to give a user directions to someone's home and told that user to turn right when the user knew they needed to turn left. That user would have a physical reaction as in their heart rate would increase. If the user were wearing a smart watch, Siri would be able to see the heart rate increase and realize, from past experiences with that user, that the information she gave to the user was incorrect. Then she would be able to correct herself.

My research focused on measuring user reaction to a smart service provided in a real-world situation using a Tangram puzzle as a mock version of an industrial assembly situation. The users were asked to follow on-screen instructions to assemble the Tangram puzzle. Their reactions were recorded through a smart watch and analyzed post-experiment. Based on the results of a Paced Stroop Test they took before the experiment, a computer algorithm would predict their stress levels for each service provided by the step-by-step instruction guide. However, the results did not turn out as expected. Therefore, the rest of the research focused more on why the results did not support Dr. Saha's previous Framework results.

... dedicated to my community,
my wonderful wife,
and my beautiful son,
Yona ...

Acknowledgements

I would first like to thank my co-advisors Dr. Tom Martin and Dr. Ben Knapp for going on this journey with me. The journey was long and at times more than complicated, but they were there to guide me through it all. Tom and Ben's guidance reminded me of my father's guidance, firm and concise but also lighthearted. It was always an enlightening and enjoyable environment when we sat down to discuss my thesis.

I would also like to thank Dr. Deba Saha. Deba was the mentor I truly needed for this research. The knowledge and insight he shared with me on a daily basis helped shape my understanding with both the small intricacies of the project and the big picture of the framework. I will always consider him a great mentor an even better friend.

I would also like to thank Dean DePauw as my outside source and mentor. Without Dean DePauw's constant help and guidance, I could not have completed this degree. The relationship I built with her will last well beyond my time here at Virginia Tech.

I also have great appreciation for my community: Native at VT, my Virginia Tech community, and the community I built at Blue Ridge Church. My community has been there to support me in more ways than they could ever know. I am forever grateful for the time we have had together.

Lastly, but certainly not least, my wonderful wife Qualla. Accomplishing this momentous task could not have been possible without her constant love and support. I want to tell her thank you for putting up with the long nights, mood swings, and constant doubt. You have been my constant rock and source of confidence.

Contents

List of Figures

List of Tables

Chapter 1 Introduction

Wearables, mobile computing, and context-aware systems have become a major focus of research in the technology community. There is a consistent drive to make environments more contextually aware. Context-aware computing has led to the exploration of physiological measurements to infer the correctness of a service being provided to a user. Measuring a user's physiological response can give the system implicit feedback without interrupting the user's cognition [1].

A term used often in this research is Context-Aware Intelligent Environment (CAIE), which is different than an intelligent environment. An Intelligent Environment (IE), as defined by Augusto et al. [2], is a space or area that a person is interacting with that has sensors, devices, and computers to make the user's life easier. The term used in this thesis experiment, Context-Aware Intelligent Environment, builds on the previous definition. CAIE "is a space in which a ubiquitous computing system has contextual awareness of a user and the ability to maintain consistent, coherent interaction across a number of heterogeneous smart devices" [3]. The goal of a CAIE is to assist the user with a seamless humanistic style interface. In order to create an environment that can succeed in this goal, we first have to know how a person would react, physically, to the assistance.

In this chapter, a brief introduction is presented as an overview of the experiment conducted for this thesis. In Chapter 2, related research and background information pertinent to this research is discussed. Chapter 3 provides a description of the

experimental setup and methodology. Chapter 4 discusses the methods used to analyze the results of the experiment. In Chapter 4, the results of the experiment are also presented in detail. Finally, Chapter 5 concludes the thesis discussing the trends and outcome, and also discussion of future works.

1.1 CAfFEINE Framework

The research in this thesis is based on research done by Dr. Saha who introduced a Framework named "CAfFEINE." CAfFEINE is an acronym for "Context-aware Affective Feedback in Engineering Intelligent Naturalistic Environments" [4]. A visual

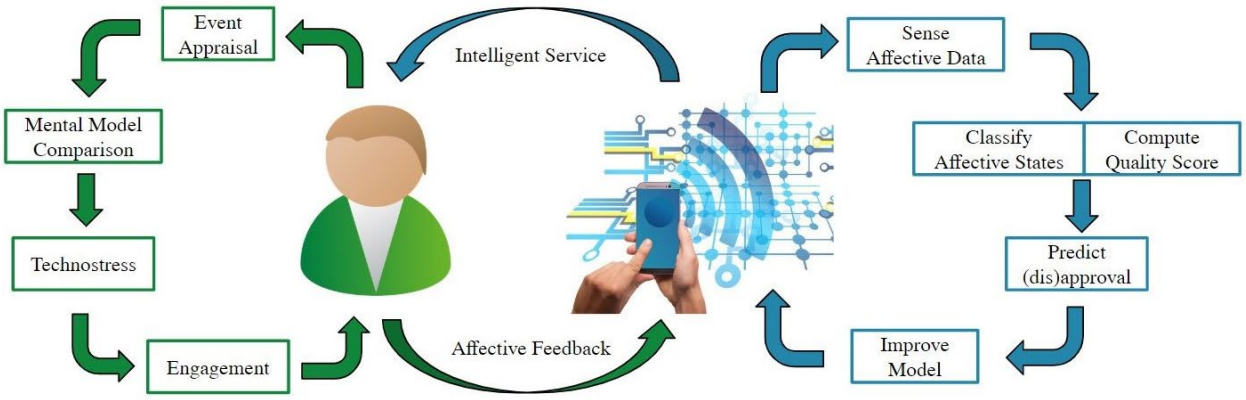

Figure 1.1 *Visual of the CAfFEINE Framework*

representation of the CAfFEINE Framework is shown in Figure 1.2. Dr. Saha's research was used to develop an algorithm that can estimate, with decent accuracy, whether a user defines a service from an intelligent environment as a "correct service" or "incorrect service."

An example of where the Framework would work in a current technology is with a voice-activated assistant like Siri. For example, a user asks Siri to call a friend named Don. Siri replies with, "Calling Mom." The user would perceive this as an incorrect

service and would have a physiological response; heart rate would increase, and possibly the user would start to sweat. If the user was wearing a smart device, such as a smart watch that was connected to the platform hosting Siri, Siri would detect the change in heart rate and sweat glands. Using the CAfFEINE Framework, Siri would reply to the user, "I'm sorry, I think I heard you wrong. What did you say?" Then the user would be able to correct Siri. The Framework is trying to model human interaction. For example, if you were to say something to your friend that may or may not hurt their feelings, you would be able to infer from their facial reaction and body language whether or not you did indeed hurt their feelings. In a sense, this is what the CAfFEINE Framework is trying to model; affect awareness. But instead of reading facial features and body language, the Framework is using physiological signals.

That being said, how on earth do you test such things? The research discussed within this thesis attempted to measure a person's reaction to a service provided by an IE, then struggled to classify those reactions. In order to test the CAfFEINE Framework, an experiment was constructed to provide a service to a user via an IE. Section 1.2 describes the experiment for the research in this thesis.

1.2 The Tangram Experiment

The CAIE used in this experiment was a mock version of an assembly line with an on-screen, step-by-step assembly guide. The initial intent of this experiment was to mimic an automated, interactive assembly instruction that could be seen in a manufacturing setting. However, because of the complexity of having to use tools and three-dimensional parts, a more simplistic model was used in order to focus on the testing of the CAfFEINE Framework. Therefore, a Tangram puzzle was used as a two-

dimensional representation of a part that needs to be assembled with assistance. The Tangram puzzle is a game that originated in China and consists of seven puzzle pieces of differing sizes and shapes [5]. The object of the game is to combine the seven shapes to create an overall silhouette given to the player, which was chosen to be a sailboat for this experiment as seen in Figure 1.1.

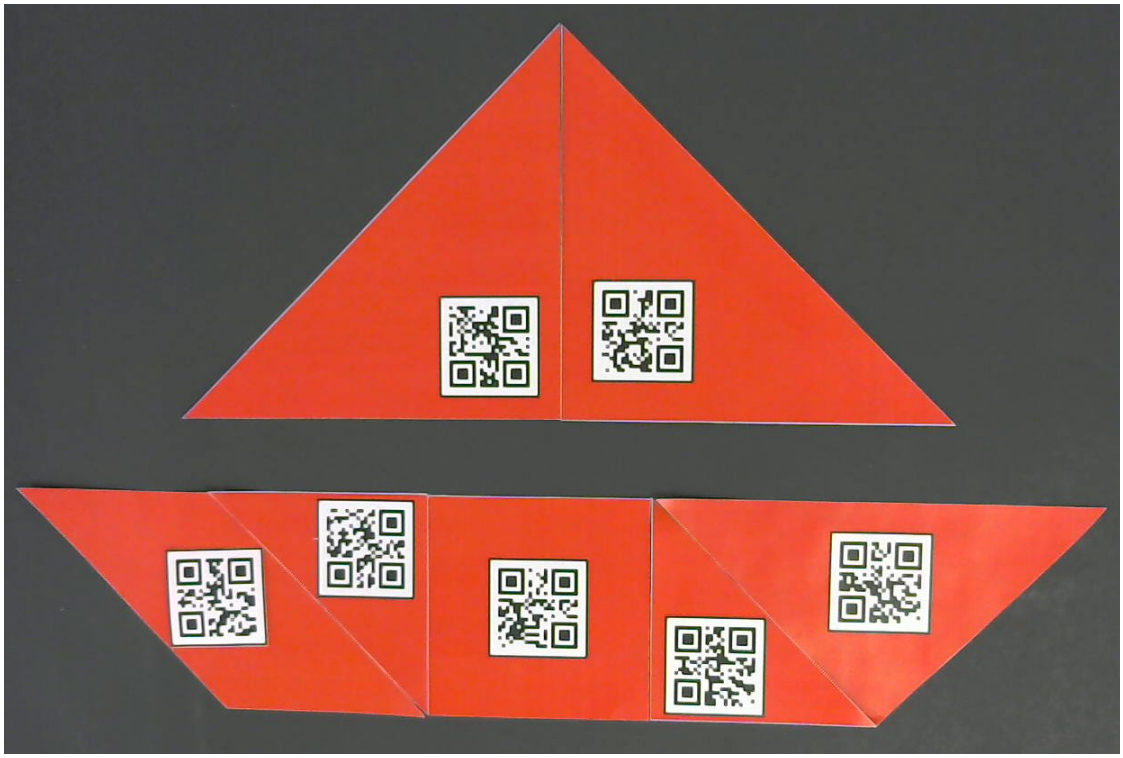

Figure 1.2 *Tangram puzzle with shapes arranged as sailboat silhouette*

The detailed experimental setup is discussed Chapter 3. However, step-by-step instructions were provided to the user on a computer screen on how to complete the puzzle. The user was asked to complete each step and then make a hand-raising gesture toward a motion-capture device. When the IE received that gesture, an overhead camera then analyzed the work space for the correctness of the user's puzzle. If the user's step

was correct, the user was asked to move on. If the step attempted by the user was incorrect, the user was asked to try that step again.

The definition of technostress is when a user receives a perceived, wrong service from an IE and has a physiological reaction. In order to capture the user's physiological response to a wrong service, a technostress stimulus had to be introduced at specified time intervals. In order to induce a technostress stimulus, specific steps in the process where the user had completed the step correctly, the IE told the user that the step was incorrect.

After data is collected from the Tangram experiment, a Support Vector Machine (SVM) will be used to predict, for each stimulus that was introduced, whether the user perceived that as a wrong service or correct service. An SVM uses data previously gathered as the training metric for future binary classification using a hyperplane between the class data [4]. Therefore, in order to train an SVM baseline, data had to be collected for each user. The baseline tests used for this research were the Paced Stroop Test (PST), white noise, and impulse. These tests are discussed in depth in Chapters 2 and 3.

1.3 Motivation

The CAfFEINE Framework, described in Dr. Saha's dissertation, outlines a context-aware intelligent environment that receives feedback implicitly from the user's physiological responses [4]. With my background being in manufacturing engineering, I wanted to construct an experiment that would both test the Framework and work within a real-world setting. The Empatica watch was used as a measuring device that could be used in a manufacturing setting without getting in the way of the user. Then intent of this research was to take the work done by Dr. Saha and move it towards a real-world setting.

Using affective computing in the manufacturing industry could streamline the process and improve turn-around time.

1.4 Focus in this Thesis

As can be seen in the literature that follows, the initial results of training the SVM on PST data then predicting on the Tangram experiment data were not as expected. Therefore, the research shifted from an analysis of the CAfFEINE Framework to a focus on the base-line data being used to train the SVM, or the PST, white noise, and impulse tests. There will be a discussion of the CAfFEINE Framework and the Tangram experiments, but the focus of this thesis is the analysis of the baseline data and how it can be used in the future to verify the usefulness of a user's data. Hopefully this will contribute to future research and analysis on data gathered for the implementation of the CAfFEINE Framework.

1.4.1 Research Question and Hypothesis

The original focus of this research was to take the research experiments that Dr. Saha had completed and migrate them to a real-world setting while verifying the results in his dissertation. For the research described in this thesis, the Tangram experiment was designed to verify the CAfFEINE Framework using a manufacturing assembly setting while also using a wearable device that would not hinder the progress of the user. In Dr. Saha's research he used multiple wearable devices to measure the user's physiological responses. However, in this research, the Empatica watch was used as a wearable device that could measure all physiological responses.

The hypothesis discussed in this thesis was, even though there was a difference in measuring devices and experimental environment, that these results would support the conclusions and results found in Dr. Saha's dissertation. The conclusion that was found in Dr. Saha's research was that an SVM could be trained on PST data, then the SVM could be used to predict on experimental data, with high accuracy, the reactions of users to a service provided by a CAIE. However, as stated previously, the research described in this thesis did not support this hypothesis.

Chapter 2 Background and Literature Review

This chapter provides background and related work on subjects and terms used for the experiment in this thesis. Section 2.1 describes the motivation for this thesis research. Section 2.2 describes the different components of the CAfFEINE Framework and related research. Section 2.2.1 describes research involving Intelligent Environments, which is one of the main components of the CAfFEINE Framework. In Section 2.2.2, related work around the concept of Affective Computing is discussed in detail. Then in Section 2.2.3, the term "technostress" is defined and past research conducted that supports the research in this thesis is outlined. Finally, in Section 2.2.4, the research conducted by Dr. Saha and his work on the CAfFEINE Framework is explained in greater detail.

2.1 Background

The research performed for this thesis was completed in support of Dr. Deba Saha's research. Figure 2.1, similar to the figure of the CAfFEINE Framework in Chapter 1, marks each step of the Framework denoted by the green squares. This research experiment was focused on the measurement of Step 2. As outlined in future sections, the research observed user interactions with an Intelligent Environment (IE) and measured their physiological response to the services provided by the IE. The physiological

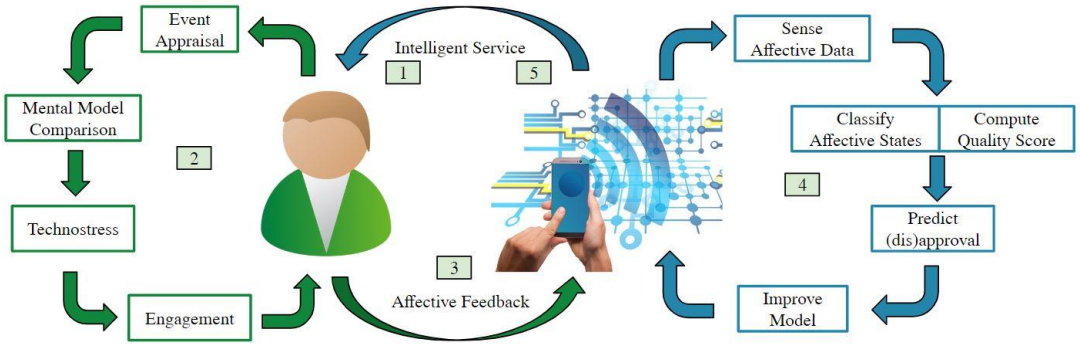

Figure 2.1 *Interaction-process Model of Service Delivery and Affect Generation in CAfFEINE Framework, derived from [1]*

response that was gathered throughout this research was electrodermal activity and blood volume pulse, both attained by each user wearing the Empatica E4 smartwatch.

2.1.1 Electrodermal Activity

Electrodermal activity (EDA) is used as a general term for "autonomic changes in the electrical properties of the skin" [6]. The Empatica smartwatch measures the skin conductance of the wearer. EDA is the most useful metric when inferring cognitive states [6]. Since EDA measurement is the most useful metric when analyzing the user's state of stress, it is an integral part for categorizing technostress in this research.

An EDA signal is made up of both a tonic and phasic component. The tonic level of skin conductance is also known as the Skin Conductance Level (SCL), and the phasic change also referred to as the Skin Conductance Response (SCR). The tonic level of the signal dictates the overall level of the signal and the change of slope over time [6], while the phasic component of the signal refers to the fast changes of the signal [6]. For each individual, or user, the tonic portion of their EDA signal "generates a constantly moving baseline" [6]. Therefore, by itself the SCL of the EDA signal cannot be used as the sole

variable to analyze the usefulness of the signal. This also led to the decisions outlined in Chapter 4 for analyzing the data; the Support Vector Machine (SVM) cannot be trained on all users but must be trained for each individual user.

2.1.2 Blood Volume Pulse and Heart Rate Variability

The Empatica smartwatch used during experimentation can measure BVP, but that measurement alone is not useful. BVP is used to derive Heart Rate Variability (HRV), which is used in the data analysis in Chapter 4. HRV is the "oscillation of the interval between consecutive heartbeats" [7]. The measurement is used as a marker for mental exertion and stress [7]. Therefore, HRV will be an added variable to the research discussed in this thesis.

2.2 Literature Review

The following sections in this chapter will discuss research being done in the fields of context-aware intelligent environments, affective computing, technostress or neuroscience involving frustration with technology, and Dr. Deba Saha's research involving the CAfFEINE Framework.

2.2.1 Intelligent Environments and Technostress

The research discussed here models automated assembly instructions to examine a user's response to a wrong service provided by a CAIE. Past and concurrent research shows when a person interacts with an intelligent system, they have an expected outcome in mind [8] [4]. When the service provided by the CAIE differs from the user's expected outcome, there is a measurable, physiological response. This will henceforth be referred to as technostress [4]. Technostress is caused when an intelligent system does not

produce the expected service to its user. The term technostress is not a claim that the user is "stressed" by definition. The term is used to denote a physiological change measured when a user is taken off guard by an unexpected outcome. "A leading cause of technostress is 'achievement stress,' which is observed to be heightened in system failures during time-pressured tasks, i.e., tasks having hard-deadlines associated with them" [4].

2.2.2 Intelligent Environments

Intelligent Environments are the building block for the CAfFEINE Framework and this thesis. The experiment conducted for this thesis uses an IE for the users to interact with, therefore providing the results discussed in Chapter 4.

Wrede and his group define IEs by defining each word individually. Their definition for "Intelligent," for this application, is Artificial Intelligence [2]. They define "Environment" as all physical space surrounding a user or person [2]. That means that any artificial or virtual environments are not included in this definition. Therefore, an Intelligent Environment is an environment that makes use of multiple "networked controllers" with software, which is "self-programming" and "pre-emptive," to manage an interactive space that enhances the user's experience [2]. This group cited M. Weiser with his prediction that "technology is gradually disappearing from our cognitive front" [2]. However, they go on to say this fact alone will not support an Intelligent Environment, but that there will need to be a "paradigm shift" for IEs to thrive [2].

There are other areas of study that will be needed to support Intelligent Environments: pervasive/ubiquitous computing, smart environments, and ambient intelligence. Pervasive/ubiquitous computing research studies computing services that

travel with the user across multiple environments and are context-aware. Ubiquitous computing is more related to the human-computer interaction, while pervasive computing focuses more on the device, networking, and production of processed data [2]. A Smart Environment is an environment that is equipped with a range of sensors and processors with which a user can interact [2]. A Smart Environment is different than an Intelligent Environment. Intelligent Environments are constantly analyzing data and attempting to infer the user's intent, while a Smart Environment is simply interactive [2]. Ambient Intelligence is the software that assists people with their daily routines with minimal distraction or interruption. Intelligent Environments are built on these areas of study and intend to integrate the Smart Environment with Ambient Intelligence based on the availability of data collected by Pervasive/Ubiquitous Computing. The IE is the building block to the CAfFEINE Framework proposed by Dr. Saha [4]. The experiment done for this thesis used Dr. Saha's Framework. Wrede defines nine principles that he and his team believe are key for an Intelligent Environment to work properly [2]. Of those nine, the following four principles are the main focus of the CAfFEINE Framework:

- "To achieve its goals without demanding from the user(s) technical knowledge to benefit from its help
- To prioritize safety of the user(s) at all times
- To have autonomous behavior
- To be able to operate without forcing changes on the look and feel of the environment or on the normal routines of the environment inhabitants" [2]

These principles can be categorized into the three other areas of study that were described earlier. The reason that they focused on AI as the intelligence part of this paper

is because the system has to be able to read human social cues, and a simple sensor-processor pair will not pick up on that. That is the reason for the Smart Environment, where multiple sensors work together to identify those cues. Presently, there is no umbrella system that contains all the necessary sensors to achieve this goal. Instead there are systems that are made for specific applications.

When focusing on the user, an IE should be able to assist a user no matter their age, mental capacity, or familiarity with technology. In other words, an Intelligent Environment must be "People Oriented," where the user makes all the decisions and the technology simply suggests decisions [2]. The environments in which these kinds of systems can be used will differ greatly. The authors discuss two kinds of environments: closed and open. A closed environment, for example, would be the user's house, office, or car. While the open environment would be streets, parks, or the ocean. They specify these two different kinds of systems because of the complexity for each. A closed environment would be easier to install sensors and software for a fully immersive IE. While the open environments would be much more difficult to place sensors and install a software that could handle the constantly changing variable of the system. Another factor to consider for the system is the IE's perception of what is actually going on. One example Wrede and his coworkers give is a pressure sensor on the seat of a couch [2]. This sensor by itself can only tell if there is weight on it. But does that mean that there is a person sitting there, a dog on the couch, or has someone set something heavy there? If it is a person present on the pressure sensor, are they sitting, are they lying, or have they fainted and need medical attention? A pressure sensor, as a single sensor, cannot predict any answers to these questions, which is why there is a need for multiple sensors in an IE.

At what point, however, are there too many sensors providing an overabundance of data? Once the appropriate number of sensors has been addressed, privacy must also be discussed. Current research is still exploring these questions as they pertain to Intelligent Environments.

In their concluding remarks, Wrede and his research team point out that the effort for a fully functional Intelligent Environment will be a multidisciplinary collaboration. When it comes to the technology, computer scientists and computer engineers will be at the forefront of the design and implementation. However, for different applications, there will be a need for professionals in different fields. An example they give is a Smart Home. Even though it seems a simple task, there will need to be input from such experts from the medical, architectural, and social work fields. It is because of this multidisciplinary involvement that they see a need for "incremental development" [2]. In other words, every piece of technology needed for a specific environment is not thrown in at once, but slowly introduced to the users with one or few new features at a time.

2.2.3 Affective Computing

Affective computing refers to the recognition, interpretation, and response to human feelings and emotions based on data gathered by sensors: "facial expressions, body language, voice, physiological responses, etc." [9]. Affective computing has been used in multiple research efforts to perfect the system. For example, Liu et al. [10] monitored "the autonomic nervous system (ANS) responses" of a user's anxiety levels during a game of Pong, and used that data to adapt the difficulty of the game. When discussing affective computing, it is important to consider recognition accuracy. Recognition accuracy, as it pertains to affective computing, is the percentage the system

recognizes a class of affect correctly compared to what is truly experienced by the user. In our research, affective computing will be when our system recognizes the user to be stressed or frustrated when we intentionally tell them they are incorrect, when in fact, the user is correct. Novak et al. also found that using more than one type of physiological sensor, such as BVP and Galvanic Skin Response (GSR) used in our experiment, increases the accuracy. In their experiment they used an adaptive game where the difficulty would increase or decrease depending on the perceived affect and the user's answer at the end of each round. In other words, at the end of each round the system would guess if the round that the user just finished was too easy or too difficult for them. If the user agreed, the difficulty would change in that direction. If the user disagreed, the difficulty would change in the other direction. The user's choice to agree or disagree with the system was the stimulus used to record recognition accuracy. For the experiment in this thesis, the stimulus was the screen that told the user if they were correct or incorrect. Their results were spread out across the spectrum from 0% to 100%. One of the contributing factors to this spread is that a user could exit the game after any round they wished without answering whether or not that round was easy or difficult for them. They found that the more frustrated the users were with the difference between the predicted affect and their perception of the round, the more likely the user was to quit the game prematurely. This can be expected; if a person is not getting the results they expect out of a system, they are less likely to use it all together. Novak et al. did discover that in order for a user to keep playing the game, there needed to be at least 80% recognition accuracy [9].

2.2.4 Technostress

In 1984, Brod defined the term technostress as a "modern disease of adaptation caused by an inability to cope with the new computer technologies in a healthy manner" [11]. Technostress is both a physiological and emotional response to stress brought on by technology. The use of information and communication technologies (ICT) in the workplace can cause employees stress. When they refer to ICT, they are referring to technologies such as "cell phones, email, and instant messaging" [12]. The stress that comes from ICT adds to the stress already experienced by employees with their general responsibilities and is affecting employee health.

Researchers have divided up technostress models into three categories: "transactional and perceived stress, biology, and occupational health" [12]. The transactional category is a comparison of workplace tasks that induce stress and specific factors that inhibit the impact of ICT. Technostress has the possibility to reduce "job satisfaction, organizational commitment, and employee outcomes" [12]. Technostress also leads to the fear of being replaced by technology or other workers who understand the technology better. Therefore, in order to alleviate technostress in the workplace, there must be constant training on all new and used technology. Because of the research being done on technostress, companies are now equipped to help employees cope with new technologies being used in the workplace.

According to other researchers [13] [14], technostress is defined as a "psychosomatic problem" [12]. These researchers believe that a user's, or employee's, physiological response to the technology in the workplace needs to be monitored. When technostress is perceived while monitoring an employee, the system can return to a "set

point" at which the employee feels more comfortable with the technology [12]. According to past research, a physiological response is noticeable before a person perceives the use of ICT to be stressful. Therefore, the system can return to a less stressful state before the user becomes perceivably stressed. There are two drawbacks to this category of research. First, it only monitors when heart rate and hormones increase, which occurs both when the user is stressed and when the user is excited that the technology is working the way the user assumes it to work. Second, there will be jobs in which the user cannot wear sensors to track their physiological responses.

Therefore, the use of technology in the workplace should be evaluated before and on a consistent basis throughout the career of an employee, using said technology. Technology in the workplace should be there to assist in a way to make the job easier, not cause stress. That is one of the main reasons that our experiment focused on an assembly/manufacturing process. Previous personal experience in a manufacturing setting provided insight as to the complexity and stressful nature of assembling product. That is where Dr. Saha's research and Framework was used.

2.2.5 Paced Stroop Test

The use of a paced Stroop test for the experiment described in this thesis is based on research conducted by P. Renaud and J. Blondin. In their paper [15], they completed three types of experiment: "self-paced, externally-paced, and fast externally-paced" [15]. Externally-paced simply means that the questions progressed at a specified interval of time instead of letting the user determine when to move on to the next question. According to [15], the heart rate of the test subjects increased from a resting state but was not dependent on the pace of the test. However, the number of skin conductance

responses varied with the pace of the test [15]. Renaud and Blondin were able to measure more frequent changes in skin conductance as the amount of time per questions decreased.

2.2.6 CAfFEINE

Dr. Saha's research revolves around Context-Aware Intelligent Environments (CAIE). As stated earlier, CAIE is an environment in which a user interacts with multiple sensors and interactive computers that will infer the user's needs and predict their state of being. As the CAIE provides services to the user, the system is still not robust enough to provide "appropriate" services at the appropriate time [4]. Achieving this correct appropriateness is the most important step for a CAIE to complete. This step in the process is where Dr. Saha and I have focused our research. The object of Dr. Saha's multiple research experiments was to measure the user's physiological response to the services provided by the system. These responses were used as implicit feedback to the system to access the correctness or appropriateness of the services provided. In order to gather data on the users' physiological response to the CAIE, wearable sensors were

worn by each user. He used real-life scenarios to introduce a CAIE system and then

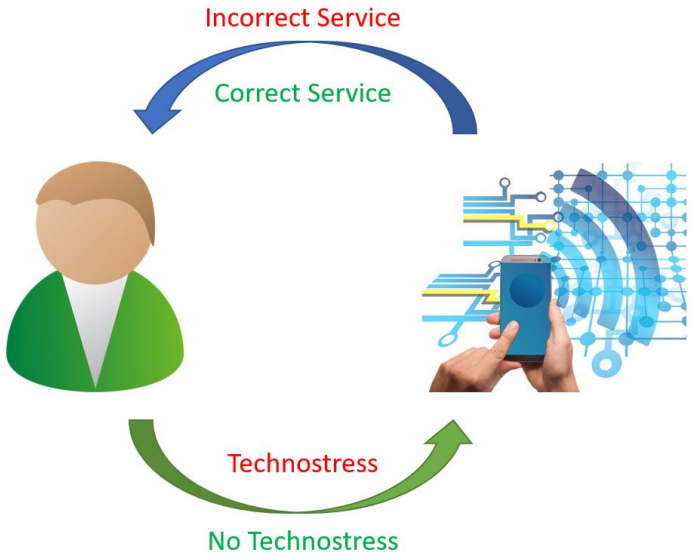

Figure 2.2 *Services provided and Technostress observed*

provided the users both correct and incorrect services, as shown in Figure 2.2 to monitor the difference between when a user does and does not experiences technostress.

After the experiments were concluded, a binary support vector machine was trained for each individual user. An SVM is a classification algorithm, regularly a binary classification, that calculates a "maximum margin classifying high-dimensional hyperplane" between each group of data [4]. The training data for each user was gathered through a Paced Stroop Test (PST). The Stroop test is a timed test where the user has to match the text color on-screen to the options provided.

As shown in Figure 2.3, the word is "BLUE" but the font color is "RED." Therefore, the user would select "RED" from the radio buttons on the left. There are 60

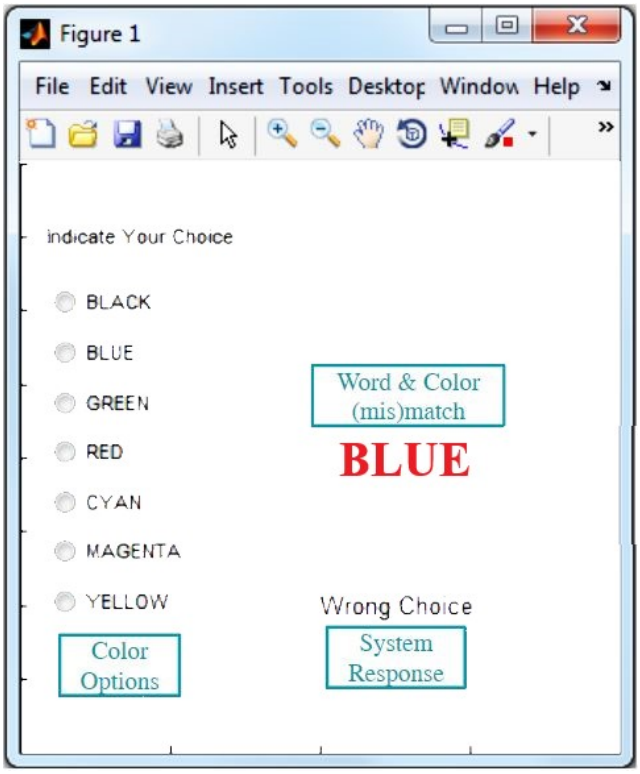

Figure 2.3 *Screenshot of Paced Stroop Test*

combinations for the user to go through. Each combination is on the screen for only three seconds. The test is broken into thirds: for the first third the word and font color do not match, the middle third has matching word and font color, and the last third is the same as the first. The results from the PST give "baseline" data to be used for training the SVM [4]. Dr. Saha found that when training the SVM on all users and then predicting on individual users, the results were not as accurate as expected. However, when training the SVM on a user-by-user basis, the prediction precision increased.

Chapter 3 Experimental Methods

This chapter presents the methods, hardware, and software used for the experiment outlined in this thesis. Section 3.1 describes the physical space in which the user interacted. Section 3.2 describes the hardware used to track the user's movement. In Section 3.3, there is a description on the use of an overhead camera and how the system validated the user's progress. Section 3.4 outlines the use of the Paced Stroop Test. Section 3.5 discusses the smartwatch used to gather physiological data on each user. In Section 6, the use of a personality test in this experiment is defined. Lastly, Section 3.7 contains a brief discussion on the demographics of the users.

The software that was used for the experiment was Matlab. Matlab was chosen because of the seamless interaction with a webcam and a Kinect for Windows and the ease of creating the user interface.

3.1 Experimental Setup

The experiment for this thesis is a user study to simulate automated instructions for industrial assembly lines, in which a context-aware intelligent environment (CAIE) helped the user perform a task. A Kinect for Windows monitored the user's movements, and an overhead camera monitored the user's progress. The experiment was purposefully designed to give the user a wrong service, as described in Chapters 1 and 2, at specified intervals and the user's reaction was measured. Throughout the entire experiment, the user wore an Empatica smartwatch.

The mock assembly part for the experiment was a Tangram puzzle. Tangram is a puzzle game for children that consists of seven unique shapes. These shapes are arranged together to form an overall silhouette that is pre-defined. There are varying difficulties of silhouettes, which led us to use this puzzle as a mock setup for an assembly line. As with an assembly line part, assembling a Tangram puzzle correctly is not always intuitive.

As stated previously, the experiment implemented an assembly instructional, which gives step-by-step instructions on how to construct the puzzle, shown in Figure 3.1. Our intelligent environment displayed each step, one at a time, on a computer screen for the user to follow. At the end of each step, the user was instructed, prior to starting the assembly process, to raise their hand. This motion signaled to the system that the user

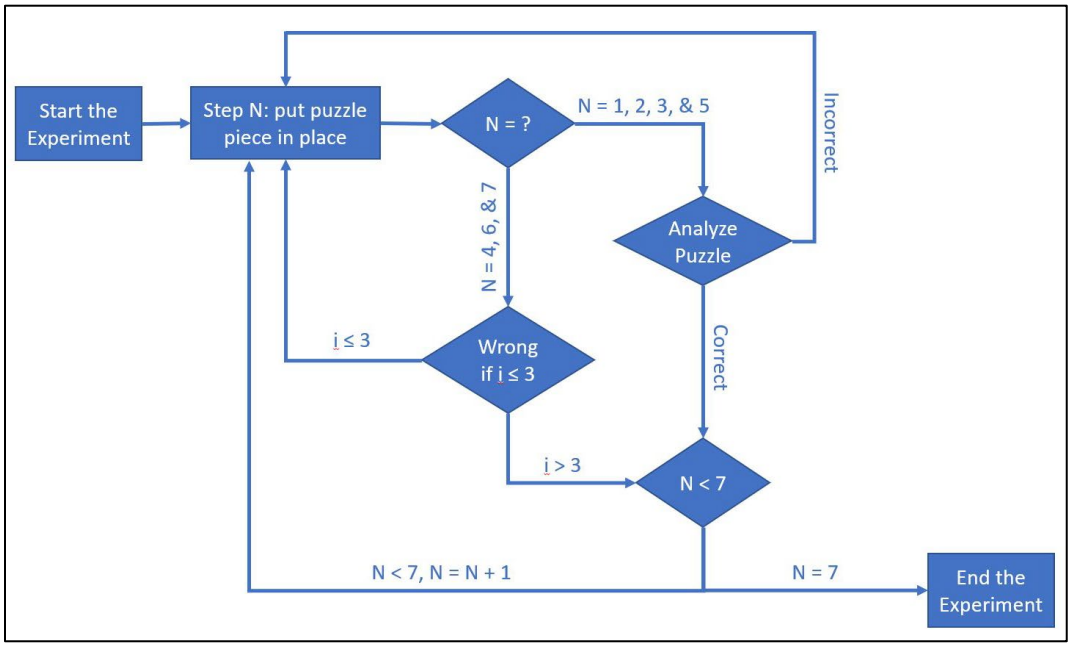

Figure 3.1 *Tangram Experiment Flow Chart*

was finished with that step. The system would analyze their progress and then let the user know, on screen, whether or not they had completed the assembly step correctly. If

incorrect, the user was asked to attempt the step correctly before moving on to the subsequent steps. There were seven steps in total, and the user would progress through until the entire assembly of the Tangram puzzle was completed.

3.2 Kinect for Windows

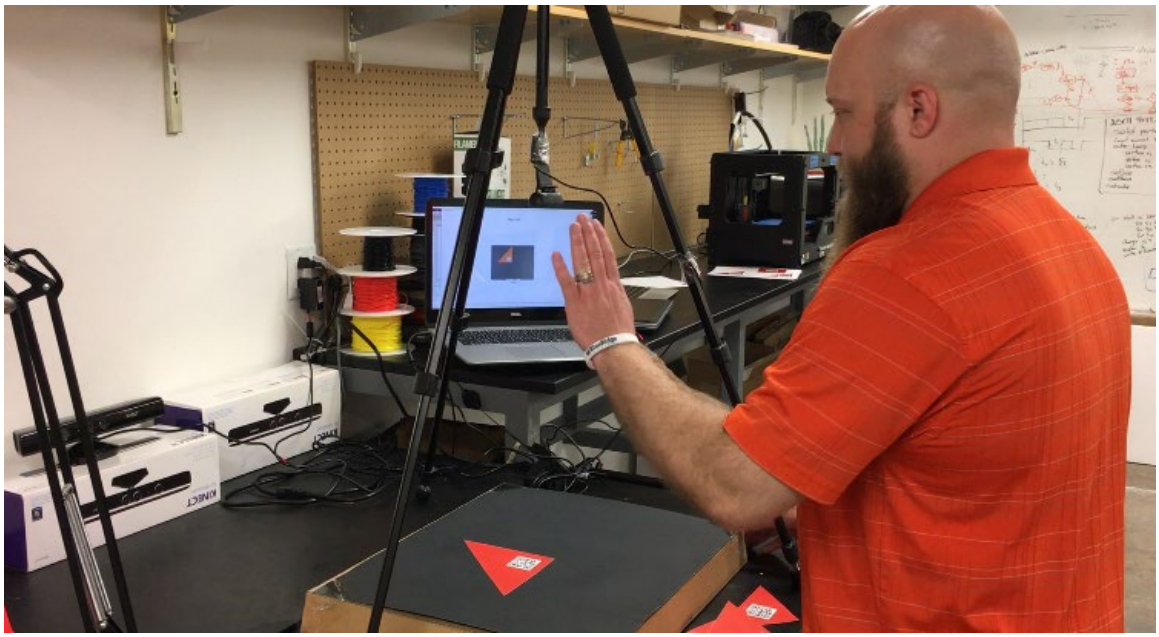

Figure 3.2 *Kinect for Windows v1*

The Kinect was positioned in front of and facing the user, as seen in Figure 3.2. The Kinect comes with built-in commands to run a "skeleton tracker" of a person's motion using infrared depth measurements. Once user's movements were trackable, the system could determine if the user was raising their hand or not. Raising of the user's hand was used as an explicit indication to the system that the user's intent was to move on to the next step. Tracking the user's arm motion was a first step in analyzing the intent of the user. Future experiments would make use of this tracking system to determine the user's intent to move forward with implicit feedback (i.e. the user stops moving their arms).

3.3 Camera and Correctness Validation

The user was asked to assemble the Tangram puzzle on a flat, black surface, referred to as the "work mat." The puzzle pieces were red, for high contrast on the black

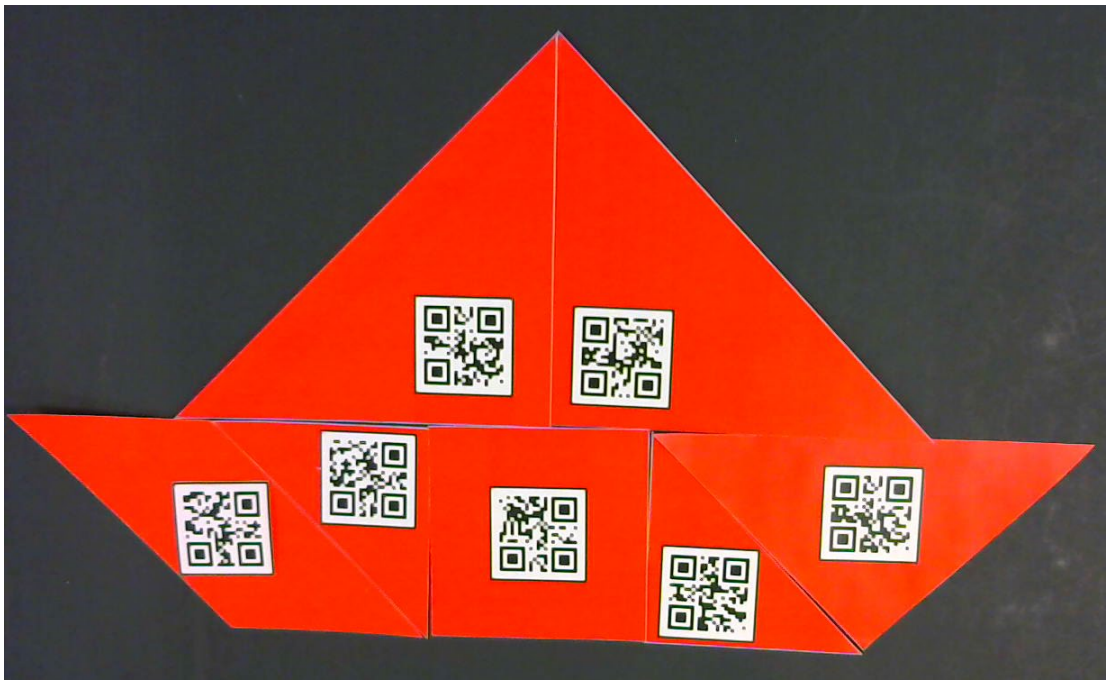

Figure 3.3 *Tangram shapes that form the silhouette of a boat*

surface. Each of the shapes had a distinct 1.5 by 1.5-inch QR code printed on it as well. These QR codes were used to identify the shape and position on the work mat. In order to validate the progress of the user during the experiment, we mounted a Logitech Webcam above the work mat. The camera took a picture of the work mat and analyzed the "correctness" of the Tangram puzzle. The algorithm was able to find the outline, area, centroid, and classification of the shapes. The software output if the current step was missing any shapes, if there were too many shapes, and if they were in the correct orientation. From those results, feedback was given to the user if their current progress was correct or if they needed to try again. The algorithm used to analyze the correctness of a user's progress was constructed within a class project with Vinidhra Sivakumar and

myself. Matlab was used for the GUI, the control of the webcam and Kinect, and also used for data analysis.

3.4 Paced Stroop Test (PST)

As described in Chapter 2, in order to classify a user's state, stressed or non-stressed, data from the Paced Stroop Test was used as training for the SVM. The Stroop Test, created by J.R. Stroop in 1935 [16], has been used in research and experiments as a "psychological or cognitive stressor" [15]. The test presented color-description words to

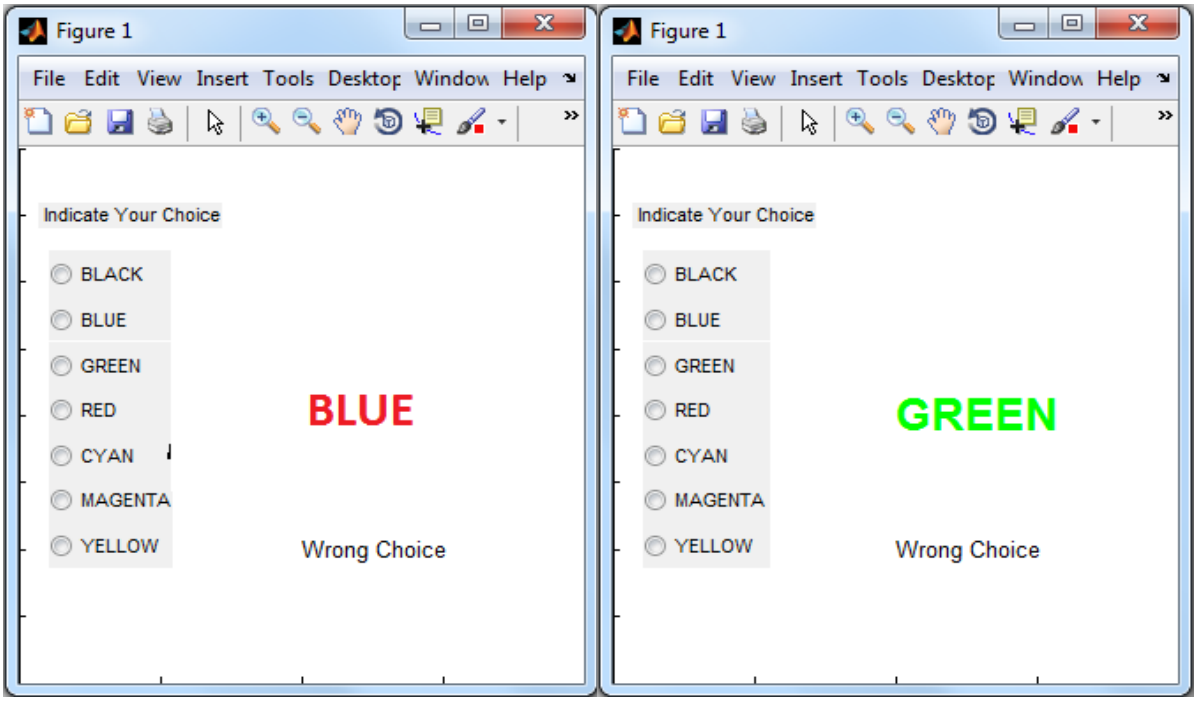

Figure 3.4 *(Right) Color-description word and font color match, (Left) Color-description word and font color do not match*

the user. The word on screen was either in matching font or in non-matching font; an example is found in Figure 3.3. The user was given three seconds to answer each word-color pair question. There was a total of 60 questions. The first and last 20 questions were non-matching, the middle 20 were matching. The first 20 questions were there to allow the user to get acclimated to how the test works, therefore the physiological data for the

first 20 questions were discarded in the data analysis. The middle 20 questions, with matching color and description, were used as a baseline for the user's non-stressed state. The final 20 questions were used for categorizing the user's stressed state of physiology.

3.5 Empatica Smartwatch

In order to gather the physiological data needed for this experiment, each user wore the Empatica E4 Smartwatch. The Empatica Smartwatch is equipped with a Photoplethysmography (PPG) sensor for measuring BVP, a pair of silver-plated electrodes for measuring skin conductance or EDA, an infrared thermopile for measuring body temperature, and a three-axis accelerometer to measure the movement of the wearer's wrist [17]. For the experiment conducted for this thesis, the results described in Chapter 4 only take EDA and BVP into consideration. EDA and BVP are discussed in more detail in Chapter 2.

3.5.1 BVP Data

EDA data was just one stream of data that was gathered by the Empatica E4 smartwatch. BVP was also gathered and used for data handling and evaluation. Blood volume pulse is the measure of arterial blood pulses, under the skin, emanating from the user's heart beating [17]. Using a combination of Toolbox for Emotional feAture extraction from Physiological signals (TEAP) [18] functions and the functions provided by Dr. Saha, heart-rate variability (HRV) features were extracted. Features such as mean heart-rate (HR), standard-deviation of N-N intervals, etc. [4]. These features, along with the EDA features, were used for training the SVM and predicting the classification of the experimental data.

User		NS	S	G-Score	F_β-score
1	NS	6	1	0.27	0.49
	S	6	1		
2	NS	6	1	0.57	0.79
	S	6	4		
3	NS	6	1	0.38	0.65
	S	7	2		
4	NS	7	0	0.00	0.00
	S	9	0		
5	NS	4	3	0.46	0.50
	S	4	3		
6	NS	7	0	0.45	0.96
	S	4	1		
7	NS	6	1	0.00	0.00
	S	6	0		

Table 3.1 *Results without using EDA*

EDA data alone could be used to train the SVM and then classify experimental data. As shown in Table 4.1, the SVM could not properly predict the results without EDA data. In order to enhance the results, more data points were needed. According to J. Kim and E. Andre [6], both EDA and ECG data correlate directly to a person's emotional state. Therefore, for the research discussed in this thesis, the EDA and heart rate data were paired for training and classification of technostress.

3.6 Personality Test

After the conclusion of the experimentation, a Ten-Item Personality Inventory (TIPI) was filled out by every user. As the title of the personality test implies, there were ten questions that the user needed to rate on a scale from 1, disagree strongly, to 7, agree strongly. The answers the users provided were then calculated into a score for each

personality type: Extraversion, Agreeableness, Conscientiousness, Emotional Stability, and Openness to Experiences. According to [19], "personality traits affect the levels of strain experienced by the bank employees." This study was based off of the work done by Davis [8]. This means that for each personality there will be a noticeable difference in their physiological response to technostress.

3.7 White Noise and Impulse

Before the beginning of any testing for the PST or the Tangram experiment, two baselines were assessed for each participant. One baseline for a non-stressed state, and

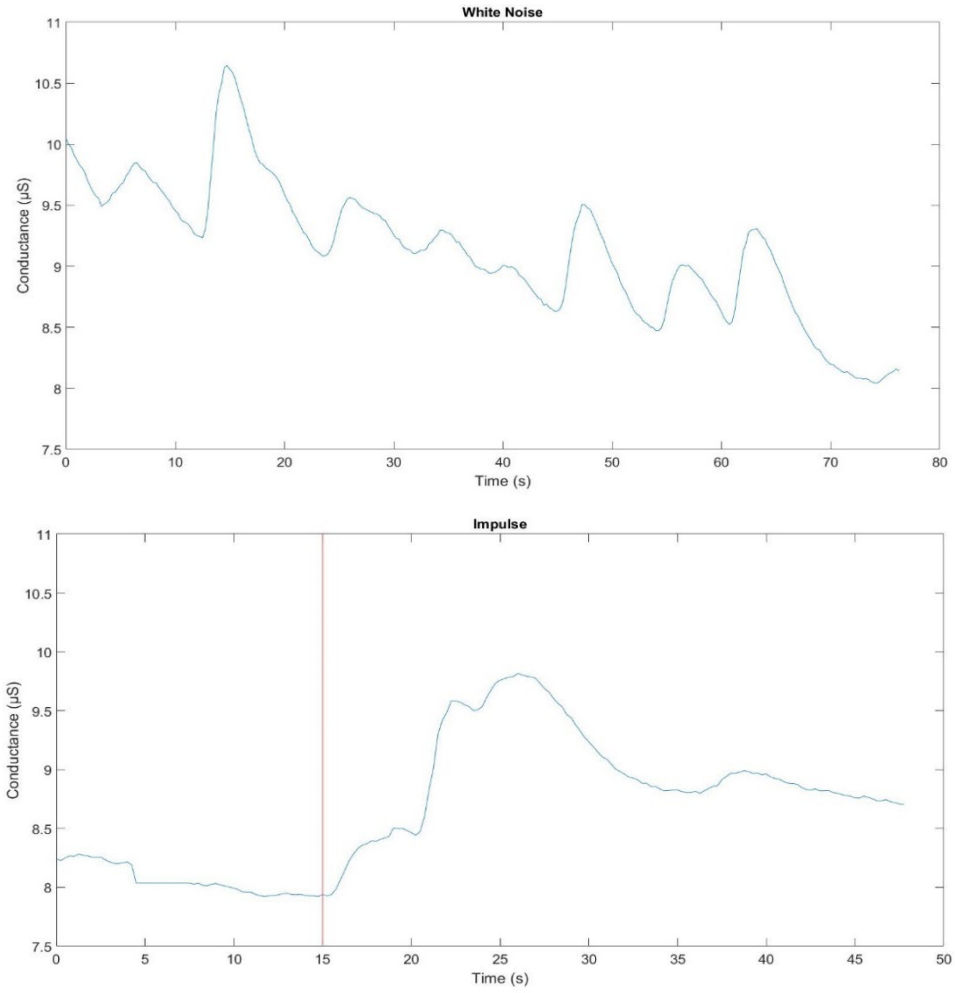

Figure 3.5 *White noise baseline reading (top) and Impulse baseline reading (bottom) with balloon pop stimulus marked*

another baseline for a stressed state. In order to measure a non-stressed state, each participant was asked to listen to 30 seconds of white noise played through noise-cancelling headphones. The raw EDA data for white noise is shown in the top graph of Figure 3.5. Following the audio clip of white noise, there was a 15-second clip of complete silence with the sound of a balloon being popped five seconds into the clip. The balloon pop represents the user's stressed state. The user's physiological reaction to the balloon pop stimulus is shown in Figure 3.5 in the bottom plot.

The non-stressed state is referred to in this thesis as White Noise (WN). The stressed state, caused by a balloon pop, measures the user's response to a stimulus [4]. The EDA graph of the raw GSR data resulting from this single stimulus will align closely to derived Impulse Response Function. Therefore, in this thesis, the data gathered from the balloon pop audio clip is referred to as Impulse.

3.8 Confusion Matrix

The results predicted by the SVM were a user-wise confusion matrix with a G-score and an F_β-score. The confusion matrix is an association of the predicted and actual results [20]. There are four quadrants shown in Table 3.2 below: in the top left is the number of correct negative predictions or true negatives (TN), in the top right is the number of incorrect positive predictions or false positives (FP), in the bottom left is the number of incorrect negative predictions or false negatives (FN), and in the bottom right is the number of correct positive predictions or true positives (TP).

User		NS	S	G-Score	F_β-score
1	NS	7	0	0.76	0.99
	S	3	4		

Table 3.2 *Confusion Matrix Example*

The G-score, also known as the Fowlkes-Mallows index, is a metric that measures the similarity between two ranking clusters [21]. The equation used to calculate the G-scores is $G = pr$, where p is the precision and r is the recall or sensitivity [4]. The F_β-score is, by definition, the harmonic mean of the sensitivity, r, and the precision, p [22]. The equation is $F_\beta = \frac{(\beta^2+1)pr}{\beta^2 pr}$, where β is a metric that controls the balance between p and r [22]. According to Dr. Saha [4], in the CAfFEINE Framework β is balanced to "reward" low FP scores, therefore the Framework uses a β value of 0.1. For both the G-scores and the F_β-score, the closer the value is to 1.00, the better the results are for that user [4] [21] [22].

Chapter 4 Analysis and Results

This chapter discusses the results and the methodology of analysis for this experiment. Section 4.1 discusses the initial results of the research using the entire PST data set for training. Those results did not support the research hypothesis. Additional analyses were performed on the data used for training the SVM to better understand why the hypothesis was not supported. For reasons discussed in Section 4.2, only the last two sections of the PST were used to train the SVM. The overall results did improve, but were not predicted accurately. Section 4.3 narrows that training data to only the last ten questions for each of those sections of the PST data set. The results were still not in support of the research hypothesis, therefore more analysis of stimuli window size was conducted. As details in Section 4.4 discuss, the window size for analyzing each individual stimulus response was adjusted from [0 6] seconds to [-1 4] seconds. Adjusting the window analysis size did improve the results, however not enough for conclusive results. Since the results thus far were not supporting the CAfFEINE Framework proposed by Dr. Saha, the SVM training data were investigated. Section 4.5 shows the results for testing the SVM using PST data as both training and predicting data. After these results provided a lack of clarity, it was decided to use different base-line data to train the SVM. In Section 4.6, the baseline data for training the SVM was switched to the White Noise and Impulse data set. These base-line data sets were not useful either. Therefore, in Section 4.7 exploration of the baseline data sets and determination factors of the data sets were examined. Finally, Section 4.8 discusses the results of each investigation as a whole in a single table. Each section discusses what was learned from each individual analysis performed.

4.1 Initial Results: Train whole PST – Predict Tangram

The physiological data gathered for this thesis research were preprocessed using Dr. Saha's altered version of Matlab subroutines originally developed by Jaimovich [4] [23]. These subroutines were designed to extract features and detect and remove any abnormalities for EDA and BVP data gathered from the Empatica E4 smartwatch [23]. As discussed in Chapter 3, BVP data were only used as support data for training the SVM. In-depth analysis of the EDA data was the focus for this research because the data varies greatly from user to user.

Once the data was preprocessed and all necessary features were extracted, an SVM classified events into two classes, stresses (S) and not stressed (NS). User-wise SVM training is used for experimental classification because every person's physiological response is different [4]. Therefore, for this experiment, a different SVM was trained using each user's baseline data and then predicted on that same user's Tangram data. However, as the rest of this chapter discusses, the hypothesis was not supported by the experimental results.

For EDA processing, Dr. Saha's feature extraction and reduction methods were used [4]. According to Saha, there are "fourteen features from GSR and ECG time-series data, which have been reported in literature as distinguishing for stress related studies" [4]. He goes on to explain that the features that are most important to analyze and note fall into a six-second window immediately after the stimulus onset. The reason the viewing window is six seconds is discussed by Figner and Murphy [24]. Their research discusses that the window has to be long enough to capture the user's reaction to the stimulus, but short enough to not capture any non-related stimuli [24].

The EDA data from user 1, which the Empatica smartwatch gathered during the PST, can be seen in Figure 4.1. Although not true for every user, the graph shows a

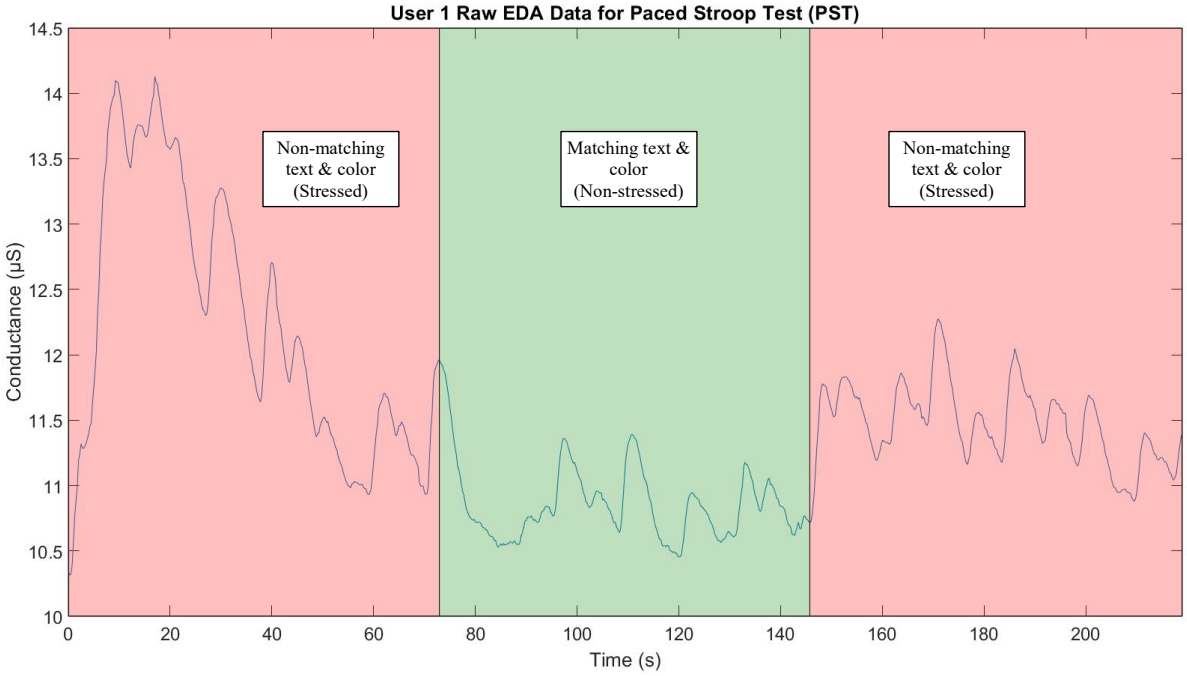

Figure 4.1 *User 1 Raw EDA data for PST with regions highlighted for matching (green) and non-matching (red) font color and font description*

distinct trend. Exploring the raw EDA data from the Empatica smartwatch, it was observed that the average for the three sections of the PST was different. In the first section, highlighted in red, the EDA data starts with higher readings and slowly tapers off. The reason for the high peaks at the beginning of this section can be attributed to the user's first interaction with the test [25] and the stress of the non-matching word description and font color [15]. Dawson et al. explain that when measuring EDA responses it is crucial to control the experiment where only one variable is changing [25]. In the PST, shown in Figure 4.1, one could argue that two variables were being changed in the first section; the user getting used to the test and the questions being presented every 3 seconds. But once the user got used to the test, only one variable was changing; the questions changing every 3 seconds.

The entire data set from the PST was used to train this SVM. Then the SVM predicted on the Tangram experiment data, resulting in the data in Table 4.1. These initial prediction results were not accurate.

User		NS	S	G-Score	F_β-score
1	NS	7	0	0.76	0.99
	S	3	4		
2	NS	4	3	0.48	0.57
	S	6	4		
3	NS	4	3	0.50	0.57
	S	5	4		
4	NS	5	2	0.54	0.66
	S	5	4		
5	NS	5	2	0.38	0.50
	S	5	2		
6	NS	7	0	0.45	0.96
	S	4	1		
7	NS	6	1	0.47	0.66
	S	4	2		

Table 4.1 *Original Results: Training SVM on entire PST and predicting on Tangram*

The SVM for User 1 predicted that the user was NS through the entire experiment. These results for User 1 were known to be incorrect because the system introduced stressful stimuli at given intervals. Similar conclusions can be drawn throughout the table. These results, along with the conclusion drawn about the first section of the PST, led to focusing on the last two sections of the PST, with each section having 20 questions.

4.2 Train 2/3 PST – Predict Tangram

Since the original results did not coincide with expected results from Dr. Saha's CAfFEINE Framework research, the focus of the research changed from testing the Framework to analyzing the data used for this research. The first consideration for analysis was the amount of data used to train the SVM.

When considering the amount of data used to train the SVM, it was observed that there would be double the data points if both red highlighted sections in Figure 4.1 were used. Due to the stress of the user getting accustomed to the test, the questions being non-matching, and the data set having two sections of non-matching and only one section matching, it was decided to ignore the first section and train on the last two sections of

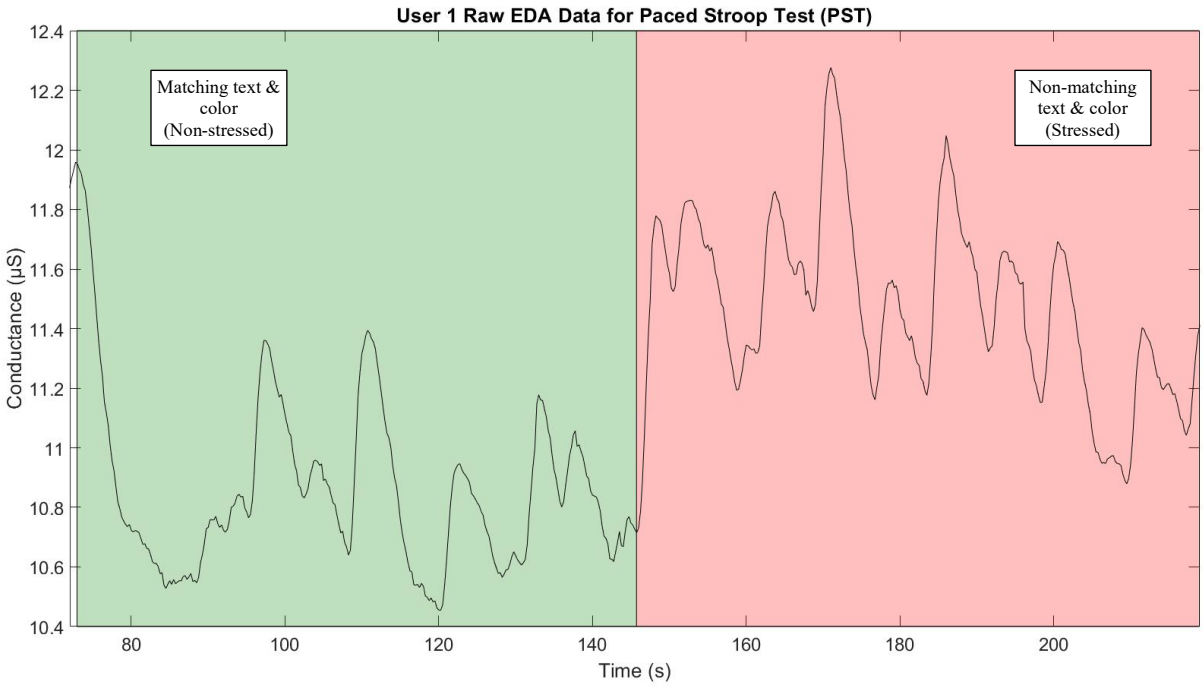

Figure 4.2 *User 1 Raw EDA data for the last two sections of the PST*

the PST in the next analysis.

Focusing on the last two sections of the PST, shown in Figure 4.2, there were some noticeable artifacts that should be pointed out. In the first half of the green section, taller peaks were observed. These peaks could have been caused by the previous non-matching section. There was a new stimulus every three seconds in the PST, which caused physiological reactions to compound [26]. Therefore, as the test was switching from non-matching to matching, the beginning of the matching section could present artifacts from the non-matching section.

As stated previously, it was observed that the average in the matching section is lower than the average in the non-matching section. The difference in averages indicated there was a difference in the user's reaction to the stimuli or the user's stress.

The phasic driver is a more accurate indication of the user's response to the stimuli [26]. In Figure 4.3, the graph is still separated into the matching and non-

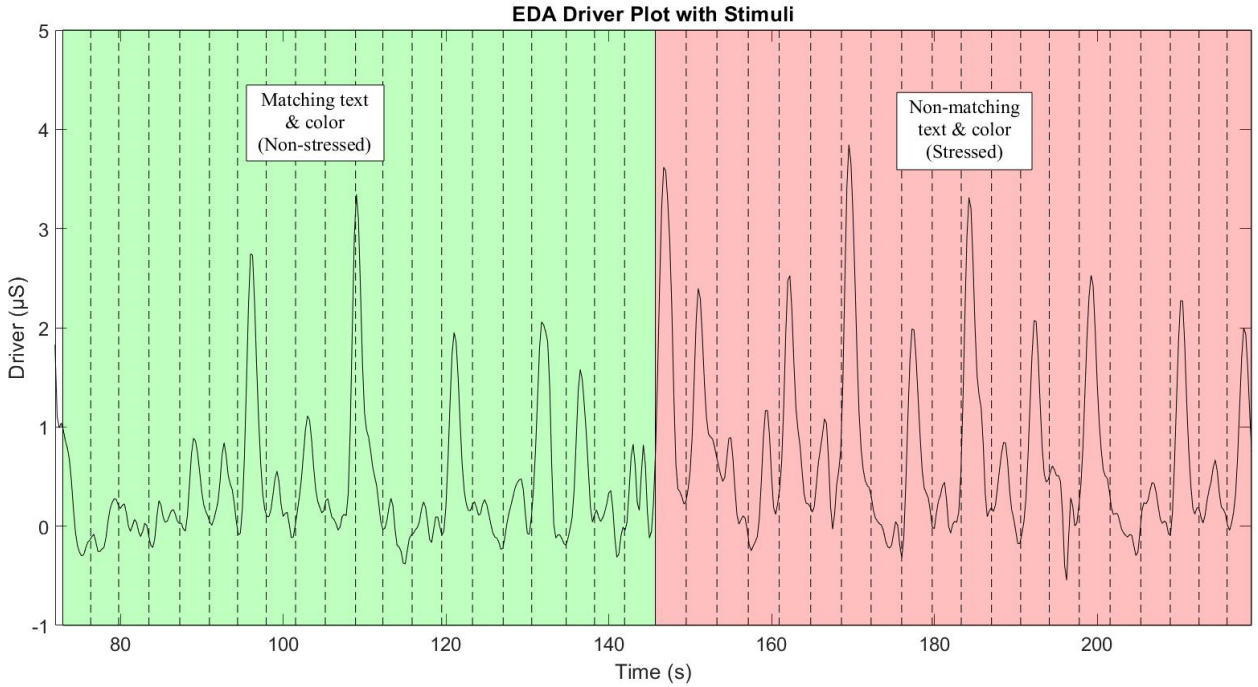

Figure 4.3 *User 1 EDA driver data for the last two sections of the PST*

matching sections, with the added vertical lines indicating each stimulus or question in

the PST. It is important to note that for almost every stimulus on the graph, there was a driver peak directly following the stimulus.

Table 4.2 shows the results of using all 20 stimuli for both the matching and non-matching sections. The overall accuracy is 60.8%. However, there were errors. The cells shaded in red should be zero because in a perfect SVM there would be no FPs or FNs. Also, the cells shaded in green should be greater than the cells in red.

User		NS	S	G-Score	F_β-score
1	NS	6	1	0.86	0.86
	S	1	6		
2	NS	4	3	0.48	0.57
	S	6	4		
3	NS	5	2	0.63	0.71
	S	4	5		
4	NS	4	3	0.50	0.57
	S	5	4		
5	NS	5	2	0.51	0.60
	S	4	3		
6	NS	5	2	0.73	0.67
	S	1	4		
7	NS	3	4	0.58	0.50
	S	2	4		

Table 4.2 *Results training on last two sections of PST and predicting on Tangram*

For example, User 1 has only one false positive and false negative, and considered accurate results within margin. On the other hand, User 2 has six FN results with only four TNs. These types of results made User 2's results less than desirable. These kinds of results, along with the compounding artifacts discussed previously, supported the decision to use the last 10 stimuli for each section of the PST. It was not obvious from the driver graph that the sections differ significantly. As a result of

physiological compounding, artifacts in the EDA data from the previous section will be present. However, if the window of focus is shifted to the last half of the stimuli in each section of the PST, the compounding artifacts from the previous section were no longer an issue, and it was easier to see the difference in the user's physiological reactions. Therefore, the analysis in this section and the following section will focus on the second half of the last two sections of the PST as seen in Figure 4.4.

4.3 Train 10 & 10 PST – Predict Tangram

The previous results, shown in Table 4.2, did not support the research hypothesis. A deeper analysis was performed on the individual matching and non-matching sections of the PST. The decisions made for this analysis are discussed below.

The data for this analysis was reduced to the last half of each section of the PST,

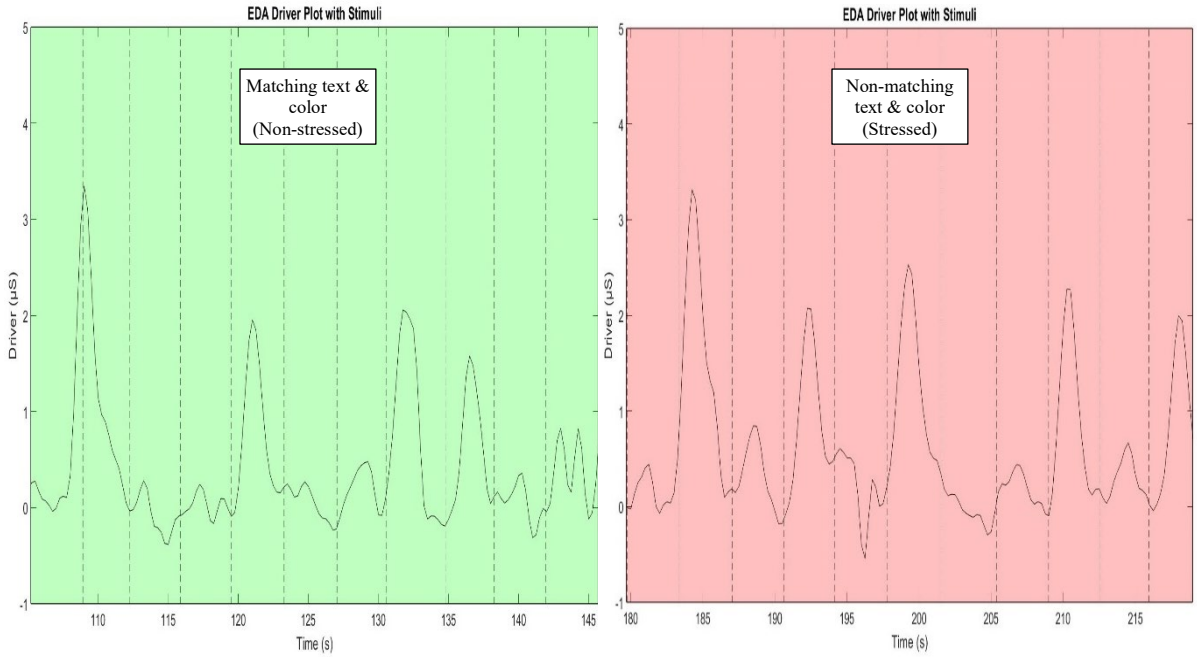

Figure 4.4 *User 1 EDA driver data for the last two sections of the PST. Using only the last 10 questions from each section.*

or the last 10 stimuli for each section. Results for the last 10 stimuli for both matching

and non-matching sections of the PST can be found in Table 4.3. The accuracy for the results when switching to the last 10 stimuli is 61.8%. The overall accuracy did increase but not consistently for all users; changing the training data did seem to alter the results negatively for some users and positively for others. Looking at User 1, even though the overall true positives and negatives are greater than the false positives and negatives, there was a negative impact on this user. Contrarily, User 2 experienced a positive impact to the overall results. Now, both the true positives and negatives are greater than the false positives and negatives. Even though the overall accuracy is better, the results for this model are still not considered acceptable.

User		NS	S	G-Score	F_β-score
1	NS	5	2	0.62	0.67
	S	3	4		
2	NS	5	2	0.74	0.78
	S	3	7		
3	NS	3	4	0.63	0.60
	S	3	6		
4	NS	6	1	0.68	0.83
	S	4	5		
5	NS	4	3	0.34	0.40
	S	5	2		
6	NS	6	1	0.52	0.66
	S	3	2		
7	NS	4	3	0.62	0.57
	S	2	4		

Table 4.3 *Results training on PST and predicting on Tangram (last half of PST sections)*

4.4 Adjusting the window size

One of the variables that was investigated for this thesis was the window size for analyzing the PST. Since the PST presented a new stimulus to the users every three seconds, the size of the window became an issue. For the research described in this thesis, Figner's window size is used; six seconds after the stimulus [24]. Therefore, the PST with stimulus every three seconds contradicts Figner's window. Initially, the PST data was analyzed using the six second window and then used to train the SVM. When the SVM then predicted on the Tangram experiment data, the results in Table 4.4 were received.

User		NS	S	G-Score	F_β-score
1	NS	4	3	0.57	0.57
	S	3	4		
2	NS	6	1	0.37	0.65
	S	8	2		
3	NS	4	3	0.50	0.57
	S	5	4		
4	NS	4	3	0.59	0.62
	S	4	5		
5	NS	4	3	0.46	0.50
	S	4	3		
6	NS	3	4	0.63	0.50
	S	1	4		
7	NS	4	3	0.72	0.63
	S	1	5		

Table 4.4 *Results using a [0 6] window for PST data*

From Table 4.4, it was determined that using the six second window was not a viable window size. Since the stimuli are presented every three seconds in the PST, as shown in Figure 4.5, a six second window would classify the user's reaction to two stimuli as a reaction to a single stimulus. Therefore, the window size needed to be

adjusted. It was discovered that in order to capture the reaction to the stimuli, considering the users' physiological state before the stimulus, and to capture any late reactions, it was best to create a window one second before the stimuli and then four seconds after each stimulus, shown as a green line in Figure 4.5. Even though this timeframe still includes the second stimulus, the window only captures reactions from the first stimulus. A user's reaction time, or onset latency, is between 1 to 3 seconds [24]. This window will allow the data analysis to consider the user's physiological state before and during a stimulus, and their physiological state just before the next stimulus. Therefore, the window defined as [-1 4] became the set window for analyzing the PST data. The results for using the [-1 4] window can be seen in Table 4.5.

User		NS	S	G-Score	F_β-score
1	NS	6	1	0.57	0.74
	S	4	3		
2	NS	5	2	0.74	0.78
	S	3	7		
3	NS	3	4	0.63	0.60
	S	3	6		
4	NS	6	1	0.68	0.83
	S	4	5		
5	NS	4	3	0.34	0.40
	S	5	2		
6	NS	6	1	0.52	0.66
	S	3	2		
7	NS	4	3	0.62	0.57
	S	2	4		

Table 4.5 *Results using a [-1 4] window for PST data*

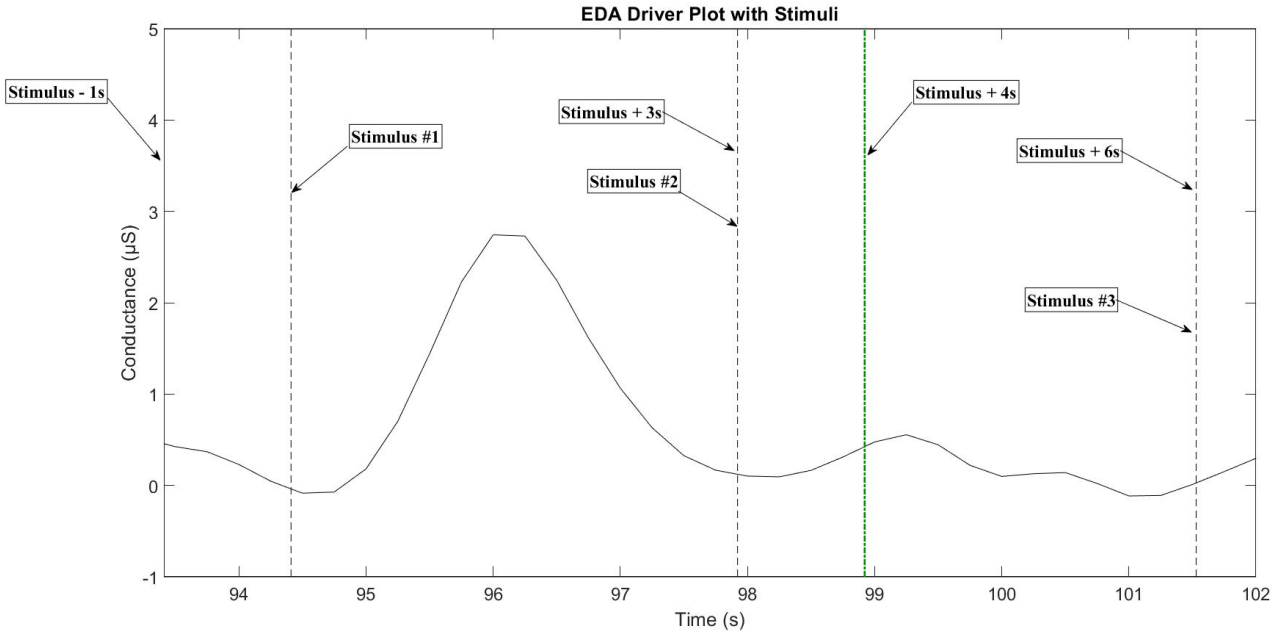

Figure 4.5 *Window size for PST stimulus analysis*

4.5 Testing the SVM: PST on PST

At this point in the research, because Dr. Saha's results were more accurate than the result found here, it was decided to test the training data. The method used to acquire the results for Table 4.5 are the same for the previous method with two exceptions. One exception was that only five points from each section of the PST were used for training the SVM. The other exception is instead of predicting on the Tangram data, the SVM predicted on the remaining five from each section of the PST. The results for this test were expected to be closer to 100% accuracy. The reason for that hypothesis was that the data used for training and predicting were from the same test, the same user, during a close period of time. As was observed from the results in Table 4.4, 100% accuracy was not the case. According to research done using a virtual reality Stroop test, they were able to use partial Stroop test data to train the SVM and predict on the remaining with 96.5%

average accuracy [27]. However, their Stroop test was not paced. The PST discussed in this thesis was paced in order to increase the user's reaction to each question. The test used in the virtual reality research attempted to increase the user's physiological response by having the users complete the Stroop test while driving a virtual reality car [27]. So the PST referred to in this thesis was paced and therefore caused issues while training the SVM.

User		NS	S	G-Score	F_β-score
1	NS	3	2	0.26	0.33
	S	4	1		
2	NS	4	1	0.91	0.83
	S	0	5		
3	NS	1	4	0.37	0.33
	S	3	2		
4	NS	2	3	0.68	0.57
	S	1	4		
5	NS	3	2	0.73	0.67
	S	1	4		
6	NS	3	2	0.45	0.50
	S	3	2		
7	NS	1	4	0.63	0.50
	S	1	4		

Table 4.6 *User-wise confusion matrix results for PST on PST*

Therefore, using the data from the PST was not considered the best data for a baseline. As an alternate, the data gathered from the White noise and Impulse baseline tests were considered for training the SVM.

4.6 Train WN & IMP – Predict Tangram

As explained in Chapter 3, two baseline tests were conducted before the PST and after the Tangram experiment. The baseline tests had the users listen to two different audio clips of white noise and a balloon pop through a set of noise-cancelling headphones. The White Noise (WN) test was then used as the users' NS baseline, while the Impulse (IMP) test data was used as the training data as the users' S state. The SVM was then used to predict the users' responses in the Tangram experiment, and the results can be seen in Table 4.6. These results showed improvement overall but still needed to be improved upon.

User		NS	S	G-Score	F_β-score
1	NS	5	2	0.62	0.67
	S	3	4		
2	NS	5	2	0.52	0.66
	S	6	4		
3	NS	4	3	0.30	0.40
	S	7	2		
4	NS	6	1	0.60	0.79
	S	5	4		
5	NS	7	0	0.65	0.99
	S	4	3		
6	NS	4	3	0.55	0.50
	S	2	3		
7	NS	7	0	0.71	0.99
	S	3	3		

Table 4.7 *Results training on White noise and Impulse, predicting on Tangram*

As discussed previously in this section, three users start to stand out as more accurately predictable users. Users 1, 5, and 7 can be observed as more accurate,

especially in Table 4.6. In the Tangram experiment there are only seven steps in which the user was correct in their construction of the puzzle and the system indicated to them that they were correct. As can be seen for Users 5 and 7, the SVM predicted seven true positives and zero false positives, which can be seen in the F_β-scores as a 99% prediction model.

One of the main reasons training with the WN and IMP data did not produce acceptable results could be the lack of data points. For the IMP data there is only one balloon pop, therefore only one stimulus. And for the WN data there are no stimuli. For training the SVM, the WN data was simply divided into multiple sections to evaluate the user's physiological response throughout the data set. So, there was a lack of data to train the SVM to predict properly. For a binary vector machine to work properly there needs to be a clear separation between data sets [28]. With only one IMP stimulus, there will not be a clear separation.

4.7 Examining the Baseline Data

During the process of deciding the White noise and Impulse as training data over the PST, a comparison of the area under the curve (AOC) was conducted. The table shown in Table 4.7 exhibits multiple comparisons between the AOC of the White noise (WN), PST matching (M), and PST non-matching (NM) data. The first two rows are the difference between the AOC of WN and the PST sections. The logic is there should be a minimum difference between the WN and M areas, and the difference between WN and M should be greater than the previous difference.

User	1	2	3	4	5	6	7
AOC WN – M	63.00	14.42	-4.10	-14.85	13.54	-43.33	0.33
AOC WN - NM	43.86	7.64	0.72	-16.37	6.45	-47.42	-23.73
ΔNM > ΔM	FALSE	FALSE	TRUE	FALSE	FALSE	FALSE	FALSE
AOC NM/M	1.59	1.48	0.46	1.04	1.56	1.06	1.48

Table 4.8 *AOC comparison between White Noise (WN), Matching (M), Non-Matching (NM)*

Along with the raw EDA data and the results from the previous confusion matrices, it was observed that users 1, 5, and 7 were better candidates for predicting a stressed and non-stressed state. That being said, Dr. Saha was able to use the data from the PST in his research to train the user-wise SVMs. Therefore, it was concluded that there was possible incorrect data gathered for this experiment. The raw PST data gathered for this thesis was analyzed in depth to verify the possibility of incorrect data. In an effort to pre-screen the data before training the SVM, multiple differences between the White noise, matching, and non-matching data were assessed. Multiple factors were observed from Table 4.6 while focused on Users 1, 5, and 7. First, the difference in AOC between WN and NM data was greater than the difference in AOC between WN and M. One more factor to notice was the ratio of matching to non-matching AOC was 1.5, or 3 to 2, for the users that were considered to be more accurate. Other ratios ranged from 0.5 to 1.0. More data must be collected before any conclusions about these ratios can be drawn.

4.8 Overview of Results

The results for each individual data analysis are outlined in Table 4.9. What can be seen from the table is that the accuracy results are all over the board. However, results improved with the focus on the last 10 questions for the two sections of the PST and adjusting the stimuli analysis window for the PST.

Table #	Accuracy (%)	Best		Worst		Average		Description
4.1	57.84	7	0	4	3	5	2	Original Results: Training SVM on entire PST and predicting on Tangram
		3	4	6	1	5	3	
4.2	60.78	6	1	3	4	5	2	Results training on last two sections of PST and predicting on Tangram
		1	6	6	3	3	4	
4.3	61.76	6	1	3	4	5	2	Results training on PST and predicting on Tangram (last half of PST sections)
		2	7	5	2	3	4	
4.4	54.90	6	1	3	4	4	3	Results using a [0 6] window for PST data
		1	5	8	2	4	4	
4.5	61.76	6	1	3	4	5	2	Results using a [-1 4] window for PST data
		2	7	5	2	3	4	
4.6	55.71	4	1	1	4	2	3	User-wise confusion matrix results for PST on PST
		0	5	4	1	2	3	
4.7	59.80	7	0	4	3	5	2	Results training on White noise and Impulse, predicting on Tangram
		2	4	7	2	4	3	

Table 4.9 *Outline of results from entirety of research*

The three columns labeled “Best,” “Worst,” and “Average” contain information for TN, FN, TP, and FP. These columns were added to Table 4.9 to present overall results for all seven users. It is important to talk about these numbers because these numbers

were used in the decision whether to accept the overall results for different training sets. For example, even though the table shows that Table 4.3 and 4.5 were more accurate, these tables did not have the top totals of TN or non-stressed responses from the users. The table presents results for FN and FP reactions highlighted in red. High FN reactions are unreliable because the user perceived that as a stressful event and it was categorized as a non-stressful event. If it were a real-world event, the wrong service would not be corrected because the CAIE calculated the event as non-stressful to the user. The same logic can be applied to the FP reactions. The user perceived the event as non-stressful, yet the CAIE categorized it as a stressful event and will try to correct the service when a correction is not necessary. Consequently, as stated throughout this chapter, the results from this research were not as expected and therefore inconclusive.

Chapter 5 Conclusions

This chapter discusses the conclusion of the experiment and the discussion of future endeavors. Section 5.1 discusses the results of the experiment, and Section 5.2 outlines the work proposed for furthering the results in the future.

5.1 Discussion of Results

In this thesis, a mock version of on-screen, step-by-step instructions was used as a CAIE to test the CAfFEINE Framework presented by Dr. Saha [4]. Using the Tangram puzzle, the user was able to follow on-screen instructions to complete the task in seven steps. What the users did not know was that the system was set up to intentionally cause errors and in turn cause technostress in the user. The users' physiological response was noticeable not only when the users were presented with a negative stimulus, but also when the users were presented with a positive stimulus. The hypothesis for this experiment was that the users' EDA graphs, and their subsequent phasic driver graphs, would show taller peaks during negative stimuli and shorter peaks during positive stimuli. Even though some user data supported that hypothesis, other user EDA graphs contradict the hypothesis.

Since the hypothesis was not supported fully, the results were inconclusive. This is why the majority of the data analysis was spent exploring multiple variables of the PST. When the PST data was found to not be a viable data set for training the SVM, the baseline WN and IMP data was used instead. However, even with WN and IMP baseline data, the SVM did not predict as accurately as hypothesized. Inconclusive results could be a product of the small number of users as there were only seven participants. Another

reason for inconclusive results could be population variability. As discussed in Chapter 3, a user's personality and state of mind, or mood, will affect the physiological data [19].

Another possible cause for inconclusive results could be in the hardware and data analysis as discussed in Chapter 4. Dr. Saha used two different devices to measure both EDA and ECG. The device used for this research was an Empatica E4, which measures both EDA and BVP data. BVP data has to be handled much differently than ECG data. Another aspect of the Empatica smartwatch is that it measures acceleration of the smartwatch. This measurement and the fact that the smartwatch could possibly move on the user's wrist can cause distortions in the data. Therefore, there could be false peaks recorded or true peaks not recorded that are caused by movement.

An additional source that could have caused issues in the final results could be the PST. As discussed in Section 4.4, Figner found that there needs to be 6 seconds between stimuli in order to get a clear physiological response [24]. Dr. Saha hypothesized that presenting a stimulus in the PST every 3 seconds would "enhance the stress-inducing capability" of the baseline test. However, the results in this thesis have shown that this may not be a good idea with the measuring device being used, the Empatica smartwatch. Additionally, the order and amount of questions in the PST could be causing issues. The first section of the PST was non-matching. As discussed in Section 4.2, the combination of a non-matching section and the user getting used to the test cause higher physiological responses than the user's response to the rest of the PST. This could be resolved if the PST was rearranged to have a matching section first that was not going to be used for analysis, but just as a section for the users to get comfortable with the interface. The table

below, Table 5.1, is a summary of possible issues discussed in this section and reflections on what caused these issues.

Issue	Type of Issue	Probable Cause
Small number of users	Experimental Design	Since Dr. Saha had 7 users, I assumed 7 would be enough for this experiment
Only one day of data collection	Psychological	Mental load varies from day to day
Measurement device data distortion	Signal Processing	Moving the arm with the watch can cause false artifacts
PST pace (3 seconds)	Psychological	Figner found that there needs to be 6s or more between stimuli [24]
PST question order	Experimental Design	Too few questions and may need to start with matching instead of non-matching

Table 5.1 *Experimental issues and probable causes*

As discussed in Chapter 4, the SVM more accurately predicted the data for Users 1, 5, and 7. Focusing on those three users in Table 4.8, it can be observed that the ratio from NM to M is a factor of 1.5. Also, User 3 is the only user where the AOC of the NM section is less than M section. These kinds of factors could be useful in the future as screening for useful and non-useful data to train an SVM. However, more users and data are needed to make a conclusion on the usability of these factors.

5.2 Discussion of Future Work

As discussed in Section 5.1, during the analysis for this thesis it was observed that not all user data appeared to be useable. Many factors could have contributed to nullification of the data. Therefore, in the future, more participants would be needed to analyze the PST data before training the SVM. It might also be prudent to retest past

participants to ensure the correct data was gathered before training the SVM for that user. However, without the difficulties encountered during the data analysis, the discovery of the need for multiple experiments on one user would not have been discovered.

Also, the hardware being used to gather physiological data must be discussed in regards to future work. The Empatica smartwatch was used to measure the users' physiological response. However, the Empatica smartwatch can frequently introduce errors in the data because of the movement of the users' wrist. Another issue with the Empatica smartwatch is the temperature measurement. When analyzing the temperature data gathered from the Empatica smartwatch, it was observed that the temperature change was minimal (0.1 to 0.5 Δ°C) [17]. Therefore, features that could have been extracted from body temperature were unable to be used in the classification process for this research. With issues, such as those listed above, on-body measurement leads to multiple errors and misclassification of the data. As such, non-wearable measurement devices should be considered for future research.

Non-wearable measurement device research into affect-interpretation has already been started by Khan, Ingleby, and Ward [29]. In this research, infrared thermal cameras were used to measure temperature changes on participants' faces [29]. Unlike the Empatica smartwatch, which only measures temperature changes within a tenth of a degree, thermal cameras can measure temperature changes more precisely. Also, instead of focusing only on specific places on the wrist, the thermal camera captures the whole face of a participant. Khan et al. also discuss Facial Thermal Feature Points (FTFPs), which were used to identify facial expression for emotional classification [29]. The use of this method and other thermal-imaging methods could be used to replace the on-body

measurement methods used in the research described in this thesis. It could be hypothesized that the data gathered by non-wearable measurement devices would be more reliable and repeatable.

Overall this research found that the CAfFEINE Framework might be useable for some users in real-world applications using a smartwatch. The Framework could easily be deployed on smartphones since the majority of the population is already wearing smartwatches that monitor heart rate. Of course, this would mean that the watches would need to be upgraded to measure EDA, but this is a discussion of future work. As research has shown, an EDA response to a stimulus does not present itself for 1 to 3 seconds [25]. Then the IE would have to classify that response. Therefore, an application would need to allow for that classification to precipitate. As you can see in Table 5.2, the first column shows time, in seconds, that a person would need a response from an IE using the CAfFEINE Framework in order for the response to be useful or safe. One example of a real-world application, that also accesses information from a smartwatch, is voice-activated assistants. As discussed in Chapter 1, currently Siri provides assistance on smartphones and if there is a wrong service provided, the user has to explicitly correct Siri. For a voice-activated assistant to use the CAfFEINE Framework it would have to wait 1 to 3 seconds for the user's physiological response, then classify the response. This is a feasible situation because most applications people would use Siri for are not life threatening or time sensitive.

Response Time Window (s)	Example Application	Framework Usable?
< 1	**Driving Assistance**	No
1 - 3	**Voice-Activated Assistants**	Yes
> 3	**Long-term calculations**	No

Table 5.2 *Possible Applications and Response Time*

However, this type of Framework would not work in all real-world applications. For example, any kind of driving assistance applications would call for quick reactions and decisions to be made. The CAfFEINE Framework, when perfected, may not process fast enough for driving situations. There are also applications that are not time dependent or there is so much time between intelligent services provided that the use of the Framework would just confuse the user. For instance, a user starts a process in an IE then walks away and starts a task outside of the IE. If the CAfFEINE Framework was to provide feedback after the user walks away, the feedback would no longer be useful or relevant to the user. However, all of these applications are based on the successful use of the CAfFEINE Framework, which the research in this thesis does not support. The Framework is programed to handle only one stimulus at a time and decide whether the user's response to that stimulus is stressed or non-stressed. The training of the SVM in the Framework is to help the SVM learn whether the physiological response from the user is stressed or non-stressed for future classification of responses. The CAfFEINE Framework is not programmed to learn if the intelligent service it provided was correct or incorrect to alter the future intelligent services it provides.

In previous work, i.e. Dr. Saha's research, the environment was controlled and different measuring devices were used. The experiment discussed in this thesis moved the research into a more natural environment, but in turn allowed for more error to be

introduced. In order for a smartwatch to be used with the Framework, the system would have to be trained multiple times before it would be able to recognize and accurately classify a user's perception of the services provided.

References

[1] R. W. Picard, *Affective Computing,* Cambridge, MA: M.I.T. Media Laboratory, 1995.

[2] J. C. Augusto, V. Callaghan, D. Cook, A. Kameas and I. Satoh, "Intelligent Environments: A manifesto," *Human-centric Computing and Information Sciences,* vol. 3, no. 1, p. 12, 2013.

[3] S. A. Shafer, "Interaction Issues in Context-Aware Intelligent Environments," *Human-Computer Interaction,* vol. 16, pp. 363-378, 2001.

[4] D. P. Saha, *A Study of Methods in Computational Psychophysiology for Incorporating Implicit Affective Feedback in Intelligent Environments,* Blacksburg, VA: Virginia Polytechnic Institute and State University, 2018.

[5] J. Slocum, J. Botermans, D. Gebhardt, M. Ma, X. Ma, H. Raizer, D. Sonneveld and C. van Splunteren, The Tangram Book, Sterling Publishing, 2004.

[6] J. J. Braithwaite, D. G. Watson, R. Jones and M. Rowe, "A Guide for Analysing Electrodermal Activity (EDA) & Skin Conductance Responses (SCRs) for Psychological Experiments," *Psychophysiology,* vol. 49, pp. 1017-1034, 2015.

[7] J. Kim and E. Andre, "Emotion Recognition Based on Physiological Changes in Music Listening," *IEEE Transactions on Pattern Analysis and Machine Intelligence,* vol. 30, no. 12, pp. 2067-2083, 2008.

[8] F. Davis, "Perceived Usefulness, Perceived Ease of Use, and User Acceptance of Information Technology," *MIS Quarterly,* vol. 13, no. 3, pp. 319-340, 1989.

[9] D. Novak, A. Nagle and R. Riener, "Linking Recognition Accuracy and User Experience in an Affective Feedback Loop," *IEEE Transactions on Affective Computing,* vol. 5, no. 2, pp. 168-172, 2014.

[10] C. Liu, P. Agrawal, N. Sarkar and S. Chen, "Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback," *Int. J. Human-Comput. Interact.,* vol. 25, pp. 506-529, 2009.

[11] C. Brod, Technostress: The Human Cost of the Computer Revolution, Reading, MA: Addison-Wesley, 1984.

[12] L. Atanasoff and M. A. Venable, "Technostress: Implications for Adults in the Workforce," *Career Development Quarterly,* vol. 65, no. 4, p. 326, 2017.

[13] B. Arnetz and C. Wiholm, "Technological stress: psychophysiological symptoms in modern offices," *Journal of Psychosomatic Research,* vol. 43, pp. 35-42, 1997.

[14] R. Riedl, "On the biology of technostress: literature review and research agenda," *ACM SIGMIS Database: the DATABASE for Advances in Information Systems,* vol. 44, pp. 18-55, 2013.

[15] P. Renaud and J. Blondin, "The stress of Stroop performance: physiological and emotional responses to color-word interference, task pacing, and paced speed," *International Journal of Psychophysiology,* vol. 27, no. 2, pp. 87-97, 1997.

[16] J. R. Stroop, "Studies of Interference in Serial Verbal Reactions," *Journal of Experimental Psychology,* vol. 18, no. 6, pp. 643-662, 1935.

[17] Empatica Technical Staff, *E4 wristband from empatica: user's manual,* Empatica, 2018.

[18] M. Soleymani, F. Villaro-Dixon, T. Pun and G. Chanel, "Toolbox for Emotional feAture extraction from Physiological signals (TEAP)," *Frontiers in ICT,* vol. 4, 2017.

[19] D. Sharma and T. K. Gill, "Technostress and Personality Traits - Are they Associated? - Evidence from Indian Bankers," *International Journal of Computer Science and Technology,* vol. 7, no. 1, pp. 106-111, 2016.

[20] S. Visa, B. Ramsay, A. Ralescu and E. van der Knaap, "Confusion Matrix-based Feature Selection," *In Proc. 22nd Midwest Artificial Intelligence and Cognitive Science Conference,* pp. 120-127, 2011.

[21] E. B. Fowlkes and C. L. Mallows, "A Method for Comparing Two Hierarchical Clusterings," *Journal of the American Statistical Association,* vol. 78, no. 383, pp. 553-569, 1983.

[22] Y. Sasaki, "The truth of the F-measure," *Tech Tutor mater,* pp. 1-5, 2007.

[23] J. Jaimovich, *Emotion Recognition from Physiological Indicators for Musical Applications,* Belfast, Northern Ireland: Queen's University Belfast, 2013.

[24] B. Figner and R. O. Murphy, *Using skin conductance in judgement and decision making research,* M. Schulte-Mecklenbeck, A. Kuehberger and R. Ranyard, Eds., New York: Psychology Press, 2011, pp. 163-184.

[25] M. E. Dawson, A. M. Schell and D. L. Filion, "The electrodermal system," in *Handbook of Psychophysiology*, 3rd ed., J. T. Cacioppo, L. G. Tassinary and G. G. Berntson, Eds., Cambridge, Cambridge University Press, 2007, pp. 200-223.

[26] M. Benedek and C. Kaernbach, "A continuous measure of phasic electrodermal activity," *Journal of Neuroscience Methods,* vol. 190, no. 1, pp. 80-91, 2010.

[27] D. Wu, C. G. Courtney, B. J. Lance, S. S. Narayanan, M. E. Dawson, K. S. Oie and T. D. Parsons, "Optimal Arousal Identification and Classification for Affective Computing Using Physiological Signals: Virtual Reality Stroop Task," *IEEE Transactions on Affective Computing,* vol. 1, no. 2, pp. 109-118, 2010.

[28] J. Nalepa and M. Kawulok, "Select training sets for support vector machines: a review," *Artificial Intelligence Review,* vol. 52, no. 2, pp. 857-900, 2018.

[29] M. Khan, M. Ingleby and R. Ward, "Automated Facial Expression Classification and Affect Interpretation Using Infrared Measurement of Facial Skin Temperature Variations," *ACM Transactions on Autonomous and Adaptive Systems,* vol. 1, no. 1, pp. 91-113, 2006.