CONFIDENCE INFERENCE IN DEFENSIVE CYBER OPERATOR DECISION MAKING

THESIS

Graig S. Ganitano, Captain, USAF

AFIT-ENG-MS-19-M-028

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

DISTRIBUTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

The views expressed in this thesis are those of the author and do not reflect the official policy or position of the United States Air Force, Department of Defense, or the United States Government. This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

AFIT-ENG-MS-19-M-028

CONFIDENCE INFERENCE IN DEFENSIVE CYBER OPERATOR DECISION MAKING

THESIS

Presented to the Faculty

Department of Electrical and Computer Engineering

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Electrical Engineering

Graig S. Ganitano, BS

Captain, USAF

March 2019

DISTRIBUTION STATEMENT A.
APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

AFIT-ENG-MS-19-M-028

CONFIDENCE INFERENCE IN DEFENSIVE CYBER OPERATOR DECISION MAKING

Graig S. Ganitano, BS

Captain, USAF

Committee Membership:

Dr. Brett J. Borghetti
Chair

Dr. Gregory J. Funke
Member

Dr. Laurence D. Merkle
Member

AFIT-ENG-MS-19-M-028

Abstract

Cyber defense analysts face the challenge of validating machine generated alerts regarding network-based security threats. Operations tempo and systematic manpower issues have increased the importance of these individual analyst decisions, since they typically are not reviewed or changed. Analysts may not always be confident in their decisions. If confidence can be accurately assessed, then analyst decisions made under low-confidence can be independently reviewed and analysts can be offered decision assistance or additional training. This work investigates the utility of using neurophysiological and behavioral correlates of decision confidence to train machine learning models to infer confidence in analyst decisions. Electroencephalography (EEG) and behavioral data was collected from eight participants in a two-task human-subject experiment and used to fit several popular classifiers. Results suggest that for simple decisions, it is possible to classify analyst decision confidence using EEG signals. However, more work is required to evaluate the utility of EEG signals for classification of decision confidence in complex decisions.

Acknowledgments

I would like to offer my deepest appreciation to my faculty advisor, Dr. Brett Borghetti, for his guidance, encouragement, and patience. I would also like to thank my committee members, Dr. Gregory Funke and Dr. Laurence Merkle, for their support.

Graig S. Ganitano

Table of Contents

Page

Abstract iv
Acknowledgments v
List of Figures ix
List of Tables xii
I. Introduction 1
1.1 Problem Statement 1
1.2 Research Questions and Hypothesis 2
1.3 Methodology 3
1.4 Assumptions 4
1.5 Limitations 5
1.6 Contributions 6
1.7 Structure of the Document 7
II. Literature Review 8
2.1 Chapter Overview 8
2.2 Current Research 9
2.2.1 Drift Diffusion Model 9
2.2.2 Neural Indicators of Decision Confidence 11
2.2.3 Behavioral Indicators of Decision Confidence 19
2.2.4 Inference of Decision Confidence Through Machine Learning 20
2.3 Research Gaps 28
2.3.1 Experimental Designs 28
2.3.2 Inferring Decision Confidence 28

2.4 Related EEG Research 29
2.4.1 Random Forests 29
2.4.2 Artificial Neural Networks 31
2.5 Summary 36
III. Methodology 38
3.1 Chapter Overview 38
3.2 Background 38
3.3 Experiment 39
3.3.1 Participants 40
3.3.2 Random Dot Kinematogram Task 41
3.3.3 Cyber Intruder Alert Testbed Task 45
3.4 Electrophysiological Data Acquisition 50
3.5 EEG Pre-Processing 53
3.6 Analysis Strategy 54
3.6.1 Event Related Potential Analysis 54
3.6.2 Machine Learning 56
3.6.3 Behavioral Analysis 67
3.7 Summary 68
IV. Analysis and Results 70
4.1 Chapter Overview 70
4.2 Random Dot Kinematogram Task 71
4.2.1 Event Related Potentials 71
4.2.2 Classification of Confidence 72

4.3 Cyber Intruder Alert Testbed Experiment Analysis 84

4.3.1 Behavior Results and Analysis 84

4.3.2 Event Related Potential Analysis 96

4.3.3 Classification of Confidence 97

4.4 Summary 108

V. Conclusions and Recommendations 110

5.1 Conclusions of Research 110

5.2 Significance of Research 112

5.3 Recommendations for Future Research 113

5.3.1 CIAT Data Segmentation 113

5.3.2 Machine Learning Improvements 114

5.3.3 ECG and EOG Analysis 116

5.3.4 Experimental Design Changes 116

Appendix A: Pre-Experiment Questionnaire 118

Appendix B: Post-Experiment Questionnaire 119

Appendix C: General Cyber Alert Investigation Workflow Handout 120

Bibliography 121

List of Figures

Page

Figure 2.1: Fully-Connected Neural Network 32

Figure 2.2: Simple Recurrent Neural Network 34

Figure 3.1: Experiment Sequence 40

Figure 3.2: Example Random Dot Kinematogram 42

Figure 3.3: Random Dot Kinematogram Task Sequence 43

Figure 3.4: CIAT Interface 45

Figure 3.5: Example Decision Prompt 47

Figure 3.6: Data Acquisition Setup 51

Figure 3.7: International 10-20 System 51

Figure 3.8: RDK task EEG data timestamped (sec) to show stimulus presentation (7680) and participant response (12800). 52

Figure 3.9: EOG electrode placement 53

Figure 3.10: ECG electrode placement 53

Figure 3.11: Visually Rejected Epoch 58

Figure 3.12: Recursive Feature Elimination with Cross-Validation 60

Figure 3.13: Example Fully connected Neural Network Architecture 61

Figure 3.14: Example Convolutional-Recurrent Architecture 62

Figure 3.15: Confusion Matrix 63

Figure 3.16: Receiver Operating Characteristic Curve 66

Figure 3.17: Theoretical Model for Participant Behaviors 67

Figure 4.1: Example ERPs for Participant 4524 (Left) and Participant 7984 (Right) 72

Figure 4.2: BACC for the Best Performing Models on the RDK Task..... 74

Figure 4.3: AUC for the Best Performing Models on the RDK Task 75

Figure 4.4: MCC for the Best Performing Models on the RDK Task..... 75

Figure 4.5: Cohen's Kappa for the Best Performing Models on the RDK Task..... 75

Figure 4.6: Confusion Matrices for the Best and Worst Performing Models 77

Figure 4.7: BACC for the CRNN fit on the RDK Task Data..... 82

Figure 4.8: AUC for the CRNN fit on the RDK Task Data 82

Figure 4.9: MCC for the CRNN fit on the RDK Task Data..... 83

Figure 4.10: Cohen's Kappa for the CRNN fit on the RDK Task Data..... 83

Figure 4.11: Number of Confident Observations versus Query Number..... 87

Figure 4.12: Number of Confident Observations versus Difficulty..... 87

Figure 4.13: Distribution of Tool Transitions for Confident and Unconfident Responses 88

Figure 4.14 : Reaction Time versus Query Number 90

Figure 4.15 : Reaction Time versus Difficulty..... 90

Figure 4.16 : Distribution of Reaction Times for Confident and Unconfident Responses 90

Figure 4.17 : Reaction Time versus Tool Transit ions 91

Figure 4.18: Tool Transitions versus Query Number..... 92

Figure 4.19 : Tool Transitions Versus Difficulty 93

Figure 4.20: Tool Transitions versus Confidence 93

Figure 4.21: Number of Correct Observations versus Query Number..... 95

Figure 4.22 : Number of Correct Observations versus Difficulty..... 95

Figure 4.23: Distribution of Tool Transitions for Correct and Incorrect Responses..... 95

Figure 4.24: BACC for the Best Performing Models on the CIAT Task..... 98

Figure 4.25: AUC for the Best Performing Models on the CIAT Task 99

Figure 4.26: MCC for the Best Performing Models on the CIAT Task..... 99

Figure 4.27: Cohen's Kappa for the Best Performing Models on the CIAT Task..... 100

Figure 4.28: Comparison of BACC When Controlling for Query 104

Figure 4.29: Comparison of AUC When Controlling for Query 104

Figure 4.30: Comparison of MCC When Controlling for Query 105

Figure 4.31: Comparison of Cohen's Kappa When Controlling for Query 105

Figure 4.32: BACC for the CRNN fit on the CIAT Task Data..... 107

Figure 4.33: AUC for the CRNN fit on the RDK Task Data 107

Figure 4.34: MCC for the CRNN fit on the CIAT Task Data..... 108

Figure 4.35: Cohen's Kappa for the CRNN fit on the CIAT Task Data..... 108

List of Tables

Page

Table 3.1: Control Variables 44

Table 3.2: CIAT Tools and Descriptions 46

Table 3.3: Response Variables for the CIAT Experiment..... 48

Table 3.4: Test Matrix 49

Table 3.5: Independent Variables for the CIAT Experiment 49

Table 3.6: Class Distribution of Observations 56

Table 3.7: Set of Testable Hypothesis..... 68

Table 4.1: Electrodes and Latencies of Observed Differences in ERPs 72

Table 4.2: Mean Performance of Frequency Band Models for the RDK Task..... 73

Table 4.3: Comparison of RDK Single Band Models to Leave one-band out Models with Respect to Highest Perfromance and Highest Perfromance Drop..... 79

Table 4.4: Intersection of Salient Features Across LR, LDA, and RF Models for the RDK Task..... 80

Table 4.5: Mean Performance of the CRNN Models for the RDK Task 81

Table 4.6: Descriptive Statistics for the CIAT Behavioral Data 85

Table 4.7: Mixed Effects Logistic Regression Model for Confidence..... 89

Table 4.8 : Linear Mixed Model for Reaction Time 91

Table 4.9 : Linear Mixed Model for Tool Transitions 94

Table 4.10: Mixed Effects Logistic Regression Model for Correctness 96

Table 4.11: Mean Performance of Frequency Band Models for the CIAT Task 97

Table 4.12: Comparison of CIAT Single Band Models to Leave one-band out Models with Respect to Highest Perfromance and Highest Perfromance Drop 101

Table 4.13: Salient Features Across LR, LDA, and RF Models for the CIAT Task 101

Table 4.14: Mean Performance of the CRNN Models for the CIAT Task 106

CONFIDENCE INFERENCE IN DEFENSIVE CYBER OPERATOR DECISION MAKING

I. Introduction

Humans and computers each have inherent strengths and weaknesses. Computers can outperform humans in tasks such as sorting and searching through large amounts data and performing computations quickly and correctly, but struggle with the uncertainty and ambiguity of decision-making as well as adapting to new or unexpected situations. Humans on the other hand, excel in situations that require understanding context and are able to adapt to new situations with greater success. Because the combined strengths of humans and computers complement their individual weaknesses, researchers have devoted their attention to the concept of human-machine teaming.

A key component of human-machine teaming is the ability of a machine to assess a human operator's ability to carry out their job at a particular moment in time, known as Operator Functional State Assessment (OFSA) [1]. If a machine can assess and understand an operator's state, it can make better decisions and ultimately drive human-machine team performance towards an optimal level. The focus of this research is on a subcategory of OFSA - inferring operator decision confidence, particularly in the realm of cyber defense.

1.1 Problem Statement

Effective cyber defense currently relies heavily upon human operators, colloquially know as cyber defense operators. One critical role played by cyber defense operators is the network analyst. These operators work collaboratively with computer algorithms to identify and respond to malicious activity and policy violations. However, the alerts

generated by these algorithms do not always correspond to an actual cyber threat [2], and so operators face the challenge of determining the validity of these alerts as part of their regular operations. Once an operator has committed to a decision about the validity of an alert, due to operations tempo and manning there is usually no manpower remaining for quality assurance activities, meaning an incorrect decision could have catastrophic consequences for the corresponding network and host systems. Since decisions have an associated level of confidence, if confidence could be accurately inferred then it could potentially be used to identify operators in situations of low confidence. Assistance could then be provided in the form of investigation review, additional monitoring by more experienced cyber operators, and tailored training experiences based on observed decision confidence patterns from previous investigations.

1.2 Research Questions and Hypothesis

This study attempts to fill the current research gap of using neural and behavioral correlates of decision confidence in combination with machine learning techniques to infer confidence in a simple decision and extend the results to more complex decisions with emphasis on the types of decisions made by cyber defense operators. The following research questions focus on these goals:

RQ1. Can electrophysiological features be used in combination with machine learning techniques to infer decision confidence in a simple decision with a performance greater than chance?

Hypothesis: Machine learning models will be able to learn the neural correlates of decision confidence and thus be able to infer decision confidence in a simple decision with a performance greater chance.

RQ2. What are the salient electrophysiological features for inferring decision confidence in a simple decision?

Hypothesis: Changes in power in the five traditional EEG frequency bands (alpha in particular) will be prominent features for inferring decision confidence.

RQ3. Can behavioral features be used in combination with machine learning techniques to infer decision confidence in a complex decision with a performance greater than chance?

Hypothesis: Machine learning models will be able to learn correlations between decision confidence, reaction time, and information seeking and thus be able to infer decision confidence in a complex decision with a performance greater than chance.

RQ4. Are the salient electrophysiological features for inferring decision confidence the same for both simple and complex decisions?

Hypothesis: Features identified as salient for a simple decision will still encode important information that can be used to infer decision confidence for complex decisions.

1.3 Methodology

A two-task human-subject experiment was designed in which electrophysiological and behavioral data was recorded for eight participants. The first task used in this experiment aimed at measuring electrophysiological data associated with confident and

unconfident simple decisions in a motion discrimination task. For this task, participants were presented with a series of random dot kinematogram (RDK) stimuli [3] and had to decide whether the global direction of dot motion for each stimulus was to the left or to the right. The next task aimed at measuring electrophysiological and behavioral data associated with the types of complex decisions made by cyber defense operators in their operational environment. The investigation was conducted using a modified version of the Cyber Intruder Alert Testbed, a synthetic task environment that simulates typical network intrusion detection tasks [2]. For this task, participants investigated 30 cyber-alerts of varying difficulty with the goal of determining the validity of each alert based on information available from various tools. For each task, the electrophysiological data were transformed into both time and frequency domain features and used to fit machine learning models of varying levels of flexibility for evaluation and comparison of both model performance and feature salience. Behavioral data from the cyber investigation was explored in order to identify patterns of behavior suitable as features for decision confidence inference as well as to provide insight into misclassifications made by models fit using the electrophysiological data.

1.4 Assumptions

In order to answer the proposed research questions, the following assumptions about the experiment design were made:

- Participants have no knowledge of the RDK stimulus order or alert content and have not been informed by a past participant prior to participating in the experiment.

- Participants are willing to assess each alert based on evidence accumulated during their investigation.
- Participants will seek to maximize their score by selecting the “I Don’t Know” option instead of guessing for RDK stimuli and alerts in which they do not know the answer.
- Brain activity associated with confident and unconfident decisions can be detected using electrophysiological measurements.

1.5 Limitations

Participants for this study were recruited solely from the Air Force Institute of Technology. Eight volunteer participants (all male) between the ages of 21-31 with a mean age of 24.7 and a standard deviation of 3.60 were recruited. All participants were United States active duty military personnel and held at least a bachelor’s degree in engineering and computer science fields. Due to the sampling bias introduced by the recruiting process, it is possible the results of this study will not generalize to a more diverse population.

Each experiment session had a strict two-hour time limit. To complete the experiment within this constraint, the experimental design only allowed for collection of 440 observations for the first task and 120 observations for the second task, per participant. Several issues arise when dealing with small datasets such as these. It may not be possible to split the dataset into training, validation, and test sets such that all sets follow the same probability distribution. In such cases, the validation set may not optimally guide the parameter search and the test set may not give a meaningful estimate of model generalization. Small datasets also increase the risk of overfitting. Reducing

model complexity can reduce overfitting, however this also limits the types of classifiers that can be utilized.

It was confirmed that during experimentation, the electrophysiological data acquisition system was periodically malfunctioning. The extent to which the malfunctioning equipment impacted the integrity of the data is unknown. Due to schedule constraints and participant availability, data collection on a replacement system did not take place.

1.6 Contributions

This work builds upon the research on decision-making and decision confidence by investigating the feasibility of machine learning models trained on behavioral and electrophysiological features as a means for inferring decision confidence for both simple and complex decisions. Specifically, this research was the first to explore decision confidence inference in a RDK motion discrimination task using both linear and non-linear machine learning models trained on electroencephalography (EEG) data. This work also represents the first to attempt decision confidence inference in complex decisions using the same techniques. For machine learning models fit using data from the RDK task, the best performing model for each participant exceeded classification performance greater than random chance with respect to four performance metrics. Additionally, frequency domain information thought to discriminate between levels of confidence were identified as important features in over half the participants. Performance of models fit using electrophysiological data from the cyber investigation task appeared to exceed random chance. However, after controlling for unintended effects of the experimental

design, the use of EEG was observed to provide little utility towards decision confidence inference. This observation highlights the importance of adhering to a set of standards when conducting a performance evaluation of machine learning models, as sole reliance on standard performance metrics can lead to inflated results.

1.7 Structure of the Document

The remainder of this document is structured into four chapters. Chapter 2 provides a thorough review of present literature focusing on neural and behavioral representations of decision confidence and their salience for inferring confidence in future decisions using machine learning. Since little research has been conducted using machine learning for decision confidence inference, this is followed by a review machine learning approaches for inferring other types of cognitive activity. Chapter 3 describes the details of the two-task human-subject experiment design as well as the techniques used to analyze the behavioral and physiological data collected during the experiment. Chapter 4 presents the results of the analysis of the behavioral and physiological data. Finally, Chapter 5 summarizes the significant findings of this research and discusses areas for future work.

II. Literature Review

2.1 Chapter Overview

Human beings possess the innate ability to subjectively evaluate their performance on a perceptual task. Even without being explicitly told the correct answer, they can identify the possibility of having made an error and are able to express a level of confidence in their decision. Over the past few decades researchers have invested a substantial amount of effort into investigating the neural and behavioral basis underlying the decision-making process [4]. However, very little research has been done regarding the neural representation of decision confidence. Experimentation has focused solely on simple two-choice decisions that are made in a matter of seconds, which raises the question of whether the results of such experimentation extend to more complex decisions such as those made by cyber defense operators during a cyber-investigation.

In the following sections, we review literature that has investigated the neural and behavioral representations of subjective confidence in cognitive tasks. Subsequently, we highlight literature that links these representations to the problem of inferring confidence in future decisions utilizing machine learning techniques. Lastly, we identify gaps in the current body of research and potential avenues for filling these gaps based on results in research on inferring cognitive activities other than decision confidence.

2.2 Current Research

2.2.1 Drift Diffusion Model

The primary model of decision-making upon which experiment designs investigating neural and behavioral representations of decision confidence are based is known as the drift-diffusion model (DDM). The DDM models the decision process for decisions that meet the following assumptions [5]:

1) The decision involves two choices.
2) The decision requires a single-stage decision process (as opposed to multi-stage processes that may be involved in reasoning tasks).
3) The decision is made quickly (mean reaction time of only a few seconds).

In the DDM, each of the two available choices has a corresponding response boundary. The DDM models decision-making as a noisy process where at each time step, evidence is accumulated for one of the two choices until a response boundary is reached at which point the decision is made in favor of the corresponding choice. More specifically, the decision variable v_t is updated according to

$$v_t = v_{t-1} + \delta + cW$$

where δ is a linear drift term that encodes the rate of evidence accumulation, and cW is Gaussian noise with mean zero and variance c^2. A decision is made when

$$-\theta > v > \theta$$

where θ is a fixed deviation from zero. Decision confidence is then thought to scale with the product of δ and θ [6]. Confidence reporting in experiments examined in this work take one of two forms: participants reporting confidence that they made the correct choice

or reporting confidence that they made an error [6]. The latter form, known as error monitoring, also fits within the DDM framework. In this case errors are detected as a re-crossing of a single response boundary or as a successive crossing of both response boundaries [7].

Current research on neural and behavioral representations of decision confidence has almost exclusively focused on simple two-choice decision tasks that can be modelled using the DDM. The following are examples of such tasks:

- Random Dot Kinematogram (RDK) Task: Participants are shown a stimulus in the form of dots in an aperture, where a percentage of dots move together in the same direction and the remaining dots move randomly. Participants must make a choice between the two possible directions of coherent dot motion.
- Grating Orientation Task: Participants are shown a stimulus in the form of a grooved surface, with gratings aligned vertically or horizontally. Participants must make a choice between the two grating orientations.
- Image Discrimination Task: Participants are shown a stimulus in the form of an image belonging to one of two categories. Participants must choose which category the image belongs two.

In contrast to these simple two-choice decision tasks, complex decisions such as those made by cyber defense operators during a cyber-investigation, often violate the assumptions listed above and represent a gap in current research.

2.2.2 Neural Indicators of Decision Confidence

A large portion of research into understanding the neuronal backings for the decision-making process and the representation of decision confidence within this process utilizes Electroencephalography (EEG). EEG is a measurement technique in which electric brain potentials, resulting from electrochemical signals being passed between neurons, are noninvasively measured via electrodes placed on the head [8]. When large populations of neurons are synchronously active, the small electric fields generated by each individual neuron sum together resulting in a field strong enough to propagate through several anatomical layers including the brain tissue, skull and skin [9]. Within the reviewed literature, two techniques dominate the analysis of EEG data for investigation of neural representations of subjective confidence: event related potential (ERP) analysis and time-frequency analysis.

2.2.2.1 Event Related Potentials

ERPs are very small positive or negative voltage deflections that appear in response to an applied stimulus. When EEG data is segmented and time-locked to the stimulus event (known as epoching the data,) if the epochs are aligned and averaged at each time point, these deflections become clear. Because the noise fluctuations are randomly distributed around zero, by taking the average across all epochs, the combination of the individual noise contaminating each signal tends to cancel out, approaching zero as the number of epochs increases [9]. The waveform that results from this process is the ERP, which can be further divided into distinct components that reflect deviations from a pre-event baseline. The peak amplitudes and latencies of these ERP

components are thought to index discrete sensory and cognitive processes that unfold over time in response to a class of events [10]. ERP components that have been shown to discriminate between different levels of confidence include P300 (P3), error-related negativity (ERN), and error-related positivity (Pe). However, since the process to extract the ERP components collapses all trials into a single observation, ERPs are generally not a suitable feature for the classification of decision confidence using machine learning discussed in 2.2.4.

2.2.2.1.1 P300

The P300 or P3 event-related component is a positive deflection that occurs when a subject detects an informative task-relevant stimulus, with a typical peak latency of 300 ms and is thought to represent the transfer of information to consciousness [11]. Kerkhof investigated the P3 ERP component manifesting from decisions on signal and non-signal presentations in a threshold detection task [12]. EEG data was collected from six participants who were asked to determine whether or not an auditory stimulus in the form of 3 seconds of wide-band white noise contained a 100 ms duration of a 1000 Hz sinusoidal signal. Participants responded after the presentation of the stimulus by pressing one of eight buttons, each indicating a level of confidence ranging from high confidence that the sinusoid was present to high confidence that it was not. Several multivariate analyses were conducted on the preprocessed EEG data. The results of these analyses indicate that the level of decision confidence is positively correlated with the quality of the associated P3s and negatively correlated with the length and the variability of the associated P3 latencies [12].

2.2.2.1.2 Error-Related Negativity

The ERN is a negative ERP component over the frontocentral region, peaking between 60-120ms following an incorrect response [7]. Selimbeyoglu et al. examined the ERN as a neural correlate of subjective confidence levels with emphasis on subjective uncertainty in an attempt to reveal the differences in processing of perception and response level errors and to discriminate between different confidence levels [13]. EEG data was collected from seventeen participants during a circle discrimination task designed to create difficulty during the stimulus processing and response selection of the decision-making process. Each participant was shown two circles of similar sizes and had to report the larger of the two. If the participant was unsure that their answer was correct or certain that they gave an incorrect answer, they reported their confidence level. The EEG data was partitioned according to the three confidence levels corresponding to participants being certain of giving a correct answer, being certain of giving an incorrect answer, and being uncertain. ERN was quantified as the mean value between 0 and 100 ms after the response. Statistical tests were carried out through a repeated measures-ANOVA. ERN amplitude was found to be statistically different between the three confidence levels, with amplitude being most negative when a participant was certain they made an error and least negative when they were certain their given response was correct [13].

2.2.2.1.3 Error-Related Positivity

The Error-related positivity or Pe event-related component is a positive deflection occurring 200–400 ms after giving an incorrect response and reflects a representation of conscious error awareness in that the amplitude of the waveform is modulated by the degree of awareness that an incorrect response was given [7]. Boldt and Yeung

investigated whether Pe varied in a graded way with subjective ratings of decision confidence given by participants in a dot count perceptual decision task [14]. In each trial, participants were shown two arrays of dots for 160 ms and asked to determine which array contained more dots. Participants had 1520 ms post stimulus to report their answer and were also asked to report their confidence level using a qualitative 6 level scale. EEG analysis focused on the 600 ms interval between which the participant gave their answer and the subsequent appearance of the confidence scale. Pe amplitude was taken as the difference between error and correct-trial waveforms manifesting in the interval of 250 - 350 ms post response. Analysis of the ERP waveforms representing the 6 levels of confidence reported in the task revealed statistically significant differences between the Pe amplitude for each pair of confidence levels, strongly suggesting that Pe amplitude is modulated as a function of decision confidence [14].

2.2.2.2 Time-Frequency Analysis

Neural oscillations contain a wealth of information as evidenced by countless studies over many decades linking specific patterns of oscillations to perceptual, cognitive, motor, and emotional processes [5]. Neural oscillations contain multiple frequencies that can be separated using signal-processing techniques such as the Fourier transform and wavelet transform and are commonly grouped into bands that are defined by logarithmically increasing center frequencies and frequency widths [6]. These bands include delta (2-4 Hz), theta (4-8 Hz), alpha (8-12 Hz), beta (15-30 Hz), and gamma (30-80 Hz). Of these frequency bands, alpha has shown particularly promising results as a neural indicator of decision confidence. Additionally, because these techniques reduce the

dimensionality of the signal down to only a few frequency components, these components may be more useful features for classification via machine learning than the original signal.

2.2.2.2.1 Fourier and Short-Time Fourier Transform

The Fourier transform of a signal $x(t)$ is given by

$$X(f) = \int_{-\infty}^{\infty} x(t) e^{-i2\pi f t} dt$$

The power spectrum of $x(t)$ is given by $S(f) = |X(f)|^2$ and describes the distribution of power into the frequency components that make up $x(t)$ [15]. There are two major limitations of using the Fourier transform for EEG time-frequency analysis. First, the Fourier transform obscures the temporal dynamics in the frequency structure of the data. Second, the Fourier transform assumes that the signal $x(t)$ is stationary, which is clearly violated by the dynamic properties of the brain reflected in EEG data. These limitations are addressed by a simple extension known as the short-time Fourier transform (STFT). The STFT uses a window function w which is nonzero for a short period of time to compute an approximation of the Fourier transform and is given by

$$X(\tau, f) = \int_{-\infty}^{\infty} x(t) w(t - \tau) e^{-i2\pi f t} dt$$

The window function is typically symmetric and has unit L_2 norm. Common windows include the Hann window, Hamming window, and Gaussian window. The power spectrum of $x(t)$ is now a function of time and computed as $S(t, f) = |X(t, f)|^2$.

2.2.2.2.2 Wavelet Transform

Though the STFT addresses issues with frequency decomposition of EEG data by the Fourier transform, it is not without its own limitations. One issue with the STFT is that it uses the same window function for all frequencies. If the size of the window is large, resolution in the time domain is degraded and if the size of the window is small, resolution in the frequency domain is degraded. An alternative to the STFT is the wavelet transform

$$W(s, t_0) = \int_{-\infty}^{\infty} x(t)\psi_{s,t_0}^{*}(t)dt$$

where $\psi(t)$ is a continuous function in both the time and frequency domain called the mother wavelet, * denotes the complex conjugate, s is the timescale and t_0 is the center of the window. Wavelets are generated using the mother wavelet

$$\psi_{s,t_0}(t) = \frac{1}{\sqrt{s}}\psi_0\left(\frac{t - t_0}{s}\right)$$

and are simply translated and scaled versions of the mother wavelet. From the equation for the mother wavelet, it is clear that when the local area contains a high frequency, the wavelet gets shorter, and when the local area contains a low frequency, the wavelet gets longer [16].

2.2.2.2.3 Alpha Oscillations

Alpha oscillations are neural oscillations in the frequency range of approximately 8 and 12 Hz. These oscillations occur over the entire scalp but are typically maximum in amplitude in the parieto-occipital areas [17]. Alpha oscillations have been implicated in perceptual uncertainty and difficulty in decision making [18]. Several studies investigated how decision confidence modulates neural signals in individuals who explicitly reported their subjective confidence in perceptual decision tasks [19], [20], [21]. A common result

between the studies was that confidence was strongly encoded in alpha oscillations. In particular, for decision falling under the assumptions of the DDM, alpha power is lower for decisions made with high confidence and alpha power is higher for decisions made with low confidence.

Kubanek et al. collected EEG data from 10 participants during a perceptual decision task [19]. For this task, participants fixated on a monitor that displayed a picture of a joystick on the right half of the screen and a picture of an eye on the left. While fixating on the monitor, participants were presented with a stereo auditory stimulus in the form of clicks. If more clicks were heard in the right ear than the left, participants pressed a button with their right hand. If more clicks were heard in the left ear than right, participants made an eye movement towards the icon of an eye. After making their choice, participants were presented with a prompt asking them to rate their confidence level in a binary manner. Time-frequency analysis was carried out by using an autoregressive model of order 15 to estimate the power spectral density for each frequency from 1 to 80 Hz. The EEG signals were evaluated in 300 ms windows sliding through the trial in 30 ms timesteps. The neural representation of choice confidence was investigated using a regression model where power at a given time and frequency was regressed on confidence, where confidence is a two-level dummy variable for sure or unsure. This regression was carried out for each timestep and frequency. The p-values of the confidence effect for each regression were compared, showing that the effect was particularly significant in the alpha band for button press choices. Further analysis revealed a negative correlation between alpha and confidence. The authors interpreted these results as alpha reflecting a variable related to a degree of a subject's confidence [19].

Graziano et al. employed a partial report paradigm designed to separate the sensory encoding stage that begins with stimulus presentation, from the retrieval stage that begins after presentation of the response cue [20]. For this task, participants first fixated on a cross located in the center of a 19-inch screen for 1000 – 1500 ms. A stimulus in the form of an 8-letter circular array around the cross was then presented for 153 ms. Following a 750 ms delay, an array of 8 dots with exactly one being red was presented in the same position as the 8-letter array for 153 msec. After waiting for another 1000 ms, participants had to report the letter that was in the same position as the red dot, as well as their confidence in their decision on a 0-100 scale. EEG data was divided into two periods corresponding to the encoding and retrieval stages and transformed to the frequency domain for each trial using a Fourier transform. The authors observed that trials in which the participant was confident were accompanied by a lower alpha power during the encoding phase [20].

Samaha et al. measured prestimulus alpha power as a trial-by-trial index of cortical excitability through a two-choice orientation discrimination task [21]. Participants were tasked to identify whether sinusoidal luminance gratings embedded in random dot noise presented within a circular aperture were rotated left or right of vertical. Each trial began with a 500 – 1000 ms fixation period, followed by stimulus presentation for 33 ms. After a 600 ms waiting period post-stimulus presentation, participants were then asked to report their confidence as one of four levels. Time-frequency analysis was performed on the preprocessed EEG data using wavelet transformation. Data from each channel and trial were convolved with a family of complex Morlet wavelets from 2-50 Hz in Hz steps with wavelet cycles increasing linearly as a function of frequency. A non-parametric single-trial

multiple regression approach was used to relate single-trial estimates of power across time and frequency to decision confidence. Additionally, a binning analysis in which decision confidence was binned into 10 deciles according to prestimulus alpha power levels obtained from a fast Fourier transform (FFT) of prestimulus data was conducted. Both the regression and binning analysis revealed a strong negative relationship between prestimulus alpha power and confidence ratings [21].

2.2.3 Behavioral Indicators of Decision Confidence

Engelke et al. analyzed the relationship between quality scores, reaction times, and confidence ratings in a subjective image quality experiment [22]. Fifteen participants were tasked to rate the quality of 80 images and report the confidence level in their decision using a five-level Likert scale. The authors tested the hypothesis that confidence of a human observer when rating the quality of an image is strongly related to the response time of the quality rating and expected that images that were harder to judge to be associated with longer response times. They found that reaction time was strongly negatively correlated with confidence ratings: reaction times were shorter when participants had high confidence in their quality score and longer when they had low confidence. Similarly, Robitza and Hlavacs investigated the relationship between participant rating times and self-reported confidence in a subjective video quality experiment [23]. For this experiment, 27 participants were tasked to rate the visual quality of 135 ten-second video clips and give their confidence about their decision using a five-level Likert scale. The authors investigated average quality rating time as a function

of confidence score which showed a strong negative correlation between rating times and subjective confidence.

Boldt and Yueng investigated the relationship between subjective confidence and information seeking in two perceptual decision tasks of varying difficulty [24]. More specifically, the authors tested the hypothesis that subjective confidence predicts information seeking in decision-making. In their study, the authors created two conditions which were matched for accuracy but differed in subjective confidence. It was found that confidence tended to be lower in a condition with high evidence variability relative to a condition with low evidence mean. Another significant finding was the observation that a participant's decision to seek more information tracked subjective confidence, but not objective accuracy. It was observed that participants consistently chose to use the available means to seek information more often when evidence variability was high than when evidence mean was low. This relationship was observed in both experiments, each consisting of a different experimental setup and task difficulty.

2.2.4 Inference of Decision Confidence Through Machine Learning

A small subset of studies moved beyond investigating the behavioral and neural encoding of subjective confidence and instead examined whether or not these representations could be used to predict a subject's confidence level. Techniques for predicting qualitative responses, such as discrete categories of confidence, fall into the machine learning problem known as classification [25]. Specifically, classification is the problem of identifying the class membership of a new observation based on a training set of data containing observations whose classes are known. In the context of estimation of

decision confidence, the classes are the different levels of confidence as defined by the specific experiment, and the observations are any of the neural representations discussed in the previous section.

2.2.4.1 Logistic Regression

Logistic regression is a parametric method of classification used to fit a linear model that directly predicts the posterior probability that a sample $X = x$ belongs to class k. However, instead of invoking Bayes' theorem and generating probabilistic models from prior information, logistic regression generates boundaries that maximize the likelihood of the data from a set of class samples [26]. In the case of binary classification, the logistic regression model is given by

$$\Pr(\mathrm{G} = 1|\mathrm{X} = \mathrm{x}) = \frac{\mathrm{e}^{\beta_0+\beta_1 X_1+\cdots+\beta_p X_p}}{1 + e^{\beta_0+\beta_1 X_1+\cdots+\beta_P X_p}}$$

To fit the model, the method of maximum likelihood is utilized to estimate the regression coefficients. The coefficient estimates are chosen to maximize the likelihood function

$$\ell(\beta) = \sum_{i=1}^{N} \left\{ y_i \beta^T x_i - \log\left(1 + e^{\beta^T x_i}\right) \right\}$$

Shih et al. investigated whether a combination of EEG, pupil dilation, heart rate, and response time data collected during a simulated crew station experiment could be used to estimate decision confidence and accuracy [27]. The authors conducted their experiment using the Small Team Reconnaissance and Urban Surveillance Missions (STRUM) multi-attribute task battery (MATB), which was designed to emulate drone operator workload. The STRUM experiment setup consists of a two-seat, multi-screen crewstation with camera feeds, satellite maps, and text message feeds. The experiment

focused on visual and auditory subtasks. For the visual subtask, an icon tinted in one of four colors was briefly shown in one of the four quadrants of the satellite map screen. After one to three seconds, the subject was presented with a cue asking them to identify either the quadrant the icon was located in or its color. The subject could either answer within six seconds or choose to skip. For the auditory subtask, a sound was played from one of three locations and the subject was asked to identify the direction in which the sound came. Once again, the subject could either answer within six seconds or choose to skip. Responses were scored as +2 for correct, -2 for incorrect, -1 for skipped, -2 for missed, and total score was transformed into monetary compensation after the experiment. Neural and physiological data was collected utilizing 205-channel EEG, 2-channel electrooculography (EOG), and a custom head-mounted eye tracker. Response time was measured as the time from presentation of cue to the time of response by the subject. EEG data was windowed based on the onset of the stimulus and on the onset of the cue, resulting in six 250 ms windows. Pupil data was windowed around the stimulus resulting in 5 two-second windows. Average heart rate was computed over a time period of 6 seconds around the stimulus. Response time was used directly. Logistic regression models were fit using every combination of the features above. Classification using multiple features was done using a two-layer hierarchical logistic regression. The EEG and pupil data were used as features for the first layer which output scores that discriminated the data between whether the subject would be correct in their decision or skip making a choice or whether they were correct or not. These scores along with heart rate and response time were then used at the second layer to output a final score for discriminating the data between conditions. The best performing models achieved an average accuracy of

70-75% and included the stimulus-windowed EEG, cue-windowed EEG, and pupil features. It was also noted that for both the audio and visual subtasks, the EEG data windowed on the cue to respond best predicted correct vs. skipped conditions for the single-feature models [27]. An issue with the results presented in this study is that the distribution of observations with respect to class membership is not given. If 70-75% of the data represents a single class, then a naïve classifier which always predicts the majority class would achieve the same results.

Based on their result that rating times were strongly negatively correlated with observer confidence, Robitza and Hlavacs investigated whether observer confidence could be inferred from rating time using a multinomial logistic regression model in which the confidence score was used as the ordinal dependent variable and rating time was used as the sole feature [23]. Before fitting the model, extreme outliers where rating times were over 10 seconds were removed. The authors observed that the probability of classifying an observation as one of the higher confidence classes decreased as rating time increased and that the probability of classifying an observation as one of the lower confidence classes increased as rating time increased.

2.2.4.2 Linear Discriminant Analysis

Linear discriminant analysis (LDA) is a parametric method of classification that attempts to find linear combinations of features that best separates the groups of observations [26]. LDA models the class densities as multivariate Gaussian distributions given by

$$\mathrm{f_k}(x) = \frac{1}{(2\pi)^{p/2}|\Sigma_k|^{1/2}} e^{-\frac{1}{2}(x-\mu_k)^T\Sigma_k^{-1}(x-\mu_k)}$$

where μ_k and Σ_k are the mean and covariance matrix of class k respectively and $\Sigma_k = \Sigma\ \forall k$. Optimal classification requires the posterior probability that a sample $X = x$ belongs to class k be known. Given π_k, the prior probability of class k, application of Bayes theorem gives the posterior as

$$\Pr(\mathrm{G = k|X = x}) = \frac{\mathrm{f_k}(x)\pi_k}{\Sigma_{l=1}^{K} f_l(x)\pi_l}$$

In the case of binary classification of classes k and l, it is sufficient to look at the log-ratio

$$\log\frac{\Pr(G = k|X = x)}{\Pr(G = l|X = x)} = \log\frac{f_k(x)}{f_l(x)} + \log\frac{\pi_k}{\pi_l}$$

$$= \log\frac{\pi_k}{\pi_l} - \frac{1}{2}(\mu_k + \mu_l)^T\Sigma^{-1}(\mu_k - \mu_l) + x^T\Sigma^{-1}(\mu_k - \mu_l)$$

which is a linear equation in x and implies that the decision boundary separating classes k and l is also linear in x. This can be generalized for any pair of classes and so the decision boundary between any pair of classes is linear and corresponds to the linear discriminant function

$$\delta_\mathrm{k}(x) = x^T\Sigma^{-1}\mu_k - 1/2\mu_k^T\Sigma^{-1}\mu_k + \log\pi_k$$

with class membership determined by the decision rule

$$\mathrm{G(x) = argmax_k}\,\delta_k(x)$$

Kubanek et al. applied their results discussed in 2.2.2.2.3 to predicting whether a subject was going to be sure or unsure of pressing a button [14]. The authors averaged the EEG alpha power over all channels in the period of statistical significance of the effect of confidence creating a single feature per observation that was input into an LDA classifier.

Their classifier achieved an accuracy of 0.60 using a hold-out test set. Like results presented by Shih et al. [27], distribution of observations with respect to class membership is not presented.

2.2.4.3 Support Vector Machines

Support Vector Machines (SVM) is a method of classification that can be used for both linearly and non-linearly separable data. Similar to LDA, SVM uses a separating hyperplane as the decision boundary that separates the two classes. The decision boundary is associated with a pair of hyperplanes that are parallel to it, with each passing through the datapoint nearest to it. The distance between these supporting hyperplanes is known as the margin. In the case where the data is linearly separable, the decision boundary is chosen so that it maximizes the margin. If the data is not linearly separable, a linear boundary may be obtained if the data is transformed to a higher dimensional space. The non-linear classification is done using a kernel function that replaces the computationally expensive inner product of the feature vectors in the higher dimensional space. Additionally, the hard-margin in the linear case is replaced by a tunable soft-margin which adjusts model flexibility [25].

Paul et al. sought to identify the neural patterns corresponding to actions with and without decision-making through classification of reference and decision trials in an instrumental reward-based learning task [28]. For this task, 13 participants were presented with a series of trials in which they chose between abstract visual images in order to accrue a small reward at the end of each trial. For reference trials, participants were presented a single image to select from, whereas they were presented two images for

decision trials. Each trial duration was approximately 4 seconds. The problem was formulated as a binary classification problem in which the two classes were whether the participant was making a decision or not. 64-channel EEG data was windowed on each trial and the mean amplitude over each 4 second window was computed for each channel, generating 64 features per trial for classification using SVM. On the individual subject level, average classification accuracy was reported as approximately 90%. Feature saliency analysis indicated that channels associated with the frontal areas of the brain were most important, consistent with the notion that these areas are implicated in the decision process [28].

2.2.4.4 Linear Spatial Integration

Despite research backing several ERP components as a robust index of decision confidence, the trial-averaged ERP is not an appropriate feature for classification as it condenses the associated observations into a single one, leading to the unfavorable situation in which the number of features is much larger than the number of observations. Parra et al. propose integrating information over space as an alternative to the trial-averaging methodology of standard ERP analysis [29]. Specifically, the method uses logistic regression to find the optimal spatial weighting such that the resulting spatial distribution of electrode activity in a given time window maximally discriminates between two conditions of interest. After finding the optimal spatial weighting, the discriminating component is averaged over the dependent samples for each trial. The resulting value ranges from 0 to 1 which can be conceptualized as the probability that the condition of interest for that trial is the first condition [30]. Improvement in signal-to-noise ratio is

achieved in single trials because the ERP component amplitude is estimated as a spatially weighted average across electrodes for each trial in much the same way as conventional ERP analysis achieves a high signal-to-noise ratio through cross-trial averaging [29].

Gherman and Philiastides utilized the spatial linear integration method to discriminate between certain versus uncertain trials to identify the temporal characteristics of the neural correlates of decision confidence during a binary, delayed-response task [31]. For each trial of the task, 19 participants had to determine whether a visual stimulus presented for 0.1 sec, displayed at one of three possible levels of sensory evidence, was either a face or a car and had 1 sec to indicate their response. Each trial began with a randomized delay between 1 and 1.5 sec and each stimulus presentation and response cue were separated by a randomized delay between 0.9 and 1.4. Correct responses were incentivized with monetary compensation and in a random half of the trials, participants were offered the option to opt out of giving a response for a smaller but sure reward. The spatial linear integration method was applied in the time range between 100 ms prior to and 1000 ms after the presentation of the stimulus. The optimal spatial weighting was identified for a 60 ms sliding training window centered in increments of 10 ms within the time range described above. Performance was assessed using the area under a receiver operating characteristic curve. The authors observed that the classifier's performance gradually increased after 300 ms and was maximum at around 600 ms with an AUC of approximately 0.75 [31].

Boldt and Yueng also trained a classifier on single-trial Pe amplitude using the spatial linear integration method to predict confidence on a single-trial level [14]. The authors' goal was to assess whether a classifier that was trained to distinguish between

objectively correct and incorrect responses could also be used to predict levels of decision confidence on a holdout set of correct responses. The authors found that the Pe-trained classifier was predictive of fine-grained differences in correct-trial confidence, suggesting that information reflected by the Pe includes both graded certainty about having made an error as well as graded certainty of having made a correct response [14].

2.3 Research Gaps

2.3.1 Experimental Designs

Since the experimental designs in the surveyed research are based on decision-making models like the DDM, it is no surprise that they inherit similar limitations. It was previously stated that the DDM is only applicable to single-stage decisions in which the mean reaction times are less than 1000 to 1500 ms [5]. This limitation begs the question of applicability to the realm of the cyber analyst's investigation of cyber alerts. The current body of research into decision confidence focuses solely on experimentation in which the decisions made by participants are discrete and forced to occur at a specific time. While experimental designs of this kind have been shown to produce EEG data that is convenient for analysis techniques such as ERP and time-frequency, it is currently unknown whether the results of conducting an EEG analysis on data generated in this manner will generalize to real-world decisions that unfold gradually as they are shaped by a continuous stream of sensory inputs.

2.3.2 Inferring Decision Confidence

Other than the studies previously mentioned, there is a lack of research into estimation of decision confidence utilizing its neural representations. In particular, despite

several studies identifying alpha power as robust index of decision confidence, there has been little research utilizing it as a feature for inferring decision confidence. Instead, existing studies have leaned on time series data windowed on events of interest as the feature of choice. The number of classification techniques utilized for inferring decision confidence was also extremely limited, with no research applying more recent machine learning approaches such as deep learning methods.

2.4 Related EEG Research

Though little research has been conducted with respect to applying machine learning approaches to inferring decision confidence from EEG data, there exists a large body of research which uses machine learning to infer other cognitive processes. The following is a review of machine learning approaches that have been successful in EEG analysis in other domains, which may be applicable to inferring decision confidence.

2.4.1 Random Forests

Random forest models are an ensemble learning method for classification in which the ensemble consists of many decision trees [32]. A decision tree algorithm recursively partitions the data into smaller subgroups until some criteria is met. At each split, the algorithm finds the best feature in the dataset to partition the data into subsets which have similar values for that feature. A random forest model is created by growing many decision trees trained on a random subset of the available features. By using a random subset of the available features, the set of trees are decorrelated, resulting in better generalization of the models. The overall prediction of the model is determined from a function of the individual predictions made by the decision trees – for example, a vote for

the most prevalent class or a computation of the regression mean of the values predicted by the trees.

Random forest models trained using EEG data have performed exceptionally well in the area of estimating pain experienced by humans. Vijayakumar et al. developed a robust and accurate cross-participant machine learning approach to quantify tonic thermal pain in healthy human subjects using a random forest model trained using time-frequency wavelet representations of independent components obtained from EEG data [33]. 64-channel EEG data was collected from twenty-five participants and was concatenated across all participants. Each datapoint corresponded to one of ten classes of pain and the overall distribution of classes was non-uniform. The EEG data was subjected to full rank Independent Component Analysis (ICA) to enable multivariate analysis by focusing on the fraction of source information available at each scalp. Each independent component was transformed using the continuous wavelet transform and Gabor wavelet. The power spectral density was computed for 60 points corresponding to a frequency range of 2 - 80 Hz. Training was done using leave-one-out cross-validation and tested on data from a hold-out test participant. Due to the non-uniform distribution of pain classes, balanced classification accuracy, F-score, and Matthew's correlation coefficients were used as performance metrics for assessing model performance. The best performing model achieved a balanced classification accuracy of 0.89, the highest among existing classifiers for this dataset. In addition to classification, the authors investigated the salience of each frequency band and found the Gamma band to be most important.

2.4.2 Artificial Neural Networks

Artificial neural networks (ANN) are machine learning models inspired by the biological neural networks of the brain [34]. ANNs specializes in learning complex data representations that are expressed in terms of other, simpler representations. Beginning at the raw data representation level, this layered representation is obtained through simple non-linear transformations from one level of representation to a higher one that is slightly more abstract [35].

2.4.2.1 Fully Connected Neural Networks

The first and simplest type of ANN is the fully-connected neural network shown in Figure 2.1. Each unit (neuron) in the hidden layer computes a weighted sum of its inputs, followed by a nonlinear activation function. The output of the nth layer is given by

$$\mathrm{x_n} = f(W_n^T x_{n-1} + b_n)$$

where f is the nonlinear activation function, x_{n-1} is the input to the nth layer, W_n is a matrix of weights that describes a mapping from x_{n-1} to x_n, and b_n is a vector of biases. The aim of the network is to modify the parameters of the model until the network maps the input to the desired output. Learning the parameters involves minimizing a loss function, which is done via an optimizer and the backpropagation algorithm [36].

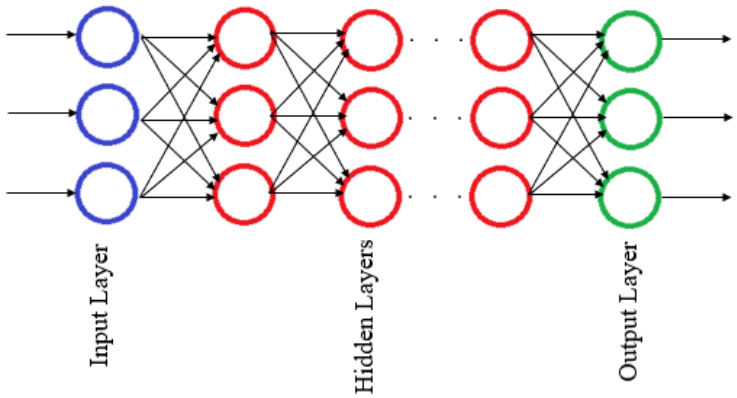

Figure 2.1: Fully-Connected Neural Network

ANNs have performed well with respect to classification of cognitive workload using non-stimulus locked EEG data and typically outperform other machine learning classification methods in this domain. Wilson et al. investigated the performance of a single, 43-node hidden-layer fully connected ANN with respect to online classification of operator workload using EEG data [37]. EEG data was collected from 8 participants performing the NASA Multi-Attribute Task Battery (MATB) task at one of three levels of workload: baseline, low, and high. Data was collected on a single day during three 5-minute sessions, each corresponding to one of the three levels of workload. The raw EEG data was transformed to the frequency domain using the FFT so that the average power could be computed in each of the five traditional EEG bands using a 10-second sliding window with 5 seconds of overlap. EEG bands included delta (1-3 Hz), theta (4-7 Hz), alpha (8-13 Hz), beta (14-30 Hz), and gamma (31-42 Hz). Data was segmented randomly such that 75.0% was used for training and 25.0% for validation. The trained network was then used for online classification of two additional blocks of the three workload levels. The mean classification accuracies were 85.0% for the baseline

condition, 82.0% for the low workload condition, and 86.0% for the high workload condition.

Christensen et al. assessed the cross-day stability of EEG data for use in classification of operator workload [38]. EEG data was collected from 8 participants performing the MATB task at two levels of workload with collection for each participant occurring over 5 days randomly distributed over a four-week period. Due the potential for a classifier trained on only one day's worth of data to key in on unstable features unique to that day, Christensen hypothesized that using multiple days in the training set would improve generalization. As in Wilson et al., feature engineering consisted of transforming the EEG data to the frequency domain and then computing the average power in each of the traditional EEG bands using a sliding window. Christensen divided the 5 days of data into various combinations of days and sessions within a day, with 50% being used for training and the remaining data used for validation and testing. Linear discriminant analysis (LDA), support vector machine (SVM), and ANN models were trained using cross validation. Christensen found that the ANN performed the best, with a classification accuracy of 83% when trained on the first 4 days of data and tested on the 5th day. A decline in the performance of all classifiers was observed as the amount of days in the training set was decreased.

2.4.2.1.1 Recurrent Neural Networks

Recurrent Neural Networks (RNN) are a type of neural network that specialize in learning sequences by maintaining a state containing information relative to what has been seen so far via a recurrent connection (internal loop) in the hidden layer [34]. The structure of a simple RNN is shown in Figure 2.2. An issue with the simple RNN is that it

is unable to retain information about inputs seen many timesteps earlier and thus unable to learn long term dependencies. This is due to gradients becoming extremely small during backpropagation, effectively preventing weights from changing value and rendering the network untrainable. The Long Short-Term Memory (LSTM) algorithm was designed to address this issue. The LSTM algorithm combats these vanishing gradients by adding mechanisms which provide control over which pieces of information to remember, which to update, and which to focus on.

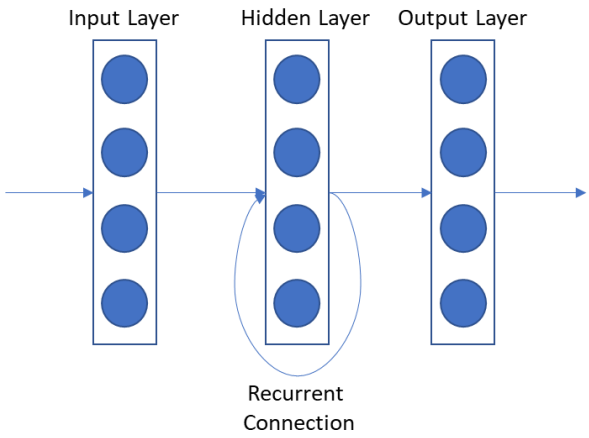

Figure 2.2: Simple Recurrent Neural Network

RNNs have been shown to handle the temporal non-stationarity of EEG signals and often outperform the machine learning models previously discussed. Hefron et al. extended the work of Christensen by investigating the use of deeply recurrent neural networks to account for the temporal dependence in EEG-based workload estimation on the same dataset [39]. Feature engineering was conducted in a manner similar to Christensen, however, Hefron also computed the variance, skewness, and kurtosis of the power distribution for each window. Hefron explored the performance of several models on all combinations of mean, variance, skewness, and kurtosis features. The data was split

such that the first four days were used for training and cross-validation while the last day was reserved for testing. The model the highest performance was a deep LSTM model which consisted of an LSTM layer with 50 hidden units, followed by an LSTM layer with 10 hidden units with a dropout of 20% on the inputs, followed by a fully-connected layer with a sigmoid activation function for classification. Hefron's deep LSTM architecture achieved a classification accuracy of 93.0%, representing a 59.0% decrease in error compared to the best published results for the dataset.

2.4.2.1.2 Convolutional Neural Networks

A Convolutional neural network (CNN) is a type of neural network that specializes in learning translation invariant spatial hierarchies of patterns [34]. Three main types of layers are used to build CNNs: The convolutional layer, pooling layer, and fully-connected layer. In the convolutional layer, local patterns are learned by convolving the input with a set of kernels. This is followed by application of an elementwise activation function such as a Rectified Linear Unit (ReLU) to produce an activation map. The pooling layer performs a downsampling operation along the spatial dimensions of the activation map. Finally, the fully-connected layer computes the class scores used to classify the input. CNNs have been shown to outperform other machine learning models including fully connected ANNs in the area of emotion classification from EEG data. Tripathi et al. used a CNN to classify human emotion using EEG data from the DEAP dataset which represents the benchmark for emotion classification research [40]. The DEAP dataset consists of 40-channel EEG data recorded from 32 participants as they watched 40 one-minute extracts of music videos and gave an online self-assessment based on arousal, valence, and dominance for each video. However, the authors restricted their

research to classifying levels of valence and arousal for both the binary (high or low) and 3 class (high, normal, or low) problem. The raw data structure was a 40 x 40 x 8064 array corresponding to trial x channel x data. The authors divided the 8064 readings per channel into batches of approximately 807 readings each. For each batch they extracted the mean, median, maximum, minimum, standard deviation, variance, range, skewness and kurtosis values. They further incorporated the same values computed over the 8064 readings along with the experiment and participant number for a total of 101 values per channel. The input to their CNN was then a 2D array of shape 40 x 101. Their CNN consisted of two convolutional layers followed by a max pooling layer with a 50% dropout on the inputs followed by a 128-node fully connected layer with a tanh activation function and 25% dropout on the inputs followed by 2-node or 3-node fully connected layer with a softplus activation. Their model used categorical cross entropy as the loss function and stochastic gradient descent as the optimizer. Their model achieved an accuracy of 0.814 and 0.734 for binary classification of valence and arousal levels and an accuracy of 0.668 and 0.576 for 3-class classification of valence and arousal levels. Their results represent a 4.51 and 4.96 and 13.39 and 6.58 percentage improvement over the best published results for this dataset.

2.5 Summary

In summary, very little research has delved into inferring decision confidence through behavioral and electrophysiological signals using machine learning approaches. Within the body of research that exists, the use of behaviors such as reaction time and information seeking and electrophysiological features such as stimulus windowed time-

series data and time-frequency representations of confidence appears promising. Despite these promising results, the fact that the analyses surveyed in this work have focused on investigating the neural representation of decision confidence for decisions that meet the assumptions of the DDM must be emphasized, as it is currently unknown whether these results will generalize to more complex decisions encountered in the operational environment.

III. Methodology

3.1 Chapter Overview

This chapter describes the methodology used for the collection and analysis of data for this research. First, the chapter discusses the research questions and hypotheses. Then, a description of the experiments that were performed, including the makeup of the participants, required assumptions and the various factors and variables which were changed is given. This is followed by a description of the data acquisition process and data wrangling techniques used to create a dataset. Finally, the analysis strategy that is used in Chapter IV is presented.

3.2 Background

The objective of this research is to determine if human electrophysiological signals and human behavioral features can be used to infer decision confidence in simple and complex decision-making environments. To complete this objective, the following research questions are investigated:

RQ1. Can electrophysiological features be used in combination with machine learning techniques to infer decision confidence in a simple decision with a performance greater than chance?

Hypothesis: Machine learning models will be able to learn the neural correlates of decision confidence and thus can be used to infer decision confidence in a simple decision with a performance greater chance.

RQ2. What are the salient electrophysiological features for inferring decision confidence in a simple decision?

Hypothesis: Changes in power in the five traditional EEG frequency bands (alpha in particular) will be prominent features for inferring decision confidence.

RQ3. Can behavioral features be used in combination with machine learning techniques to infer decision confidence in a complex decision with a performance greater than chance?

Hypothesis: Machine learning models will be able to learn correlations between decision confidence and reaction time and information seeking and thus can be used to infer decision confidence in a complex decision with a performance greater than chance.

RQ4. Are the salient electrophysiological features for inferring decision confidence the same for both simple and complex decisions?

Hypothesis: Features identified as salient for a simple decision will still encode important information that can be used to infer decision confidence for complex decisions.

3.3 Experiment

The experiment conducted for this research consisted of two tasks, corresponding to simple and complex decision-making environments, accomplished during a single 2-hour period. A diagram of the experiment sequence is shown in Figure 3.1. The first phase was a modified two-alternative forced choice (2AFC) task [41] using random dot kinematograms (RDK) [3]. The use of a 2AFC experiment falls in line with previous

research in that it fits into the experimental paradigms based on the DDM [5], and thus the data generated could be utilized to extend the results of multiple observational studies identifying neural correlates of decision confidence. Additionally, no research has been conducted that uses machine learning for decision confidence inference using EEG data collected from an RDK motion discrimination task. The design of the second phase looked to extend these results even further by simulating a more realistic decision-making environment akin to what a cyber defense operator experiences during every day operations in which the assumptions of the drift diffusion model no longer hold. For both phases, Electroencephalography (EEG), Electrooculography (EOG), and Electrocardiography (ECG) data was collected while participants completed the tasks. Pre- and post-experiment questionnaires were given to each participant on the experiment day and can be found in Appendix C and D respectively.

Figure 3.1: Experiment Sequence

3.3.1 Participants

A total of 8 male participants were recruited for this research. All participants were voluntary military or government civilian personnel. Participant age ranged from 21 to 31 with a mean age of 24.8 and standard deviation of 3.60. All participants had at a minimum, a Bachelor's Degree, and used electronic devices on a daily basis for both work and personal use. Two participants had previously completed courses in cyber security

education and one participant had earned several cyber security certificates. Participants were not compensated for their participation. Exclusion criteria included: inability to use a mouse and keyboard, visual impairment or inability to view information on a computer screen, and specific motor, perceptual, or cognitive conditions which precluded them from operating a computer. Additionally, all participants had to consent to the placement of electrodes on their head, face, and chest. Participant consent was obtained prior to starting participation in the study.

3.3.2 Random Dot Kinematogram Task

Participants performed a perceptual decision-making task in which they had to determine the global direction of motion (left or right) of dots in an RDK. The experiment interface was created using the PsychoPy API [42]. An example RDK is shown in Figure 3.2. The RDK was displayed on a 15-inch monitor with a resolution of 3840 x 2160 pixels and refresh rate of 60Hz. Participants sat in a comfortable chair 60 cm in front of the monitor. Each RDK consisted of an aperture with a 10 cm diameter creating a visual angle of 9.5° which 200 white dots (2 x 2 pixels) moved on a black background. A subset of dots (signal dots) within the aperture moved coherently in either the left or right direction, while the remaining dots (noise dots) each followed a random, but constant direction. The motion coherence level for each RDK was defined as the number of signal dots divided by the total number of dots. All dots moved at a speed of 6°/s and had a limited lifetime of 200 ms.

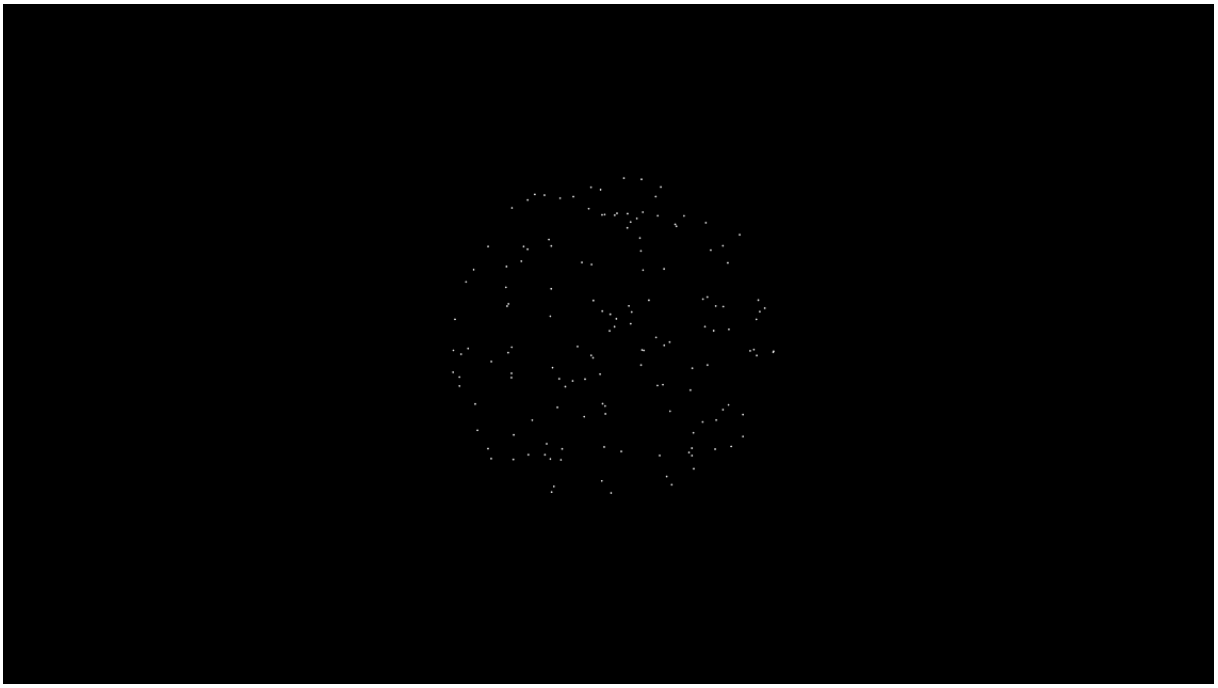

Figure 3.2: Example Random Dot Kinematogram

The experiment sequence is shown in Figure 3.3. Each trial began with the participants fixating on a cross for 1500ms before the stimulus presentation. The stimulus was presented for 400ms followed by a forced delay of 1000ms to allow for the evoked response in EEG to unfold without motor contamination [21]. The motion coherence level for each stimulus was randomly selected from seven levels, with the distribution of these levels intended to produce approximately 50% discrimination accuracy. After the forced delay, participants were prompted for their decision. Participants were given the option to use a right or left-handed decision input configuration for the entire task. Participants pressed the ‘A’ or ‘J’ key to indicate global motion to the left, the ‘D’ or ‘L’ key to indicate global motion to the right, or the ‘S’ or ‘K’ key to opt-out if they could not identify the direction of global motion. A scoring system was implemented to encourage participants to opt-out during low confidence trials [31]. Participants were awarded 1 point

for each correct answer, -1 points for each wrong answer, and 0 points if they chose to opt-out or did not input a response before time expired. Each participant completed a total of 440 trials.

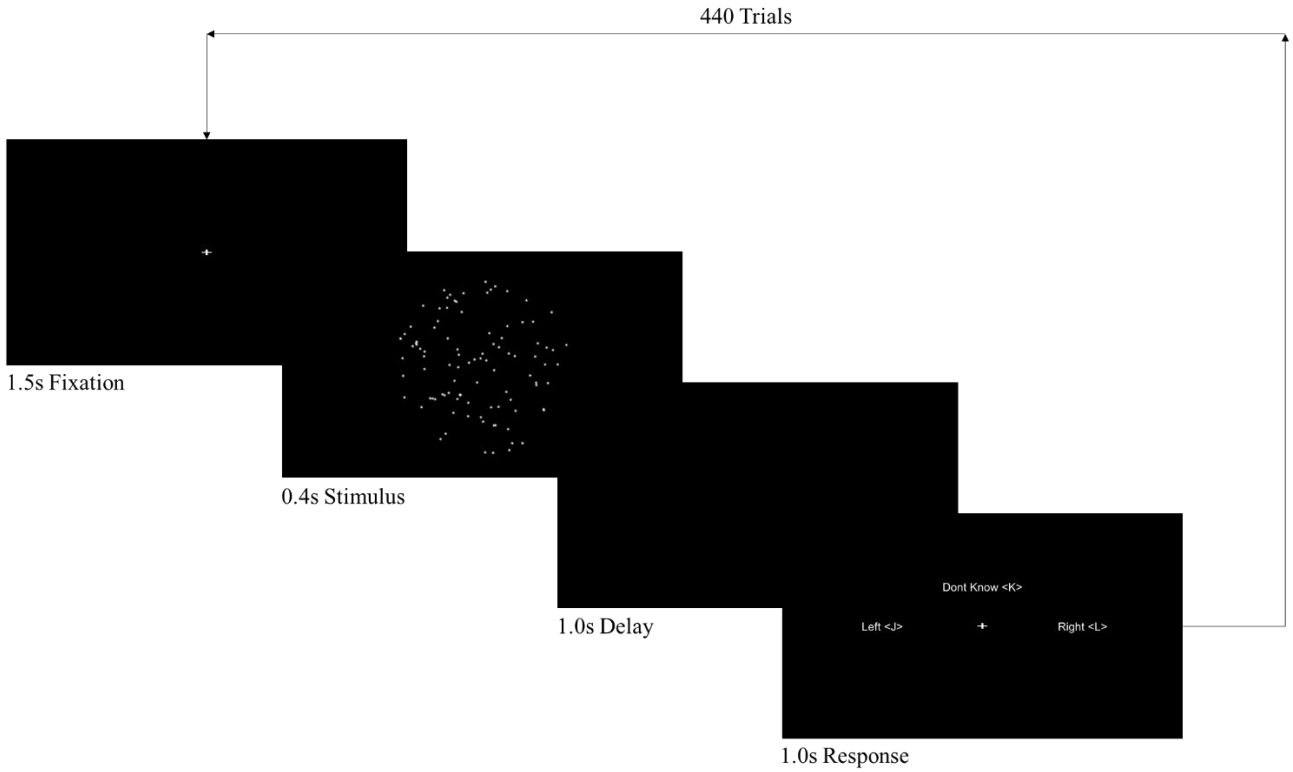

Figure 3.3: Random Dot Kinematogram Task Sequence

3.3.2.1 Response Variables

Participant decision confidence is the sole response variable for this experiment. Decision confidence is the degree to which a participant believes that their decision is correct. For this experiment, decision confidence is treated as a categorical variable where the participant is either confident or not confident. If the participant selects “Left” or “Right”, their decision is labelled as confident and if they select “Don’t Know” or run out of time it is labelled as not confident.

3.3.2.2 Independent Variables

The motion coherence level of the RDKs is the sole independent variable for this experiment. Seven levels of coherence (10%, 20%, 30%, 40%, 50%, 70%, and 80%) were chosen based on research by Pilz et al. which examined motion coherence and direction discrimination in healthy adults [43]. The motion coherence levels were approximately evenly distributed over the 440 trials and order was determined through randomization and then held constant for each participant.

3.3.2.3 Control Variables

There are five control variables for this experiment: aperture size, number of dots, dot speed, dot lifespan, and stimulus duration. Table 3.1 provides a summary of the control variables for this experiment.

Table 3.1: Control Variables

Factor	Desired Experimental Level	How Controlled
Aperture size	10 cm diameter	Experiment configuration
Number of dots	200	Experiment configuration
Dot speed	6°/s	Experiment configuration
Dot lifespan	5 frames	Experiment configuration
Stimulus duration	400 ms	Experiment configuration

Aperture size was chosen to minimize participant eye strain. The number of dots and dot speed was set based on pre-trial experimentation such that no individual dot could be easily tracked by a participant. Dot lifespan was set so that the distribution of dots within the aperture was approximately uniform. The stimulus duration was chosen based on the research by Pilz et al. [43].

3.3.3 Cyber Intruder Alert Testbed Task

Participants performed a simulated cyber investigation typical of a first line computer network defense analyst. The investigation was conducted using a modified version of the Cyber Intruder Alert Testbed (CIAT) synthetic task environment (STE) [2]. An example of the CIAT interface is shown in Figure 3.4. Participants investigated 30 cyber-alerts designed by a subject matter expert, where each alert had one of four levels of difficulty. The goal of each alert investigation was to determine the validity of the alert based on information available from various tools. Table 3.2 outlines the available tools and their functionality.

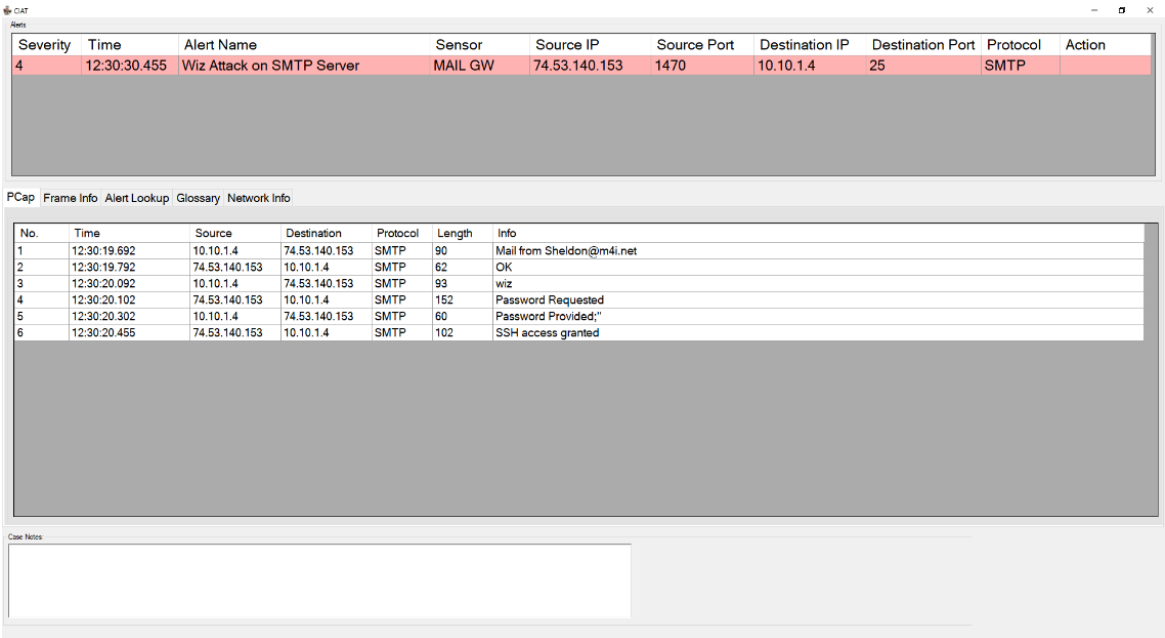

Figure 3.4: CIAT Interface

Each alert investigation lasted for 2 minutes. Every 30 seconds, participants were queried via a popup which asked them to assess the current alert by selecting one of three options via a button-press as shown in Figure 3.5. Participants selected “Threat” if they

believed the alert was legitimate, “False Alarm” if they believed the alert was generated in error, or “I Don’t Know” if they were unsure at the time of the prompt. Participants had 5 seconds to submit their answer, which did not count towards the 2-minute trial time. If time expired before the participant could submit their answer, it was logged as a “I Don’t Know”. The same scoring system from the first phase was utilized to encourage participants to select “I Don’t Know” when their confidence was low. The test matrix for the CIAT experiment is shown in Appendix A. During the course of the experiment, the timing of every mouse click and keyboard input was recorded for each participant. This data was then reconstructed into a workflow and timeline that could be used to replay participant behavior during the investigation of every alert, including which tools were accessed and how long they were accessed for.

Table 3.2: CIAT Tools and Descriptions

Tool	**Description**
Packet Capture (PCap)	Displays raw packet information
Frame Information	Provides more detailed information corresponding to the rows of the PCAP tool, including additional log information
Alert Lookup	Provides a description of each alert with triggering information
Glossary	Defines common terms encountered during cyber investigations
Network Information	Contains information about whether certain IP addresses are known to be safe or dangerous

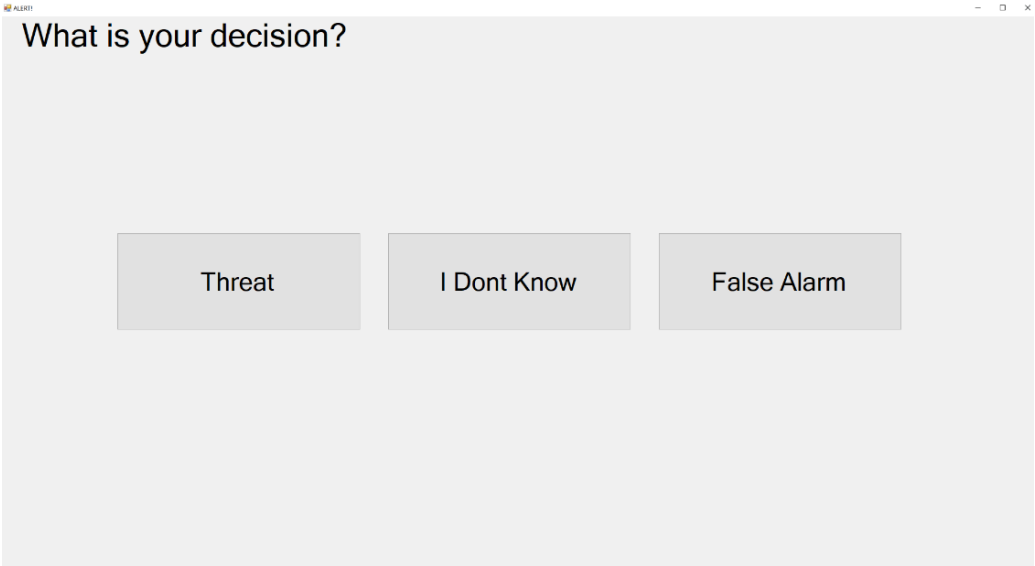

Figure 3.5: Example Decision Prompt

3.3.3.1 Response Variables

There are six primary response variables for this experiment: decision confidence, reaction time, number of tool transitions, EEG, ECG, and EOG. As in the first phase, decision confidence is the degree to which a participant believes that their decision is correct. For this experiment, decision confidence is treated as a categorical variable where the participant is either confident or not confident. If the participant selects “Threat” or “False Alarm”, their decision is labelled as confident and if they select “I Don’t Know” or run out of time it is labelled as not confident. Response time is the length of time taken for the brain to perceive and react to a stimulus. For this experiment, it is measured as the difference between the time at which the participant was prompted for a decision and the time at which they selected an option. Number of tool transitions is the number of times in which the participant switched between the available tools. EEG, ECG, and EOG are the electrophysiological signals to be recorded. Table 3.3 summarizes the response variables for this experiment and their associated measure.

Table 3.3: Response Variables for the CIAT Experiment

Response Variable	Type	Measurement
Decision Confidence	Categorical	[Confident, Not Confident]
Reaction Time	Numerical	Time (ms)
Number of Tool Transitions	Numerical	Quantity
EEG	Numerical	Voltage
ECG	Numerical	Voltage
EOG	Numerical	Voltage

3.3.3.2 Independent Variables

There are two independent variables for this experiment: alert difficulty and query number. Alert difficulty is categorical and has four levels: Easy, Medium, Hard, and Very Hard. The difficulty level for each alert was determined by a subject matter expert based on four factors: information availability, information needed, and information inconsistency. Information availability was measured as the amount of information relevant to the current alert that was available in the tools, information needed was measured as the number of tools that were required in order to accurately assess the alert, and information inconsistency was measured as the amount of conflicting information among the tools. The final distribution of difficulty for the 30 alerts is 10 Easy, 8 Medium, 6 Hard, and 6 Very Hard. Alert order was determined through randomization and then held constant for all participants and is summarized in the test matrix given in Table 3.4. The decision query number is also categorical with four levels and represents the amount of time participants have to investigate an alert before making a decision. Queries 1, 2, 3, and 4 occur at 30 seconds, 1 minute, 1 minute and 30 seconds, and 2 minutes of investigation time respectively. Table 3.5 summarizes the independent variables for this experiment and their associated measure.

Table 3.4: Test Matrix

Alert Number	Alert Difficulty	Truth
1	EASY	THREAT
2	EASY	THREAT
3	HARD	FALSE ALARM
4	VERY HARD	FALSE ALARM
5	HARD	THREAT
6	EASY	THREAT
7	EASY	FALSE ALARM
8	HARD	FALSE ALARM
9	EASY	THREAT
10	EASY	FALSE ALARM
11	MEDIUM	FALSE ALARM
12	VERY HARD	FALSE ALARM
13	MEDIUM	THREAT
14	EASY	THREAT
15	EASY	FALSE ALARM
16	EASY	THREAT
17	VERY HARD	FALSE ALARM
18	MEDIUM	FALSE ALARM
19	VERY HARD	FALSE ALARM
20	MEDIUM	FALSE ALARM
21	MEDIUM	THREAT
22	VERY HARD	FALSE ALARM
23	EASY	THREAT
24	HARD	FALSE ALARM
25	MEDIUM	THREAT
26	MEDIUM	THREAT
27	MEDIUM	FALSE ALARM
28	VERY HARD	FALSE ALARM
29	HARD	FALSE ALARM
30	HARD	THREAT

Table 3.5: Independent Variables for the CIAT Experiment

Independent Variable	**Type**	**Measurement**
Alert Difficulty	Categorical	[EASY, MEDIUM, HARD, VERY HARD]
Query Number	Categorical	[1, 2, 3, 4]

3.3.3.3 Control Variables

There are two control variables for this experiment: The number of alerts, and the alert time limit. The 30 alerts and their corresponding difficulties were designed by a subject

matter expert so as to give an approximately equal confidence distribution among the trials [2]. The time limit of two minutes per alert was imposed in order to better facilitate analysis of the electrophysiological data with respect to the research questions. If participants have an unlimited time to perform their investigation, it becomes significantly harder to identify areas in the electrophysiological data that correspond to decision-making and decision confidence. Thus, extracting salient features from non-stimulus aligned data is left as future work.

3.4 Electrophysiological Data Acquisition

For each phase of the experiment EEG, ECG, and EOG data was collected using the Cognionics Mobile-72 system [44], which is capable of collecting up to 72 channels of electrophysiological data. The electrophysiological data collection setup is shown in Figure 3.6. EEG data was collected using the 64 Ag/AgCl electrodes on the EEG headset. The layout of the electrodes is shown in Figure 3.7. Note that throughout the data collection process, a periodic malfunction raised the noise floor at random time points

Each participant had their head measured so as to identify an appropriately fitting headset. Participants wore the headset with the ground electrode placed on the nape of the neck and the reference electrode on the right mastoid.

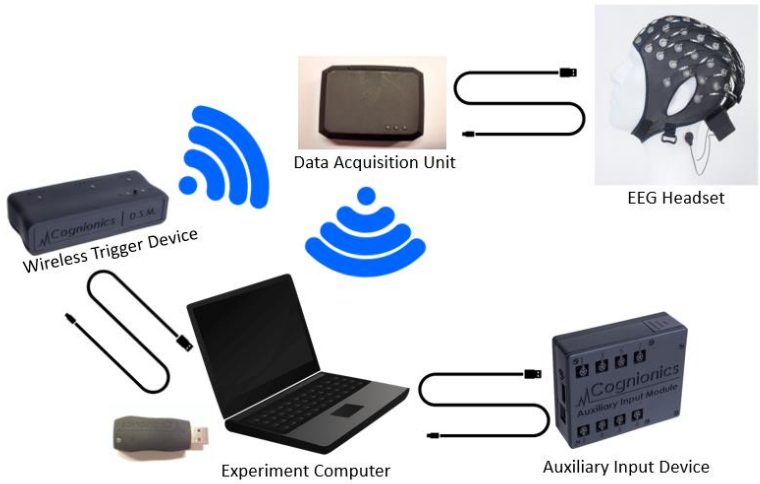

Figure 3.6: Data Acquisition Setup

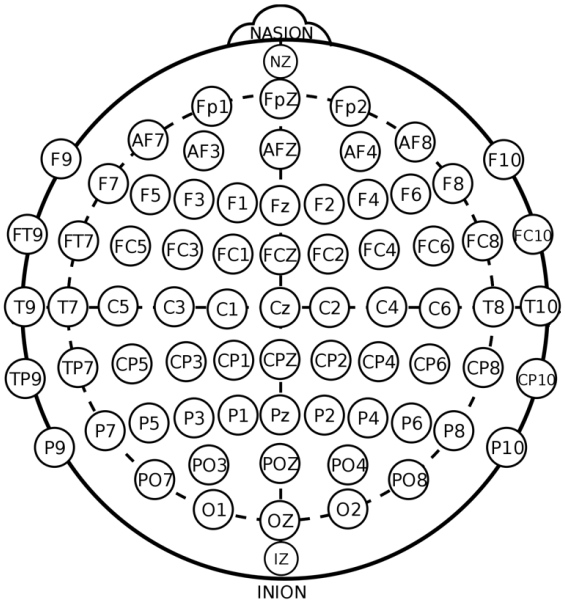

Figure 3.7: International 10-20 System

The EEG headset is connected through a wired connection to the Data Acquisition Unit (DAQ), which wirelessly transmits the EEG measurements to the data acquisition software on the experiment computer. Stimulus presentation and participant actions were

time stamped with unique trigger values using the wireless trigger device. An example is shown in Figure 3.8.

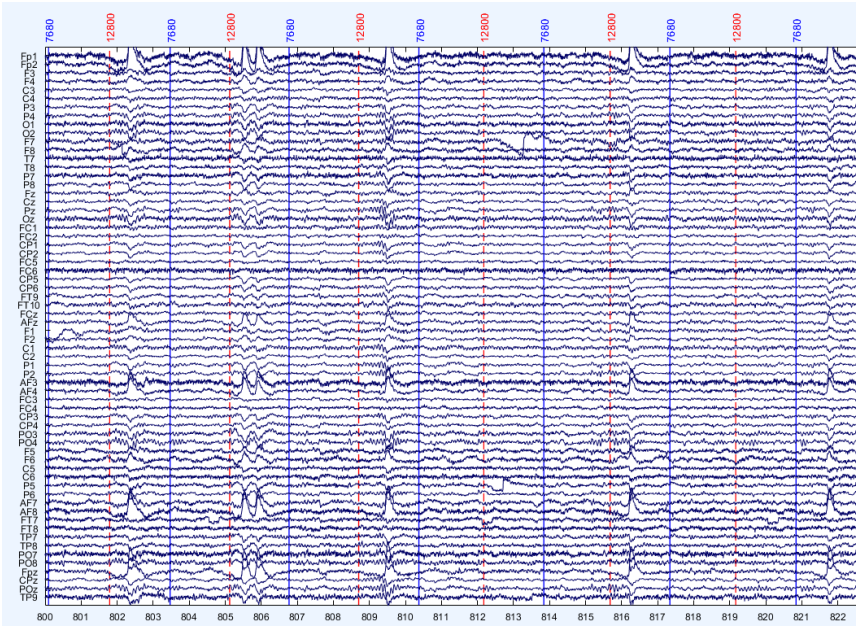

Figure 3.8: RDK task EEG data timestamped (sec) to show stimulus presentation (7680) and participant response (12800).

EOG and ECG data was collected through 8 auxiliary inputs using the Auxiliary Input Module, which was connected directly to the experiment computer. The placement of EOG and ECG electrodes is shown in Figure 3.9 and Figure 3.10 respectively. To measure EOG, two pairs of electrodes were utilized, with one pair being affixed to the participant's temples, and the other to the nasion and under the participant's left eye. A single pair of electrodes placed on the participant's chest was utilized to measure ECG. Lastly, a single electrode was placed on the participant's left clavicle for use as a shared ground. Data was collected at a sampling rate of 1000Hz and saved in the BioSemi Data Format (.BDF).

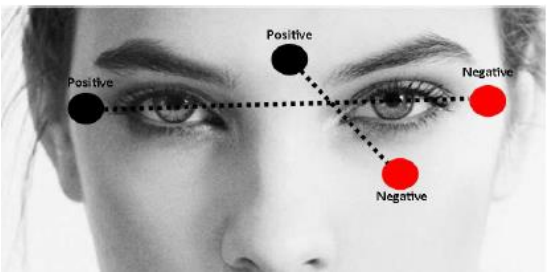

Figure 3.9: EOG electrode placement

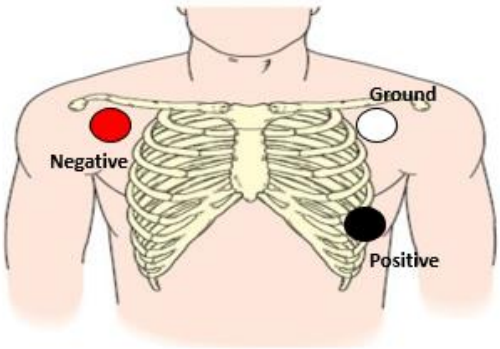

Figure 3.10: ECG electrode placement

3.5 EEG Pre-Processing

EEG data contains both oscillations generated by the brain activity of interest as well as noise introduced by a diverse range of artifacts such as eye-blinks, muscle movements and environmental noise. Preprocessing refers to any transformations or reorganizations of the data that facilitate analysis [9]. All EEG data was preprocessed using EEGlab version 14 [45] following the PREP pipeline [46]. A summary of the preprocessing pipeline is given below:

1) Data was downsampled from the collection sampling rate of 1000 Hz to 256 Hz. This was done to speed up computation as well as aid in independent component analysis (ICA) by cutting off unnecessary high-frequency information.
2) A high-pass filter at 1 Hz was applied to remove low frequency drift.

3) Channel location data was imported to allow for re-referencing.
4) A notch filter at 60 Hz was applied to suppress line noise.
5) Bad channels were removed using the Artifact Subspace Reconstruction (ASR) algorithm through the EEGLab clean_rawdata plugin [47].
6) Removed channels were interpolated using spherical interpolation to minimize potential bias when re-referencing.
7) The reference was changed from the mastoid to the channel average.
8) ICA was performed to identify independent components associated with eye-blinks.
9) Independent components associated with eye-blinks were removed based on the VEOG channel through the EEGlab icablinkmetrics plugin [48].

3.6 Analysis Strategy

The following section outlines the electrophysiological analysis and machine learning techniques used to fit the various classifiers investigated in this research, as well as the methods and metrics used to evaluate both classifier performance and feature saliency.

3.6.1 Event Related Potential Analysis

To determine if the ERP components discussed in chapter 2 could be used to distinguish between the confident and unconfident experimental conditions and associate this ability with specific regions of the brain, a statistical analysis of the ERPs for each participant was conducted. Because the EEG data is sampled at multiple time points for each of the 64 channels, statistical analysis of the ERPs is a multiple comparisons problem (MCP). That is, the statistical analysis involves simultaneous statistical tests at each

(channel, time) pair. When dealing with an MCP, the family-wise error rate (FWER) must be controlled. The FWER is the probability under the null hypothesis of no difference between the confident and unconfident conditions of falsely concluding that there is a difference at one or more (channel, time) pairs. As the number of statistical tests increases, so does the FWER. Consider the case of 30 hypotheses to test at a significance level of α = 0.05. The probability of observing at least one significant result due to chance is

$$\begin{aligned} P(\text{at least one significant result}) &= 1 - P(\text{no significant results}) \\ &= 1 - (1 - 0.05)^{30} \\ &\approx 0.79 \end{aligned}$$

Thus, there is a 79% chance of observing at least one significant result in 30 hypothesis tests even if all of tests are not actually significant. In the case of ERP analysis, the number of tests is on the order of several thousand and so the probability of observing at least one significant result due to chance is close to 100%. Methods for controlling FWER such as the Bonferroni correction [9] often involve adjusting α in such a way that the probability of observing at least one significant result due to chance is below the desired level of significance. However, with an extremely large sample size, these methods result in a statistical test that is too conservative. For this research, a more sensitive nonparametric method developed by Maris and Oostenveld [56] is utilized . A summary of the nonparametric method is given below:

1) For every (channel, time) pair, compare the confident and unconfident ERPs by means of a t-value.
2) Select all (channel, time) pairs whose t-value is larger than the 95th quantile of the Student's t-distribution.

3) Cluster the selected (channel, time) pairs on the basis of temporal and spatial adjacency.
4) Calculate cluster-level statistics by taking the sum of the t-values within a cluster.
5) Take the largest of the cluster level statistics.

3.6.2 Machine Learning

The objective of inferring participant decision confidence was formulated as a binary classification problem where the two classes were whether the participant was confident or not confident and where the distribution of the two classes was imbalanced as shown in Table 3.6. All analysis was conducted using python and its associated statistical packages. Machine learning model development was done using Scikit-learn, TensorFlow and Keras frameworks.

Table 3.6: Class Distribution of Observations

Participant	Task	Confident Observations	Not Confident Observations	Percent Confident
2863	RDK	238	202	54.1
2863	CIAT	87	33	72.5
3233	RDK	250	190	56.8
3233	CIAT	91	29	75.8
4318	RDK	297	143	67.5
4318	CIAT	86	34	71.7
4524	RDK	231	209	52.5
4524	CIAT	73	47	60.8
7984	RDK	393	47	89.3
7984	CIAT	84	36	70.0
8079	RDK	373	67	84.7
8079	CIAT	97	23	80.8
8477	RDK	304	136	69.9
8477	CIAT	87	33	72.5
9658	RDK	183	257	41.5
9658	CIAT	99	21	82.5

3.6.2.1 Feature Extraction and Data Segmentation

Before a classification model can be fit, the raw data must be transformed into an appropriate format. For traditional machine learning models, domain specific knowledge is used to manually create features from the data. In the case of deep learning models, the data must be transformed into a format the model expects, such as a sequence for RNNs or an image for CNNs. The following is a description of the process used to transform the RDK and CIAT task data into formats suitable for classification.

3.6.2.1.1 Frequency Domain Features

The raw EEG data for the RDK task was segmented into epochs spanning from -1s to 2s relative to stimulus onset. Each epoch was visually inspected and epochs containing large amounts of noise relative to the entire dataset were rejected. An example of an epoch that was rejected is given in Figure 3.11. No more than 2 percent of trials were rejected for the RDK task per participant and no more than 3 percent of trials were rejected for the CIAT task. The data from each epoch was then transformed into features in the five traditional EEG bands by taking the data from each channel and convolving with a family of complex Morlet wavelets spanning 30 frequencies over the logspace from 3 to 50 Hz. The time range for each wavelet was from -1s to 1s and the number of cycles in each wavelet increased logarithmically from 3 to 10 cycles in conjunction with the frequencies. The mean power in each band was obtained by squaring the absolute value of the mean of the resulting complex time series over the epoch. This produced up to 320 features for each trial and up to 440 observations per participant.

The raw EEG data for the CIAT task was segmented into epochs spanning from -1s to 5s relative to the appearance of a decision prompt. Epochs were inspected and rejected in

the same manner as the RDK task. The data from each epoch was transformed to the frequency domain through the wavelet transform using the same parameters as in the RDK task. However, to increase the number of samples, mean power in each band was computed using a 3s window with an overlap of 1.5s resulting in up to 320 features and up to 360 observations per participant.

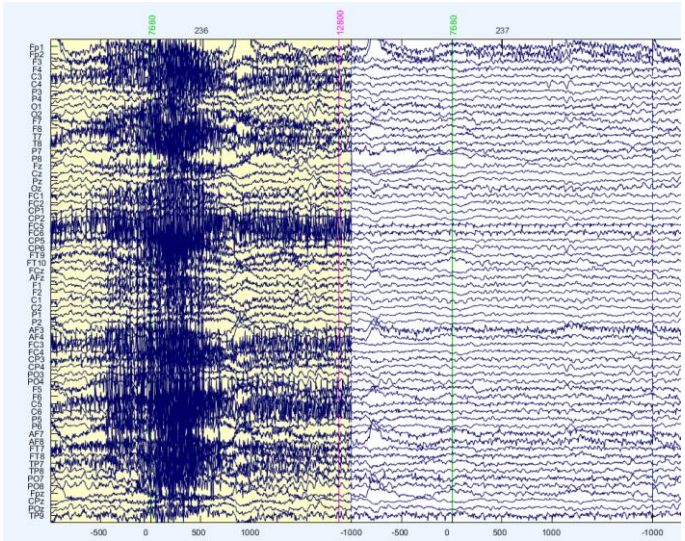

Figure 3.11: Visually Rejected Epoch

3.6.2.1.2 Time Domain Features

The raw EEG data for both tasks were segmented into epochs in the same manner as the frequency domain feature engineering process. Each epoch was split into 1-second windows with no overlap. This resulted in 64 features and approximately 1,320 observations per participant for the RDK task and 64 features and 720 observations for the CIAT task.

3.6.2.2 Classification Models

Several methods of classification were investigated for inferring participant decision confidence that generate both linear and non-linear decision boundaries. Based on their success in inferring cognitive activities other than decision confidence, logistic regression, LDA, random forest, fully-connected ANN, and convolutional-recurrent neural network (CRNN) classifiers were fit using the EEG data for both the RDK and CIAT tasks. Prior to fitting the classifiers, observations were randomly divided into a test set (30%) and non-test set (70%) for training and validation. Observations were stratified such that the class imbalance was the same across the training validation and test sets.

3.6.2.2.1 Logistic Regression and Linear Discriminant Analysis

Logistic regression and LDA classifiers were the first type of classifiers used to fit models using the EEG data. In the case of logistic regression, since the class distribution for both tasks was imbalanced, the standard log-likelihood equation was replaced by a weighted one where the weights were inversely proportional to the class frequency [49]. For both types of classifiers, the best features were selected using recursive feature elimination (RFE) in which features were selected by recursively considering smaller and smaller sets of features based on Mathew's Correlation Coefficient (MCC) computed using 5-fold cross validation. An example in which RFE was used to select 26 features is shown in Figure 3.12.

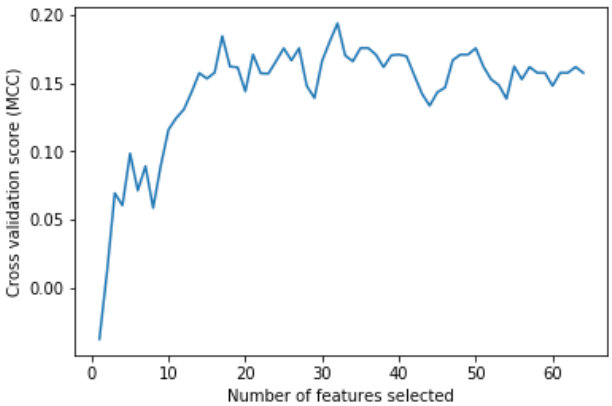

Figure 3.12: Recursive Feature Elimination with Cross-Validation

3.6.2.2.2 Random Forest Classification

A random forest classifier was trained for each task on all 320 EEG features and the best parameters were chosen based on the average MCC computed using 5-fold cross validation. The number of features considered when looking for the best split was varied from 1 to 30. For the number of trees, every integer from 1 to 500 was investigated. Maximum depth of an individual tree was varied from 1 to 20. The importance of each feature was determined as the total decrease in node impurity averaged over all trees in the ensemble [50].

3.6.2.2.3 Fully-Connected Neural Network

The first of the ANN models that were implemented was a simple fully-connected network. Hyperparameter values explored include 1, 2, and 3 fully connected layers with a ReLu activation function and 32, 64, 128, 256, and 512 hidden nodes per layer. All models used the binary cross-entropy loss function and Adam optimizer. Learning rate was tuned by exploring negative powers of 10 from 0.01 to 0.000001 with a decay of 0.000001. Model selection was done using validation-based early stopping with a patience

of 10 epochs and a delta of 0.00001. A dropout of 20% on the inputs was used for regularization. Networks were trained using mini-batch gradient descent with a batch size of 32 observations. The architecture that resulted in the best performance among all participants is shown as an example in Figure 3.13.

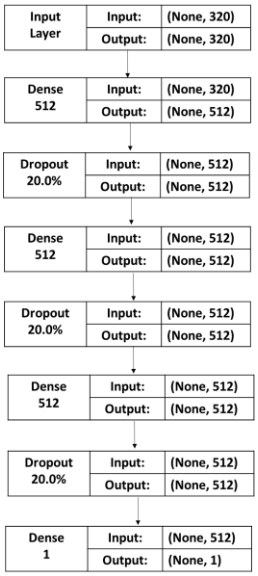

Figure 3.13: Example Fully connected Neural Network Architecture

3.6.2.2.4 Convolutional-Recurrent Neural Network

In contrast to the other classifiers which were fit using frequency transformed EEG data, a convolutional-recurrent architecture was fit for each task using the time-series EEG data described in section 3.6.1.1.1. This architecture consists of three components: The 1D convolutional layers exploit the local patterns in the temporal domain which are then used as inputs to the LSTM layer to account for the temporal dependencies between the frames. The final component is a fully-connected layer that predicts the participant's confidence. Hyperparameter tuning consisted of varying the number of layers, number of output filters, and kernel width in the convolutional component, the number of hidden units in the LSTM component and the learning rate. Hyperparameter values explored in

the convolutional component include 2, 3, and 4 layers, 32, 64, and 128 output filters, and kernel widths of 5 and 10. The number of hidden units in the LSTM layer was tuned by exploring powers of 2 ranging from 32 to 512. Learning rate was tuned by exploring negative powers of 10 from 0.01 to 0.000001 with a decay of 0.000001. RMSprop was used as the optimizer due to being well-suited to handling non-stationary environments [51]. All models used a binary cross-entropy loss function. Batch normalization and a dropout of 25% on the inputs were used for regularization. Selection was done using validation-based early stopping with a patience of 10 epochs and a delta of 0.00001. Networks were trained using mini-batch gradient descent with a batch size of 32 observations. The architecture that performed best among all participants is given in Figure 3.14 as an example.

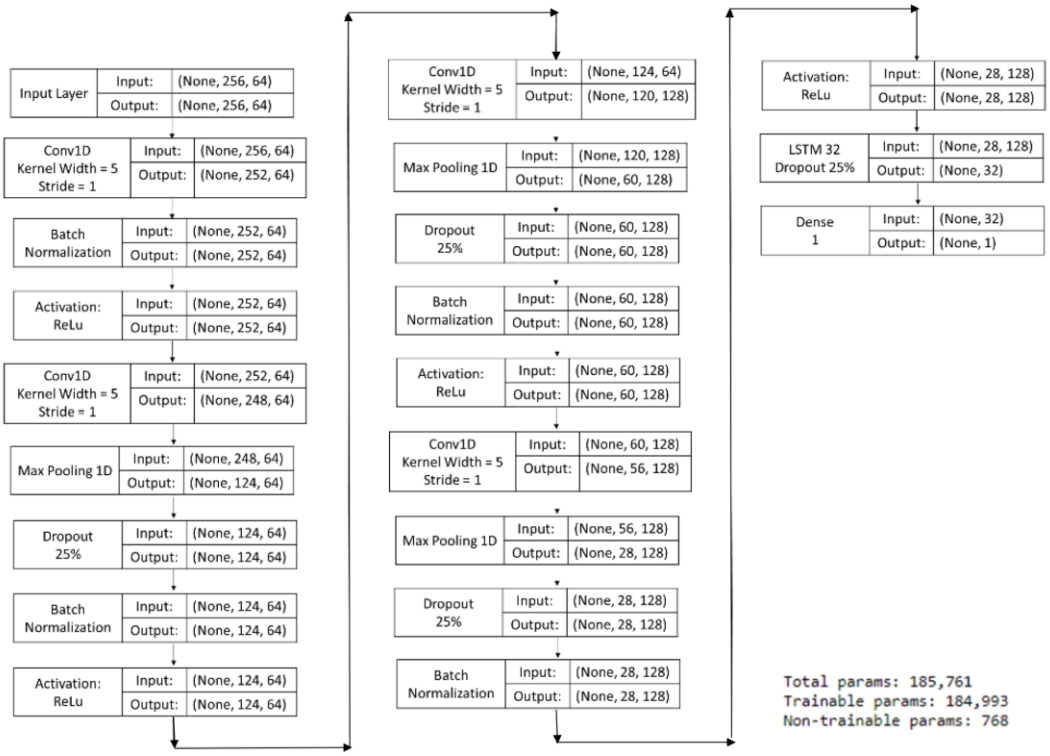

Figure 3.14: Example Convolutional-Recurrent Architecture

3.6.2.3 Performance Metrics

Many metrics exist to evaluate the performance of classification models. However, the usefulness of a metric varies with the classification problem being solved. It is also unlikely that a single metric can completely describe all facets of a classifier's performance. Thus, the following performance metrics were chosen to best capture the performance of the classifiers discussed in the previous section.

3.6.2.3.1 Confusion Matrix

A confusion matrix displays information regarding the actual class labels versus the predictions made by a classifier and provides additional insight into the misclassifications that were made. Figure 3.15 provides an example confusion matrix.

	Predicted Confident	Predicted Not Confident
Actual Confident	True Positive	False Negative
Actual Not Confident	False Positive	True Negative

Figure 3.15: Confusion Matrix

For a binary classifier, one class is labelled as the positive and the other as the negative. Using this notation, a True Positive (TP) occurs when both the predicted and actual class are the "confident" class, a False Positive (FP) occurs when the predicted class is the "confident class" and the actual class is the "not confident class", a False Negative (FN) occurs when the predicted class is the "not confident" and the actual class is the "confident class", and a True Negative (TN) occurs when both the predicted class and actual class are the "not confident class". The following metrics can be computed from the entries of a confusion matrix:

- Balanced Accuracy (BACC): The conventional accuracy can be a misleading metric if the distribution of observations over classes is imbalanced. Since this is the case for the datasets of this study, accuracy is not used and is instead replaced with balanced accuracy. BACC addresses the issue of falsely suggesting above-chance generalizability by reducing to the conventional accuracy when the classifier performs equally well on either class, but drops to chance if the conventional accuracy is high only due to the classifier taking advantage of the imbalanced data [52]. BACC is given by the equation

$$\mathrm{BACC} = \frac{\frac{\mathrm{TP}}{\mathrm{P}} + \frac{\mathrm{TN}}{\mathrm{N}}}{2}$$

- Recall: The proportion of predictions in which the classifier correctly classifies the "confident" class relative to the total number "confident" class observations. Recall is given by the equation

$$\mathrm{Recall} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}$$

- False Positive Rate (FPR): The proportion of predictions in which the classifier incorrectly classifies the "not confident" class as the "confident" class relative to the total number "not confident" class observations. FPR is given by the equation

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{FP} + \mathrm{TN}}$$

- Specificity: The proportion of predictions in which the classifier correctly classifies the "not confident" class relative to the total number of "not confident" class observations. Specificity is given by the equation

$$\text{Specificity} = \frac{\text{TN}}{\text{FP} + \text{TN}}$$

- Precision: The proportion in which the classifier correctly classifies the "confident" class relative to the total number of "confident" class predictions. Precision is given by the equation

$$\text{Precision} = \frac{\text{TP}}{\text{TP} + \text{FP}}$$

- Negative predictive value (NPV): The proportion in which the classifier correctly classifies the "not confident" class relative to the total number of "not confident" class predictions. NPV is given by the equation

$$\text{NPV} = \frac{\text{TN}}{\text{TN} + \text{FN}}$$

- Matthews Correlation Coefficient (MCC): MCC is a robust measure of the quality of the classifier when it is trained and evaluated on an imbalanced dataset and can be interpreted as a measure of correlation between the actual and predicted classes. MCC can take any value from -1 to 1 where values greater than or equal to 0.4 indicate good agreement between the observed and predicted class labels. MCC is given by the equation

$$\text{MCC} = \frac{TP(TN) - FP(FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}}$$

3.6.2.3.2 Receiver Operating Characteristic Curve

The Receiver Operating Characteristic (ROC) curve is a graph showing the performance of a classifier at all classification thresholds. The ROC plots the TPR given versus the FPR. A typical ROC curve is shown in Figure 3.16.

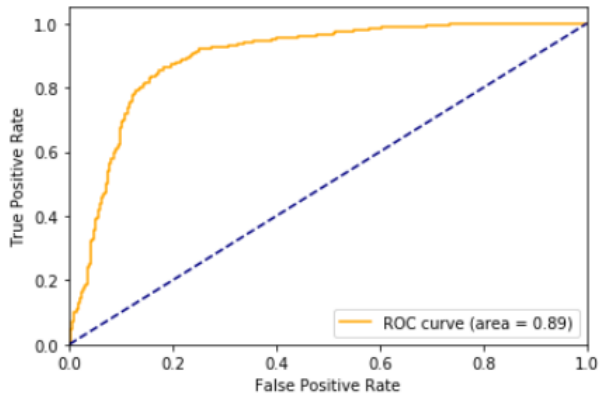

Figure 3.16: Receiver Operating Characteristic Curve

A common metric used for evaluating the performance of a classifier is computing the area under the ROC (AUC). This provides an aggregate measure over all classification thresholds. The AUC falls between 0 and 1. An AUC of 1 corresponds to a classifier with perfect predictions, while an AUC of 0.5 corresponds to a classifier performing no better than random chance.

3.6.2.3.3 Cohen's Kappa

A measure of how much homogeneity or consensus there is between the labeled data and the classifier that considers the probability of random agreement according to the frequency of each class. Cohen's Kappa can take any value from -1 to 1 and is interpreted as follows: values ≤ 0 indicate performance no better than random chance, 0.01-0.20 as slight, 0.21-0.40 as fair, 0.41-0.60 as moderate, 0.61-0.80 as substantial, and 0.81-1.00 as almost perfect to perfect [53]. Cohen's Kappa is given by the equation

$$\kappa = 1 - \frac{1 - p_o}{1 - p_e}$$

3.6.2.3.4 Binary Cross Entropy

A measure of how much extra information is required to derive the actual class labels from the predicted class labels. Binary cross entropy is given by the equation

$$\sum_i -y_i \log(\hat{y}_i) - (1 - y_i)\log(1 - y_i)$$

3.6.3 Behavioral Analysis

To investigate the main effects of the independent variables and the importance and strength of association between the dependent variables in the CIAT task, the theoretical model shown in Figure 3.17 and set of null hypotheses given in Table 3.7 were formulated.

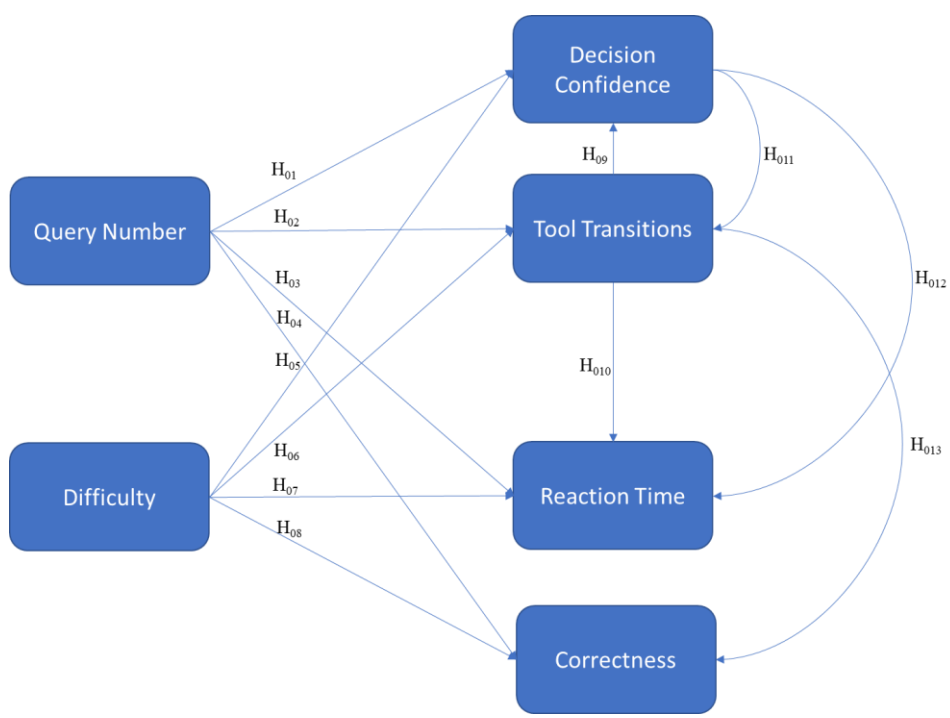

Figure 3.17: Theoretical Model for Participant Behaviors

Table 3.7: Set of Testable Hypothesis

	Hypothesis
H_{01}:	The query number does not have an effect on participant decision confidence
H_{02}:	The alert difficulty does not have an effect on participant decision confidence
H_{03}:	The query number does not have an effect on participant tool transitions
H_{04}:	The alert difficulty does not have an effect on participant tool transitions
H_{05}:	The query number does not have an effect on participant reaction time
H_{06}:	The alert difficulty does not have an effect on participant reaction time
H_{07}:	The query number does not have an effect on participant correctness
H_{08}:	The alert difficulty does not have an effect on participant correctness
H_{09}:	Participant tool transitions do not have an effect on decision confidence
H_{010}:	Participant tool transitions do not have an effect on decision confidence
H_{011}:	Participant decision confidence does not have an effect on tool transitions
H_{012}:	Participant decision confidence does not have an effect on reaction time
H_{013}:	Participant tool transitions do not have an effect on correctness

To test the hypotheses of Table 3.7, Generalized Linear Mixed Models (GLMMs) were fit for each dependent variable. The GLMM is an extension of the Generalized Linear Model (GLM) which is the unification of both linear and nonlinear regression models which allows for response variables from nonnormal distributions [54]. The GLMM extends the GLM by including both fixed and random effects. The inclusion of random effects allows for the control of non-independence in the data being analyzed. In the case of the CIAT experiment, observations at the participant level are not independent as there are individual differences between participants which may have influenced their behaviors. Statistical inference on model parameters is done using either the likelihood ratio test or Wald inference [54]. Model fitting and statistical inference was done using the Statsmodels API [55].

3.7 Summary

In summary, this chapter explained the methodology that was used for data collection and analysis of decision confidence in both simple and complex decisions. First, the

experimental design for two experiments were explored in detail. The experiments consisted of an RDK task in which participants judged the global motion of dots in an aperture and a simulated cyber investigation in which participants evaluated the validity of machine-generated alerts. Next, details on the setup and procedures for collecting electrophysiological and behavioral data were presented. This was followed by a description the preprocessing and segmentation used to create datasets for analysis. Finally, the chapter concluded with formulating the problem of inferring decision confidence as a binary classification problem and the techniques used for evaluating classifier performance as well as the statistical analysis of the behavioral and ERP data.

IV. Analysis and Results

4.1 Chapter Overview

This chapter provides an in-depth look at the data exploration process and analysis of the results obtained from both the RDK and CIAT tasks. The chapter is divided into two major sections. The first section covers the results and analysis of the electrophysiological data collected from the RDK task. This includes the results of the ERP analysis and a performance evaluation of machine learning models fit using the electrophysiological data. The results in this section serve to answer if electrophysiological features can be used in combination with machine learning techniques to infer decision confidence in a simple decision with a performance greater than chance and what those salient features are. The second section covers the results and analysis of both the behavioral and electrophysiological data collected during the CIAT task. First, the results of the behavioral data exploration are presented. This is followed by a performance evaluation of machine learning models fit using the behavioral data. Finally, a performance evaluation of machine learning models fit using the electrophysiological data as well as the results of the ERP analysis are presented. The results in this section serve to answer if behavioral features can be used in combination with machine learning techniques to infer decision confidence in a complex decision with a performance greater than chance and if the salient electrophysiological features for inferring decision confidence are the same for both simple and complex decisions.

4.2 Random Dot Kinematogram Task

4.2.1 Event Related Potentials

Statistically significant differences (cluster corrected p-value < 0.05) between ERPs corresponding to the confident and unconfident conditions were observed in frontal, central and parietal electrodes for two of the eight participants. Table 4.1 lists the electrodes and corresponding latencies at which the differences were observed. ERPs corresponding to the confident and unconfident conditions for electrode FC1 for participant 4524 and electrode C2 for participant 7984 are shown in Figure 4.1. The significant differences are highlighted in yellow. For participant 4524, a difference in negative voltages occurring on average from 791 ms to 979 ms after stimulus onset was observed in four frontal and four central electrodes. There are no known ERP components that match this description. For all eight electrodes during the time period of significance it was observed that the voltage for the confident condition was more negative than for the unconfident condition. For participant 7984, a difference in positive voltages occurring on average from 625 ms to 799 ms after stimulus onset was observed in six central and two parietal electrodes. This is likely the P300 component which peaks approximately 300 ms to 800 ms post stimulus onset. For all eight electrodes during the time period of significance it was observed that the voltage for the confident condition was more positive than for the unconfident condition. This observation is consistent with results presented by Kerkhof [12].

Table 4.1: Electrodes and Latencies of Observed Differences in ERPs

Participant 4524		Participant 7984	
Electrode	**Latency (ms)**	**Electrode**	**Latency (ms)**
F1	[785 1001]	C2	[625 820]
F2	[785 969]	CP4	[625 780]
F3	[781 957]	CPz	[625 820]
F4	[800 950]	CZ	[625 820]
FC1	[800 1002]	FC2	[625 742]
FC2	[797 1000]	FC4	[628 820]
FC3	[780 950]	FC6	[625 780]
FCZ	[800 1000]	FCz	[625 813]

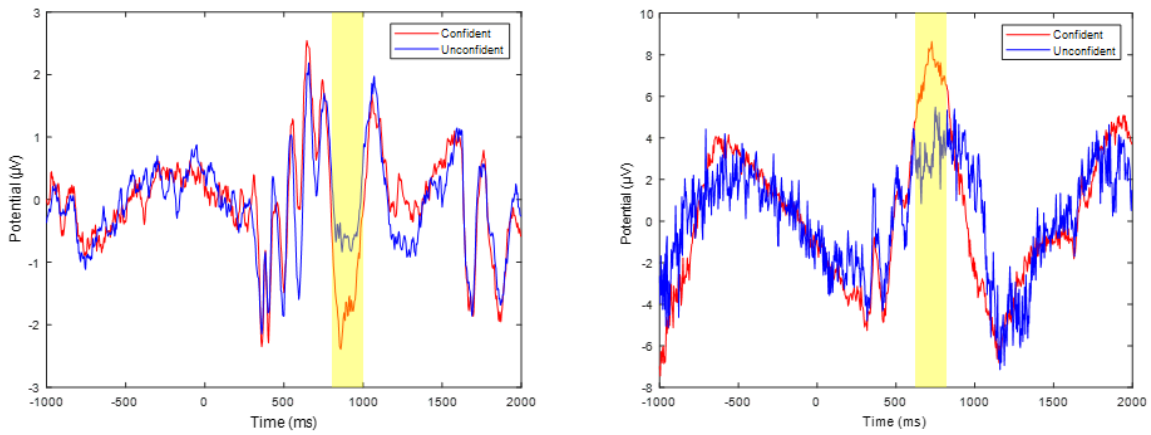

Figure 4.1: Example ERPs for Participant 4524 (Left) and Participant 7984 (Right)

4.2.2 Classification of Confidence

To evaluate classifier performance and determine the best machine learning model for classifying decision confidence for the RDK task, LR, LDA, RF, and fully connected ANN models were trained and tested using the mean power features from each of the five traditional EEG bands for each of the eight participants. This resulted in a total of 32 models that were evaluated and compared. Model performance was evaluated using four metrics: BACC, AUC, MCC, and Cohen's Kappa. For each participant, the model in which three of the four performance metrics were highest is reported as the model with the

best performance. Mean results across participants are given in Table 4.2. Each performance metric indicates performance fairly greater than random chance.

Table 4.2: Mean Performance of Frequency Band Models for the RDK Task

Metric	**Mean**	**95% Lower CL**	**95% Upper CL**
BACC	0.704	0.659	0.749
AUC	0.697	0.660	0.734
MCC	0.399	0.299	0.493
Cohen's Kappa	0.386	0.283	0.489

The best performing model and the corresponding performance metrics for each participant are shown in Figures 4.2 through 4.5. The best performing model for each participant exceeded the random chance value of 0.5 for BACC and AUC and the best performing model for seven of the eight participants exceeded the random chance value of 0 for MCC and Cohen's kappa. Models fit using a fully connected ANN were consistently the best across participants, providing the best performance for seven participants, only performing worse than the random forest model for a single participant (4524). The highest BACC among the best performing models was 0.753, 95% CI [0.708, 0.798], which was associated with participant 8477's fully connected ANN. The highest AUC was 0.782, 95% CI [0.782, 0.819] which was associated with participant 9658's fully connected ANN. The lowest BACC and AUC were 0.586, 95% CI [0.541, 0.631] and 0.632, 95% CI[0.595, 0.669] respectiviely which were both associated with participant 7984's fully connected ANN. The highest MCC and Cohen's kappa among the best performing models was 0.514, 95% CI [0.417, 0.611] and 0.507, 95% CI [0.404, 0.610] respectively. Both metrics were associated with participant 4318's fully connected ANN and indicate fair to moderate performance when compared to random chance. The lowest MCC and Cohen's kappa were 0.095, 95% CI [-0.002, 0.192] and

0.056, 95% CI [-0.047, 0.159], respectively. These metrics were once again associated with participant 7984's fully connected ANN and indicate performance no better than random chance.

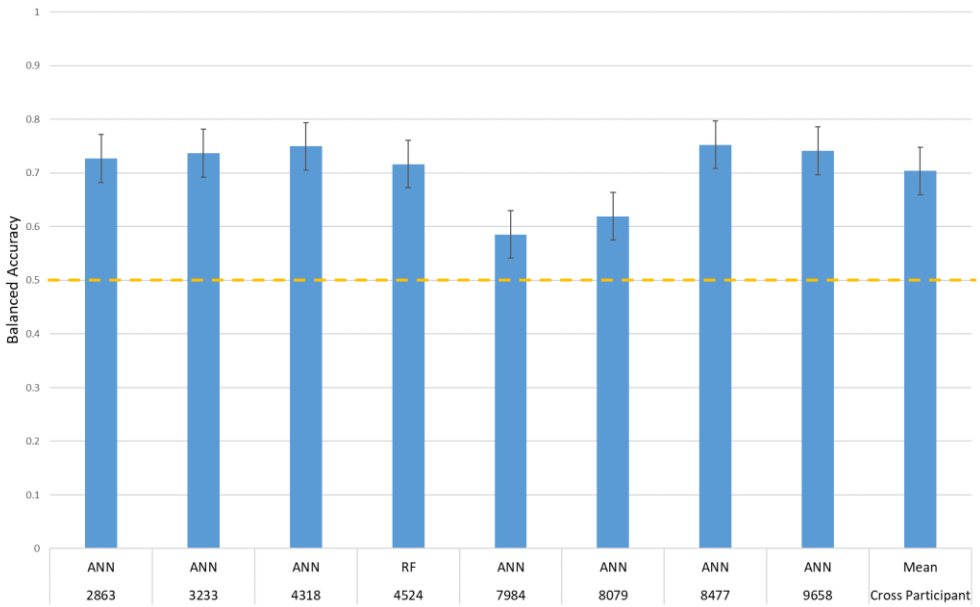

Figure 4.2: BACC for the Best Performing Models on the RDK Task

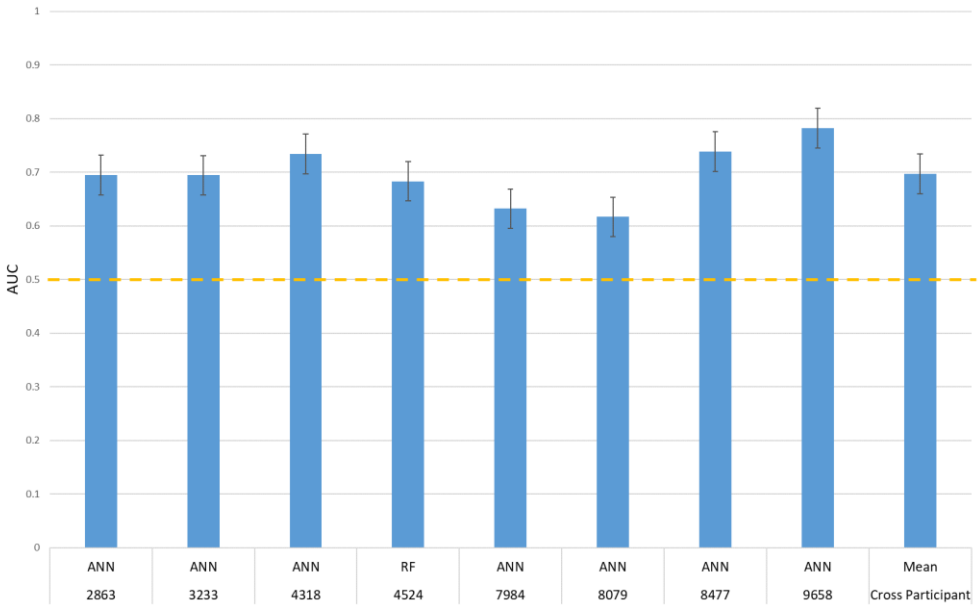

Figure 4.3: AUC for the Best Performing Models on the RDK Task

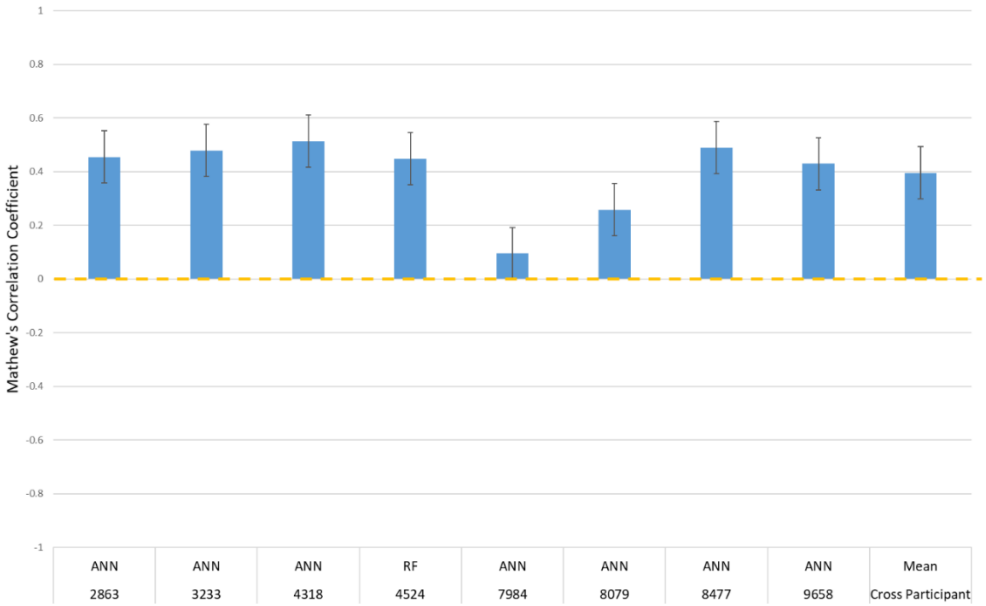

Figure 4.4: MCC for the Best Performing Models on the RDK Task

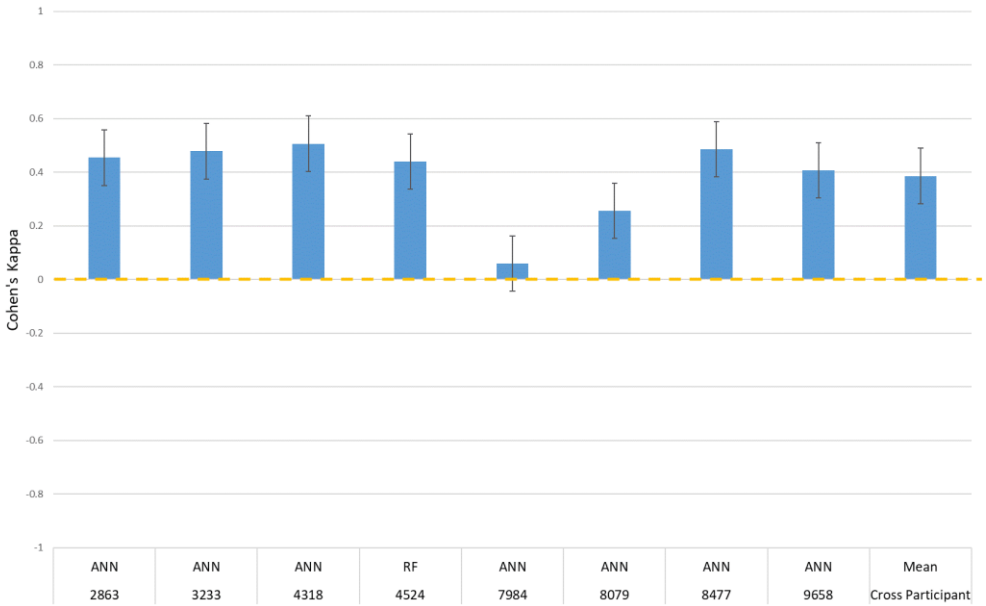

Figure 4.5: Cohen's Kappa for the Best Performing Models on the RDK Task

It is likely that the extreme class imbalance for participant 7984 shown in Table 3.6 contributed to the lower performance when compared to the best performing models of the other participants with respect to every performance metric, as it has been established that a class imbalance can have a detrimental effect on both convergence during the training phase and generalization of a model on the test set [57]. The confusion matrices shown in Figure 4.6 provide some insight as to why this is. The top matrices correspond to the best across each participant's top performing model and the bottom correspond to the worst across each participant's top performing model with respect to BACC. Test sets for both participants contain 87 observations but the distribution of classes is significantly different. The test set for participant 8477 contains 30 'Not Confident' observations, whereas the test set for participant 7984 contains only 7. If the classifier for participant 8477 had one additional "Not Confident" misclassification, recall would drop by 3 percent causing BACC to drop by 1.5 percent. However, if the classifier for participant 7984 made the same misclassification, recall would drop 14.2 percent causing BACC to drop by 7.1 percent. For this reason, class weighting was utilized to attempt to counter the class imbalance problem. However, increasing the weighting of the minority class any further would result in a proportional number of misclassifications of the majority class.

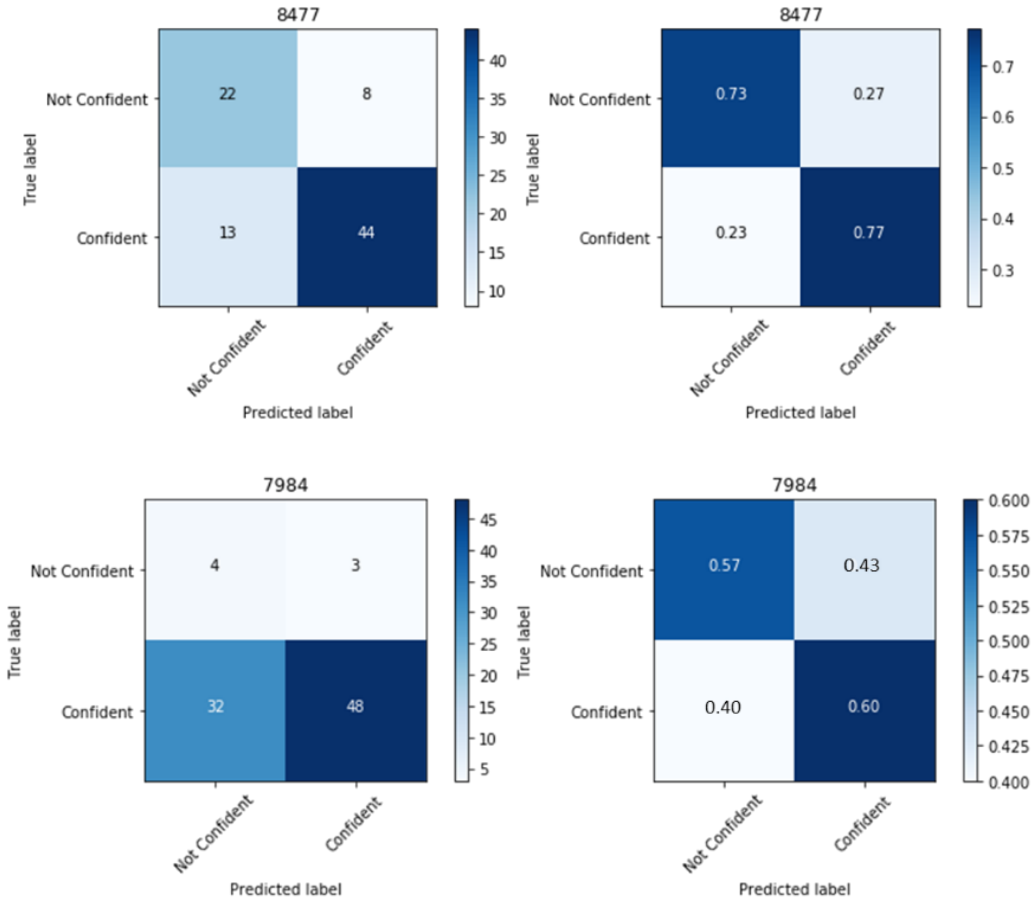

Figure 4.6: Confusion Matrices for the Best and Worst Performing Models

To determine the contribution of the individual frequency bands towards classification of decision confidence, five single frequency band models were fit for each participant and compared against a paradigm where each participant's best performing model architecture was trained and evaluated using the frequency information from all but one band. This process resulted in eighty additional models for comparison. Table 4.3 displays each of the single band models compared to the leave-one-band-out models ranked by best performance and the largest decrease in performance, respectively. For all participants, models fit using features from all five frequency bands performed better than models fit using only individual bands or by leaving out any individual band, suggesting

that all frequency bands contribute some information towards classifying decision confidence. In half the participants, the model fit using only the alpha band features resulted in the best performance and the model in which the alpha band features were excluded resulted in the largest performance drop. This suggests that the alpha band may contribute the most information, which is in-line with previous results presented by Kubanek [19], Graziano [20], and Samaha [21]. For participant 2863, the model fit using only the delta band features resulted in the best performance and the model in which the delta band features were excluded resulted in the largest performance drop. In the literature, oscillations in the delta band are typically associated with slow-wave sleep and anesthesia, when no conscious functions take place. However, more recent research has shown that the magnitude of coherent oscillations in the delta frequency band between parietal and frontal cortices is modulated by different decision alternatives and that in conditions not requiring decision making, delta band coherences are typically reduced [58]. For participant 8079, the model fit using only the theta band features resulted in the best performance and the model in which the theta band features were excluded resulted in the largest performance drop. The power of theta oscillations is thought to be correlated with several cognitive processes: left parietal theta is correlated with memory recognition, central theta is correlated with decision making, and widespread theta is correlated with memory load [59]. There was no agreement between performance of the single band model and the leave-one-band-out model for two participants. For three participants, the model fit using only the gamma band features resulted in the worst performance and the model in which the gamma band features were excluded resulted in the smallest performance drop. This suggests that the gamma band contributes the least amount of

information for classifying decision confidence. Gamma is thought to be related to the integration of information as well as attention and working memory processes, but also completely overlaps with the spectral bandwidth of muscle activity [60]. Since no muscular artifact correction methods were performed when preprocessing the EEG data, an argument could be made that the classifiers may be detecting differences in muscular artifacts associated with a decision input rather than a participant's confidence. However, the observed low feature utility of the gamma band suggests that the models are unlikely to be learning muscle movements rather than neural representations of decision confidence.

Table 4.3: Comparison of RDK Single Band Models (Column Header 1) to Leave-one-band-out (Column Header 4) Models with Respect to Highest Perfromance and Highest Perfromance Drop

	Participant															
	2863		**3233**		**4318**		**4524**		**7984**		**8079**		**8477**		**9658**	
Rank \ Bands	1	4	1	4	1	4	1	4	1	4	1	4	1	4	1	4
1	Δ	Δ	A	A	Θ	B	A	A	A	A	Θ	Θ	A	A	Γ	A
2	A	B	Δ	Θ	Δ	Θ	Δ	Δ	Γ	B	A	Δ	Δ	Δ	A	Δ
3	Θ	A	Γ	B	A	A	Θ	B	Δ	Γ	Γ	A	B	Θ	B	Θ
4	B	Θ	B	Δ	B	Δ	B	Θ	B	Δ	Δ	Γ	Θ	Γ	Δ	Γ
5	Γ	Γ	Θ	Γ	Γ	Γ	Γ	Γ	Θ	Θ	B	B	Γ	B	Θ	B

To further investigate the salient features of a simple decision and validate the rankings shown in table 4.3, feature importance was extracted using the random forest models fit on all 320 features and compared with the feature lists generated by the logistic regression and LDA models fit using RFE. Table 4.4 lists the intersection of salient features across the logistic regression, LDA, and top 15 features ranked by the random forest models for each participant. Table 4.3 and 4.4 are in general agreement with each other. For all participants, mean power features from the frequency bands associated with

the largest drops in performance are included by the logistic regression and LDA models and also in the top 15 features ranked by the random forest model. Across participants over half the important alpha band features are associated with the lower-central and parietal regions of the brain, which are regions in which the alpha band has been shown to be able to discriminate between confidence levels [19]. Similarly, over half the important delta and theta band features are also associated with the expected regions of the brain. Unfortunately, there does not appear to be any consistency across participants with respect to the specific channels selected.

Table 4.4: Intersection of Salient Features Across LR, LDA, and RF Models for the RDK Task

Participant							
2863	**3233**	**4318**	**4524**	**7984**	**8079**	**8477**	**9658**
Cz Δ	CP2 A	T8 Θ	AF3 A	C6 A	C5 Θ	Fz A	O1 Γ
C1 Δ	CPz A	FT9 Θ	FPz A	CP2 A	F4 Θ	CPz Δ	F7 A
CP2 A	C1 Δ	TP7 Δ	FP1 A	F8 Γ	P3 Θ	P1 B	FC3 B
C4 A	Cz A	F3 Θ	FP2 A	O1 Γ	CP4 Θ	C6 Δ	POz Γ
P6 Δ	P03 Θ	Cz Θ	Fz Δ	Cz Δ	FZ A		C2 Γ
Cz A	F5 Γ	PO8 Δ	FP1 Δ	O1 B	P1 Δ		FC5 B
		C6 Θ	AFz Δ		CPz Δ		PO3 Γ
		P1 A	FC1 Δ		CP2 Θ		Cz B
		C4 Θ			C6 Δ		
		FT10 Γ					

A drawback of the models fit using frequency domain information is that they require a substantial amount of preprocessing or suffer from reduced performance. Since the eventual goal is to field systems capable of inferring operator decision confidence in real-world, real time environments, the dependency of these models on preprocessing is impractical. Thus, classification using the time series information described in section 3.6.1.1.1 via a CRNN was also investigated. Each CRNN was trained using the process

described in Section 3.6.2.2.4. Models were able to achieve 100% training accuracy after approximately 100 epochs. However, models corresponding to the lowest validation loss took an average of 21.2 epochs to train. The average time per epoch was 7.6 seconds. Performance of the CRNN models is shown in Figure 4.7 to Figure 4.10. Mean results across participants are given in Table 4.5. These values indicate that across participants, the CRNN did not perform better than random chance.

Table 4.5: Mean Performance of the CRNN Models for the RDK Task

Metric	**Mean**	**95% Lower CL**	**95% Upper CL**
BACC	0.534	0.504	0.563
AUC	0.518	0.483	0.554
MCC	0.060	0.003	0.118
Cohen's Kappa	0.059	0.002	0.116

The best performing model had a BACC of 0.642, AUC of 0.628, MCC of 0.274, and Cohen's kappa of 0.271, which were the highest values for each metric across participants. This model was associated with participant 8477, who also had the overall best performing model fit using frequency band information, and was the only model to perform at a level above random chance. The worst BACC, MCC, and Cohen's kappa were 0.501, 0.001, and 0.001 respectively and were associated with participant 9658. The worst AUC was 0.457 and was associated with participant 8097. For all participants the CRNN model performed substantially worse than their best performing model fit using frequency band information, with an average decrease in performance of 0.170 for BACC, 0.179 for AUC, 0.336 for MCC, and 0.327 for Cohen's kappa. Analysis of the residuals did not reveal any patterns of misclassification other than the tendency to predict the confident class for the majority of observations. A more thorough hyperparameter search

may improve performance. However, it is more likely that number of samples available for training isn't large enough for the network to learn anything meaningful.

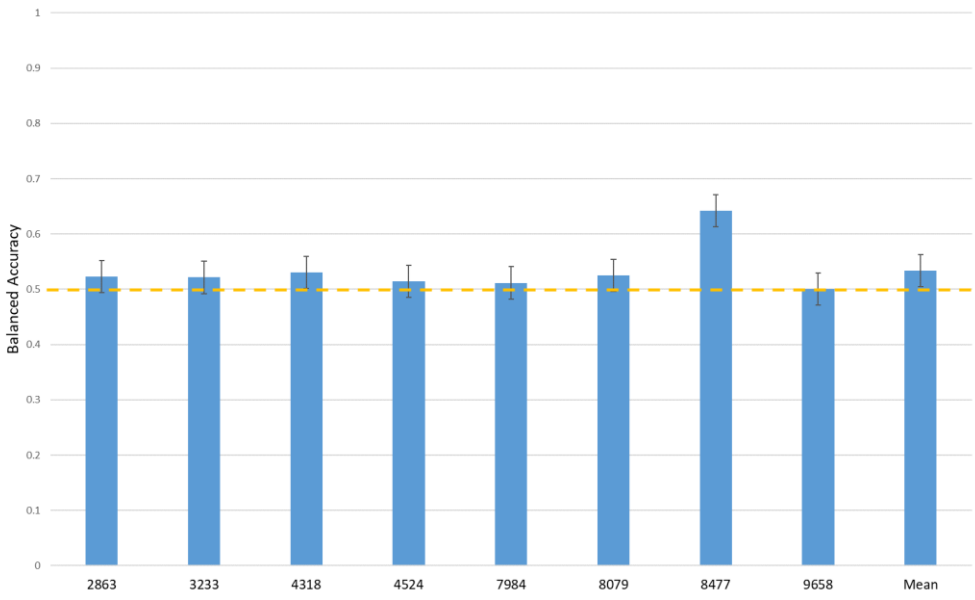

Figure 4.7: BACC for the CRNN fit on the RDK Task Data

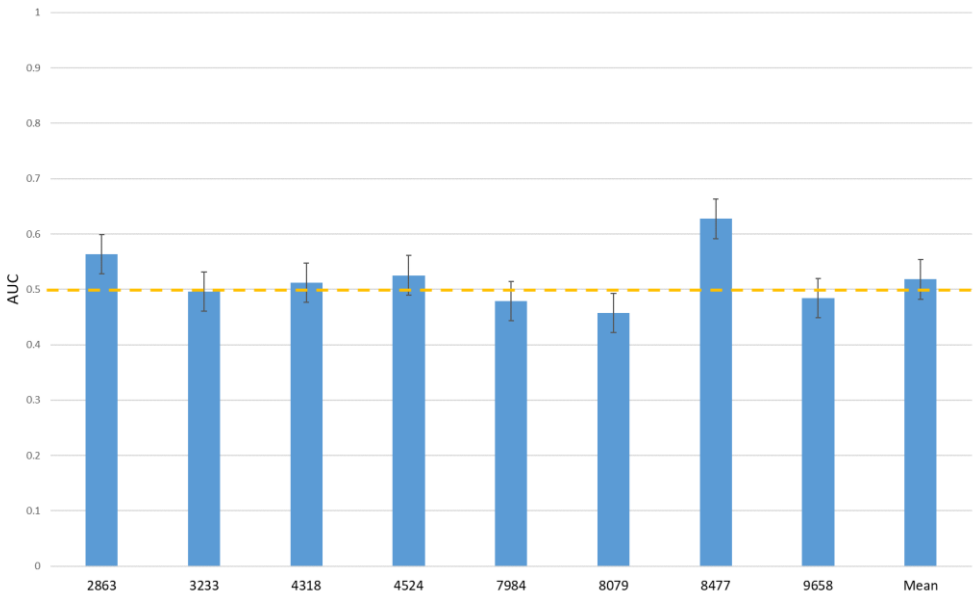

Figure 4.8: AUC for the CRNN fit on the RDK Task Data

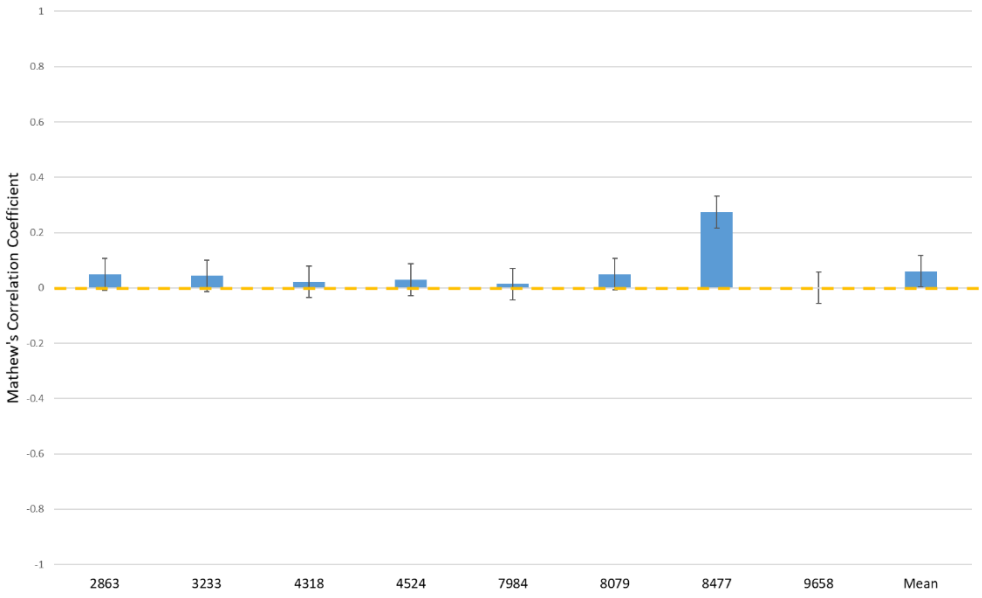

Figure 4.9: MCC for the CRNN fit on the RDK Task Data

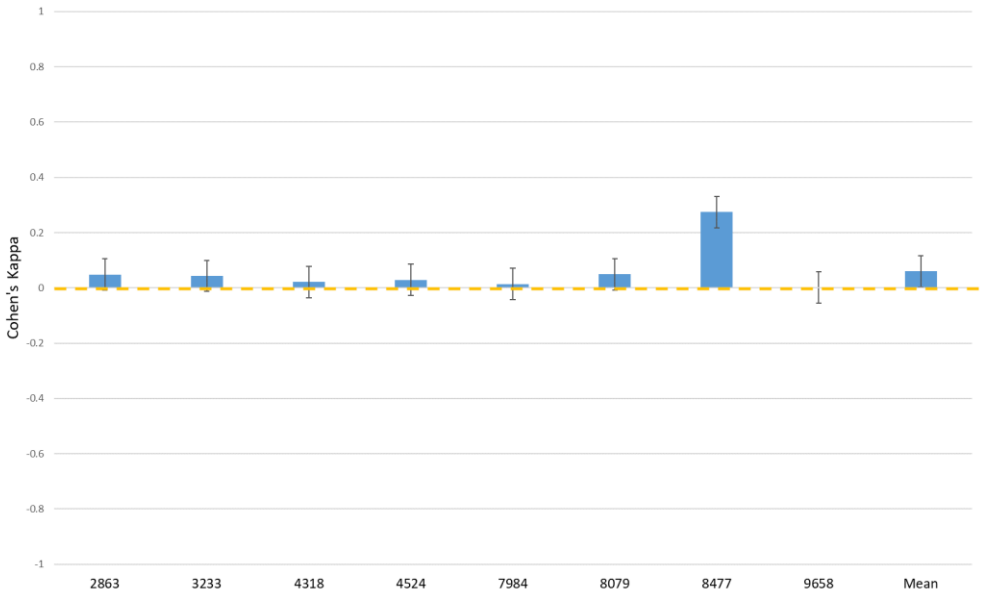

Figure 4.10: Cohen's Kappa for the CRNN fit on the RDK Task Data

4.3 Cyber Intruder Alert Testbed Experiment Analysis

4.3.1 Behavior Results and Analysis

Initial exploration of the behavioral data began with investigating the descriptive statistics given by Table 4.6. The distribution of difficulty across alerts was chosen so that a roughly equal number of confident and unconfident responses were obtained. However, participants were often more confident (73.33% of responses) than they were unconfident (26.67% of responses). Participants were also more confident than they were correct, being correct only 48.12% of the time. Reaction times for confident responses were slightly longer than for unconfident responses with a mean difference of 61.96ms. Similarly, reaction times for correct responses were also slightly longer than for incorrect responses with a mean difference of 34.52ms. The number of tool transitions was slightly less for confident responses than for unconfident responses with a mean difference of 1 transition, while tool transitions were roughly the same for correct and incorrect responses. Additionally, once a participant became confident, they typically did not lose confidence in a later decision as this occurred in only 2% of decisions across participants. Similarly, there was only one instance in which a participant changed their answer with respect to the Threat versus False alarm alternatives.

Table 4.6: Descriptive Statistics for the CIAT Behavioral Data

	Participant								
Statistic	2863	3233	4318	4524	7984	8079	8477	9658	Cross-Participant
Percent Confident	72.50	75.80	71.70	60.80	70.00	80.80	72.50	82.50	73.33
Percent Unconfident	27.50	24.20	28.30	39.20	30.00	19.20	27.50	17.50	26.67
Percent Correct	55.00	55.00	40.00	47.50	47.50	45.80	42.50	51.70	48.12
Percent Incorrect	45.00	45.00	60.00	52.50	52.50	54.20	57.50	48.30	51.88
Mean Reaction Time Confident	778.18	846.97	954.74	835.59	979.55	957.21	1129.30	1057.77	942.41
Std Dev. Reaction Time (ms) Confident	264.61	248.86	341.42	213.14	249.60	227.54	305.96	259.36	263.81
Mean Reaction Time (ms) Unconfident	744.27	808.28	1001.09	846.77	892.67	925.22	1022.06	803.24	880.45
Std Dev. Reaction Time (ms) Unconfident	167.86	194.50	406.83	218.90	200.40	338.23	305.72	163.62	249.51
Mean Reaction Time (ms) Correct	773.41	852.52	1002.17	849.81	969.86	914.10	1151.71	1051.90	945.68
Std Dev. Reaction Time (ms) Correct	285.25	256.32	372.72	193.26	239.15	338.37	294.60	229.40	276.13
Mean Reaction Time (ms) Incorrect	763.30	819.41	945.01	831.06	938.67	958.47	1061.45	971.88	911.16
Std Dev. Reaction Time (ms) Incorrect	176.16	210.68	352.41	233.42	238.38	231.29	314.81	290.35	255.94
Mean Tool Transitions Confident	1.10	1.47	2.12	1.96	1.96	2.58	1.64	2.33	1.90
Std Dev. Tool Transitions Confident	1.68	1.58	2.30	1.86	1.38	2.66	1.81	2.29	1.95
Mean Tool Transitions Unconfident	2.61	2.21	2.91	2.19	1.89	5.43	2.48	3.48	2.90
Std Dev. Tool Transitions Unconfident	1.69	1.32	1.93	1.78	1.15	2.02	1.10	2.04	1.63
Mean Tool Transitions Correct	1.03	1.42	2.10	1.89	1.91	5.25	1.73	2.35	2.21
Std Dev. Tool Transitions Correct	1.59	1.57	2.28	1.83	1.31	1.95	1.95	2.28	1.84
Mean Tool Transitions Incorrect	2.11	1.93	2.50	2.19	1.97	2.70	1.99	2.72	2.26
Std Dev. Tool Transitions Incorrect	1.89	1.49	2.19	1.82	1.32	2.74	1.46	2.28	1.90

4.3.1.1 Decision Confidence Modelling

To determine whether the query number, difficulty, and number of tool transitions had an effect on decision confidence, the data was explored using several visualization techniques and then used to fit a GLMM to test for the significance of the predictors. Figure 4.11 and 4.12 display violin plots of the number of confident observations versus the query number and versus difficulty respectively. A violin plot combines the box plot and density trace into a single diagram by plotting the density trace symmetrically to the left and right of the box plot [61]. The box plot portion of the diagram displays information about the distribution of the data based on five values: minimum, first quartile, median, third quartile, and maximum. The central rectangle spans the first

quartile to the third quartile known as the interquartile range. The circle inside the rectangle shows the median and the "whiskers" above and below the rectangle show the locations of the minimum and maximum. The density trace supplements the box plot by showing the distribution shape of the data. Figure 4.11 strongly suggests that query number had an effect on confidence. The spread of the interquartile range corresponding to query 1 is the largest and does not overlap with the spread for any other query, indicating that there is a difference between query 1 and the other queries with respect to confidence. The distribution of confidence for query 1 appears uniform with a median of 3 confident observations. Compared to the violin plots of the other queries, the median for query 1 is significantly lower, indicating that more participants were unconfident at the first query than for the later queries. The interquartile ranges for queries 2, 3, and 4 all overlap. However, the medians of each of these queries do not overlap with the interquartile ranges of any other and so there is likely a difference between the queries. The distribution for query 2 also appears uniform and has a median value of 6 confident observations. When compared to the other queries, it appears that more participants were likely to be confident for query 2 than for query 1 and that more participants were likely to be unconfident for query 2 than for query 3 and 4. The distribution of data for queries 3 and 4 is concentrated around most participants being confident indicating that by this point, participants were likely to be confident in their decision.

It is harder to discern a relationship between confidence and difficulty from Figure 4.12. The interquartile ranges for the different difficulty levels overlap and the distribution of data for the easy and hard difficulties and the medium and very hard difficulties are very similar to each other. However, the median for the medium and very hard difficulties

do not overlap with the interquartile range of the easy and hard difficulty, suggesting that there may be a difference between these difficulty levels.

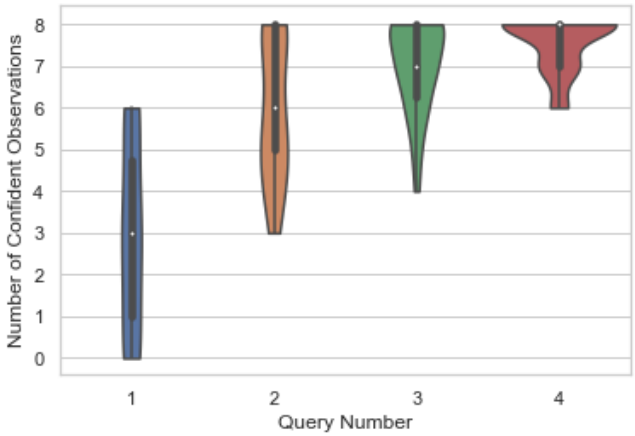

Figure 4. 11: Number of Confident Observations versus Query Number

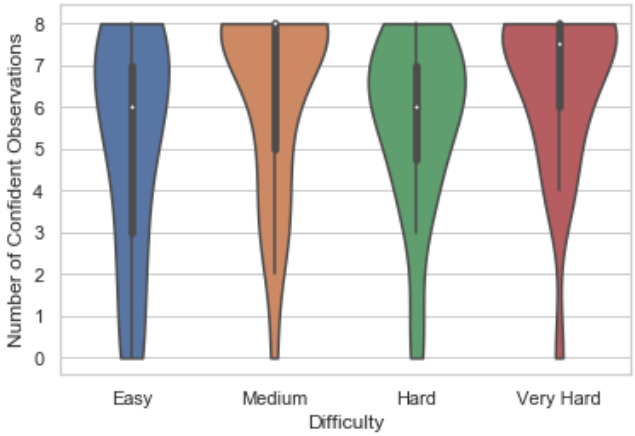

Figure 4.12: Number of Confident Observations versus Difficulty

Figure 4.13 shows histograms of tool transitions for both the confident and unconfident responses. Since there is almost a complete overlap between the two distributions, it is unlikely that there is relationship between confidence and the number of tool transitions.

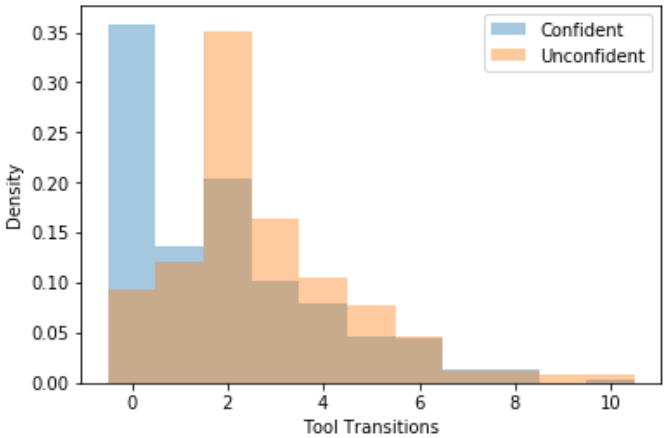

Figure 4.13: Distribution of Tool Transitions for Confident and Unconfident Responses

Since confidence is a binary response variable, a mixed effects logistic regression model was chosen to test the significance of query number, difficulty, and tool transitions while controlling for the individual differences of the participants. The results of the logistic regression are shown in Table 4.7 and are in agreement with the data exploration. Individual differences in participants accounted for 16.6% of the total variance. Query number was significant (p-value = 2e-16), indicating that query number had an effect on confidence. The positive coefficients for query number suggests that the probability of participants being confident increases with the amount of time they have to gather evidence for their decision. Difficulty was also significant (p-value = 1.38e-05). The positive coefficient for difficulty suggests that the probability of participants being confident increases with the difficulty of the alert. This is an interesting observation as it is in contradiction with the results presented by Borneman [2]. Tool transitions was not significant (p-value = 0.747) indicating that there is no relation between confidence and the number of times a participant switched between tools.

Table 4.7: Mixed Effects Logistic Regression Model for Confidence

Fixed Effects	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
Intercept	-2.40919	0.35993	-6.694	2.18e-11
Tool Transitions	-0.01529	0.04743	-0.322	0.747
Query Number	1.25574	0.10150	12.371	2e-16
Difficulty	0.34374	0.07906	4.348	1.38e-05
Random Effects	% of Total Variance			
Participant	14.3			

4.3.1.2 Reaction Time Modelling

The next avenue for behavioral data exploration and analysis was to determine whether the query number, difficulty, number of tool transitions, and confidence had an effect on reaction time. Figure 4.14 and 4.15 display violin plots of reaction time in milliseconds versus the query number and versus difficulty respectively. The overlap in the spread of interquartile ranges between all pairs of queries and small distance between medians suggests that it is unlikely that there is a difference in reaction times between the queries. However, in the distribution of the data for each query the distributions for the earlier queries appear to be denser at lower reaction times than the later queries. The violin plots for reaction time versus difficulty are almost indistinguishable, which suggests that it is unlikely that difficulty had an effect on reaction time. Figure 4.16 displays histograms of reaction times for both the confident and unconfident responses. Since there is almost a complete overlap between the two distributions, it is unlikely that confidence had an effect on reaction time. Similarly, Figure 4.17 displays reaction time versus number of tool transitions. The red trend line resulting from regressing reaction time on tool transitions suggests that the number of tool transitions does not have an effect on reaction time.

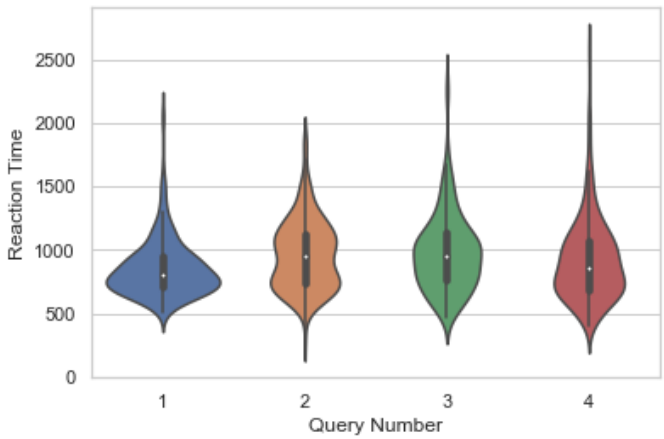

Figure 4.14 : Reaction Time versus Query Number

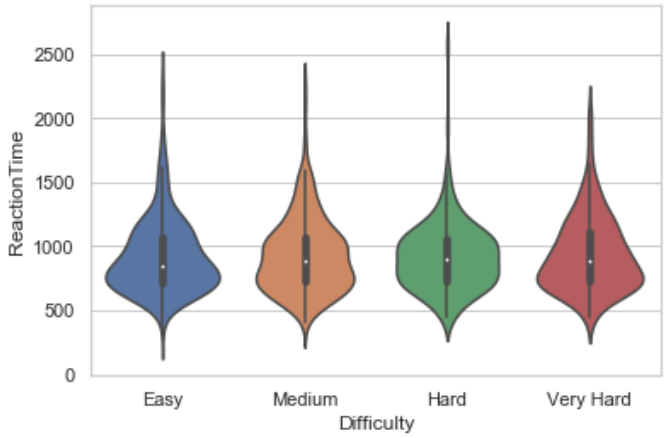

Figure 4.15 : Reaction Time versus Difficulty

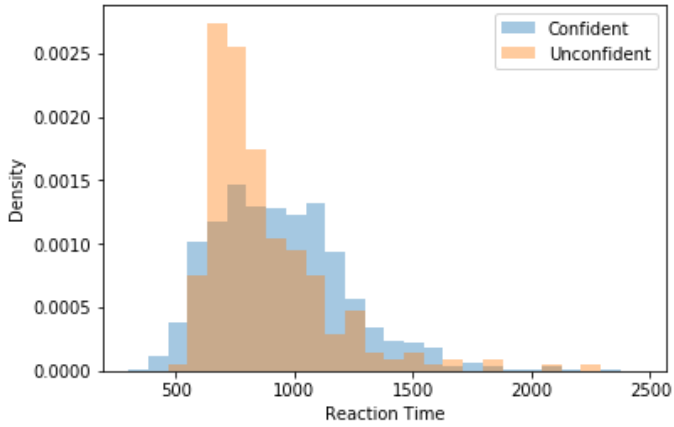

Figure 4.16 : Distribution of Reaction Times for Confident and Unconfident Responses

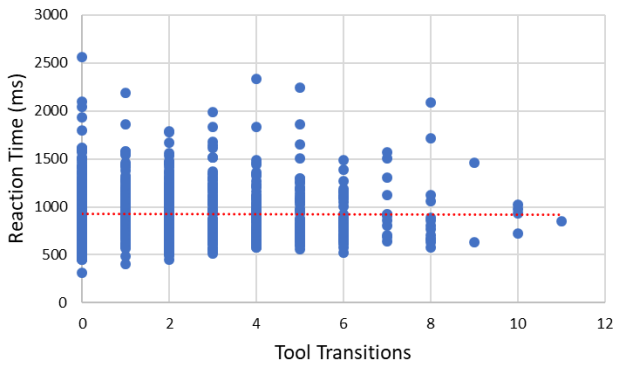

Figure 4.17 : Reaction Time versus Tool Transitions

A linear mixed model was fit to test the significance of query number, difficulty, confidence, and tool transitions while controlling for the individual differences of the participants. The results of the model are shown in Table 4.8. Individual differences in participants accounted for 13.3% of the total variance. None of the coefficients of the predictors were significant (p-value > 0.05) suggesting that query number, difficulty, confidence, and tool transitions did not have an effect on reaction time.

Table 4.8 : Linear Mixed Model for Reaction Time

Fixed Effects	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
Intercept	868.556	49.970	17.382	0.000
Query Number	12.014	9.524	1,261	0.207
Difficulty	1.553	7.818	0.199	0.843
Confidence	38.870	22.840	1.702	0.089
Tool Transitions	-0.764	4.782	-0.169	0.873
Random Effects	% of Total Variance			
Participant	13.0			

4.3.1.3 Tool Transitions Modelling

Figure 4.18 and 4.19 display violin plots of the number of tool transitions versus the query number and versus difficulty respectively. From Figure 4.18, it appears that the query number may have an effect on the number of tool transitions. The spread of the interquartile ranges for all pairs of queries overlap but there is sufficient separation between the medians. The distribution of the data suggests that tool transitions are lower for later queries. The violin plots for tool transitions versus difficulty are almost indistinguishable, which suggests that it is unlikely that difficulty had an effect on tool transitions. Similarly, the violin plots shown in Figure 4.20 are consistent with the results of the confidence modelling, suggesting that confidence does not have an effect on the number of tool transitions.

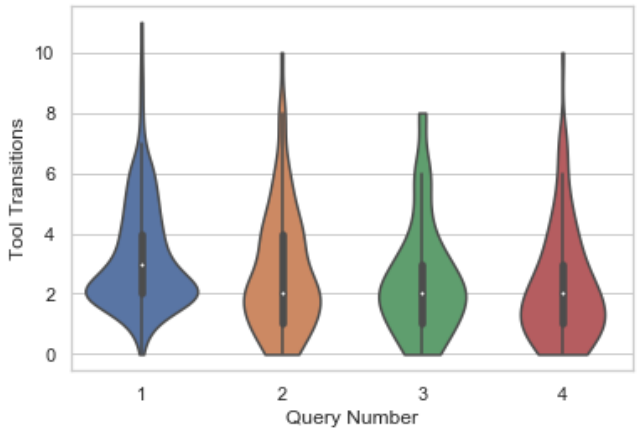

Figure 4.18: Tool Transitions versus Query Number

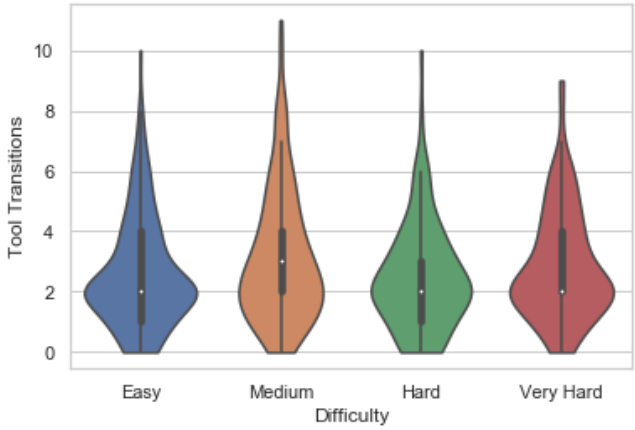

Figure 4.19 : Tool Transitions Versus Difficulty

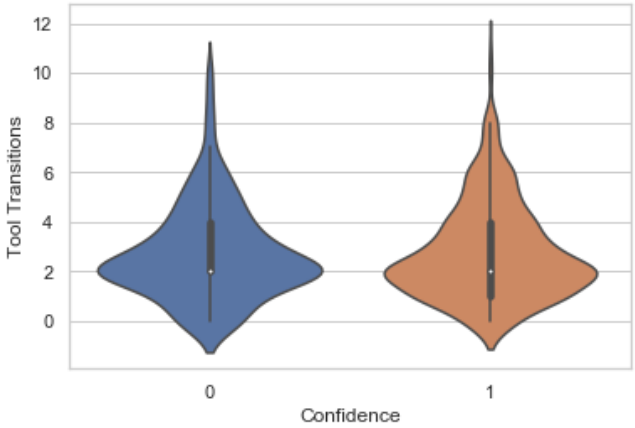

Figure 4.20: Tool Transitions versus Confidence

Similar to the reaction time modelling, a linear mixed model was fit to test the significance of query number, difficulty, and confidence while controlling for the individual differences of the participants. The results of the model are shown in Table 4.9. Individual differences in participants accounted for 6.8% of the total variance. Query number was significant (p-value = 0.000). The negative coefficient suggests that tool transitions decreased as query number increased, which is in agreement with the data exploration. As expected, neither difficulty or confidence were significant.

Table 4.9 : Linear Mixed Model for Tool Transitions

Fixed Effects	Estimate	Std. Error	z value	Pr(>|z|)
Intercept	4.03612	0.25747	15.676	0.000
Difficulty	-0.01581	0.15471	-0.102	0.919
Query Number	-0.72287	0.06017	-12.014	0.000
Confidence	-0.03891	0.05298	-0.734	0.463
Random Effects	% of Total Variance			
Participant	6.8			

4.3.1.4 Correctness Modelling

The last relationship investigated was the effect of query number, difficulty and number of tool transitions on participant correctness. Figure 4.21 and 4.22 display violin plots of the number of correct observations versus the query number and versus difficulty respectively. Figure 4.21 suggests that query number may have an effect on correctness. In the interquartile range, the spread for all pairs of queries overlap except 1 and 4, but there is sufficient separation between all pairs of medians. The distribution of the data suggests that participants were more likely to be correct for later queries. Similarly, Figure 4.22 suggests that difficulty may have had an effect. Looking at interquartile range, it is unlikely that there is a difference between the easy and medium or medium and hard difficulties. However, it is likely that there is a difference between the easy and very hard difficulties, as the distribution of data suggests that participants were more likely to be correct when responding to an easy alert than for a very hard alert. Figure 4.23 displays histograms of tool transitions for both correct and incorrect responses. Since there is almost a complete overlap between the two distributions, it is unlikely that tool transitions had an effect on correctness.

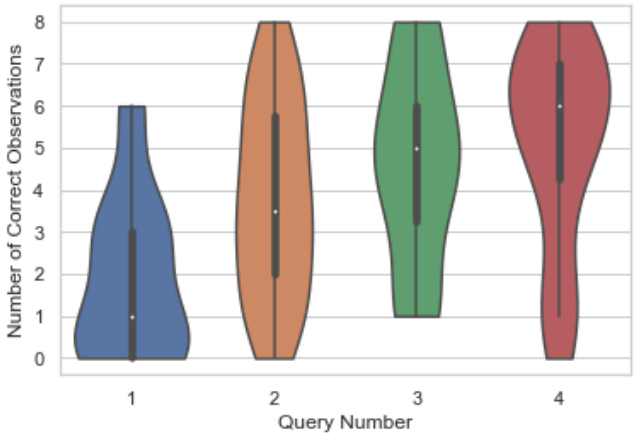

Figure 4.21: Number of Correct Observations versus Query Number

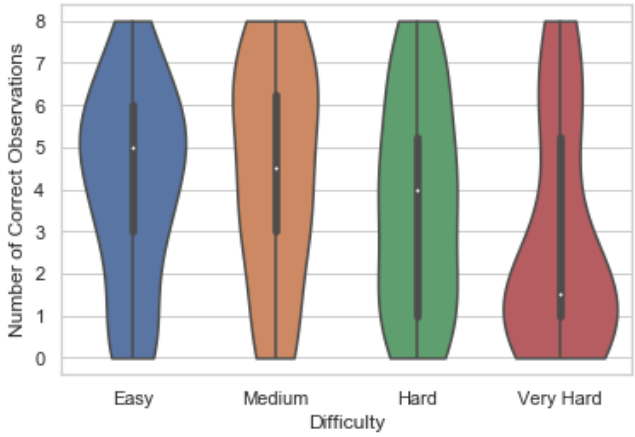

Figure 4.22 : Number of Correct Observations versus Difficulty

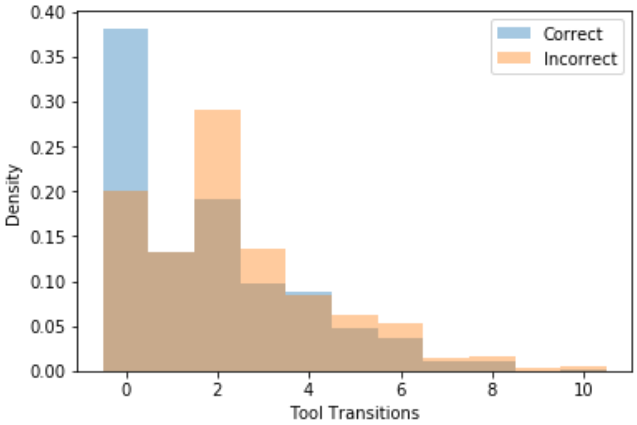

Figure 4.23: Distribution of Tool Transitions for Correct and Incorrect Responses

Similar to the confidence model, since correctness is a binary variable, a mixed effects logistic regression model was chosen to test the significance of the effects while controlling for the individual differences of the participants. The results of the logistic regression are shown in Table 4.10. Individual differences in participants accounted for 1.3% of the total variance. Consistent with observations made during the data visualization, query number was significant (p-value = 0.0027), indicating that correctness varied with query number. The positive coefficient for query number suggests that the probability of a participant being correct increases with the amount of time they have to gather evidence for their decision. Difficulty was also significant (p-value = 6.65e-05), indicating that correctness varied the difficulty of an alert. The negative coefficient indicates that the probability of a participant being correct decreased with the level of difficulty of an alert. The number of tool transitions was not significant (p-value = 0.3304).

Table 4.10: Mixed Effects Logistic Regression Model for Correctness

Fixed Effects	**Estimate**	**Std. Error**	**z value**	**Pr(>|z|)**
Intercept	-0.78620	0.26212	-2.999	0.0027
Query Number	0.53289	0.6816	7.819	5.34e-15
Difficulty	-0.24459	0.06133	-3.988	6.65e-05
Tool Transitions	-0.03577	0.03675	-0.973	0.3304
Random Effects	% of Total Variance			
Participant	1.3			

4.3.2 Event Related Potential Analysis

No statistically significant results were observed in any of the eight participants. It is likely that due to the small number of trials and class imbalance, not enough averaging was done to attenuate the noise so that the ERP becomes clear.

4.3.3 Classification of Confidence

The Electrophysiological analysis for the CIAT data was conducted in the same manner as for the RDK data. LR, LDA, RF, and fully connected ANN models were trained and tested using the mean power features from each of the five traditional EEG bands for each of the eight participants, resulting in a total of 32 models that were evaluated and compared. Model performance was evaluated using BACC, AUC, MCC, and Cohen's Kappa and the model in which three of the four performance metrics were highest was reported as the model with the best performance. Mean results across participants are given in Table 4.11. Each performance metric indicates performance fairly greater than random chance.

Table 4.11: Mean Performance of Frequency Band Models for the CIAT Task

Metric	Mean	95% Lower CL	95% Upper CL
BACC	0.641	0.608	0.673
AUC	0.635	0.601	0.669
MCC	0.261	0.200	0.322
Cohen's Kappa	0.247	0.184	0.310

The best performing model and the corresponding performance metrics for each participant are shown in Figure 4.24 to Figure 4.27. Each of these models exceeded the random chance value of 0.5 for BACC and AUC and 0 for MCC and Cohen's kappa, though model performance for most participants was substantially lower than for the RDK task with a mean decrease in BACC, AUC, MCC, and Cohen's kappa of 0.063, 0.062, 0.135, and 0.139 respectively. However, the best performing models for participants 7984 and 8079 actually exceeded the performance of their best models for the RDK task. Like the RDK task, models fit using a fully connected ANN were consistently the best across participants, providing the best performance for six of the eight participants. The RF

model provided the best results for the two remaining participants. The highest BACC and AUC among the best performing models were 0.729 and 0.716 with 95% CIs [0.696, 0.762] and [0.682, 0.750] respectively, and were associated with participant 7984's fully connected ANN. The highest MCC and Cohen's kappa among the best performing models were 0.419 and 0.404 with 95% CIs [0.358, 0.48] and [0.341, 0.467] respectively, which were also associated with participant 7984's fully connected ANN. The lowest BACC and AUC were 0.576 and 0.538 with 95% CIs [0.543, 0.609] and [0.504, 0.572] respectiviely which were both associated with participant 8477's fully connected ANN. The lowest MCC and Cohen's kappa were 0.142 and 0.106 with 95% CIs [0.081, 0.203] and [0.043, 0.169] respectively which were also associated with participant 8477's fully connected ANN.

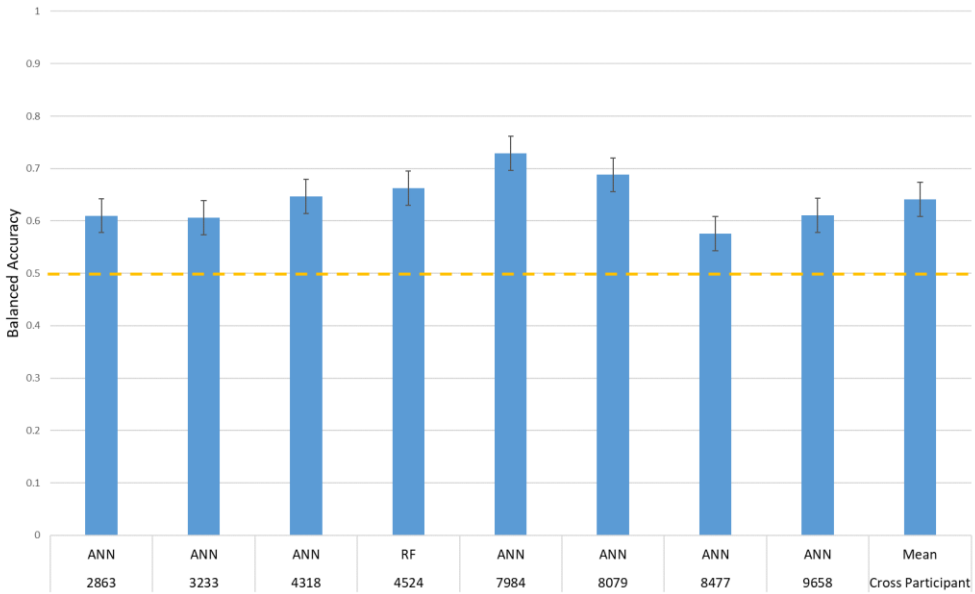

Figure 4.24: BACC for the Best Performing Models on the CIAT Task

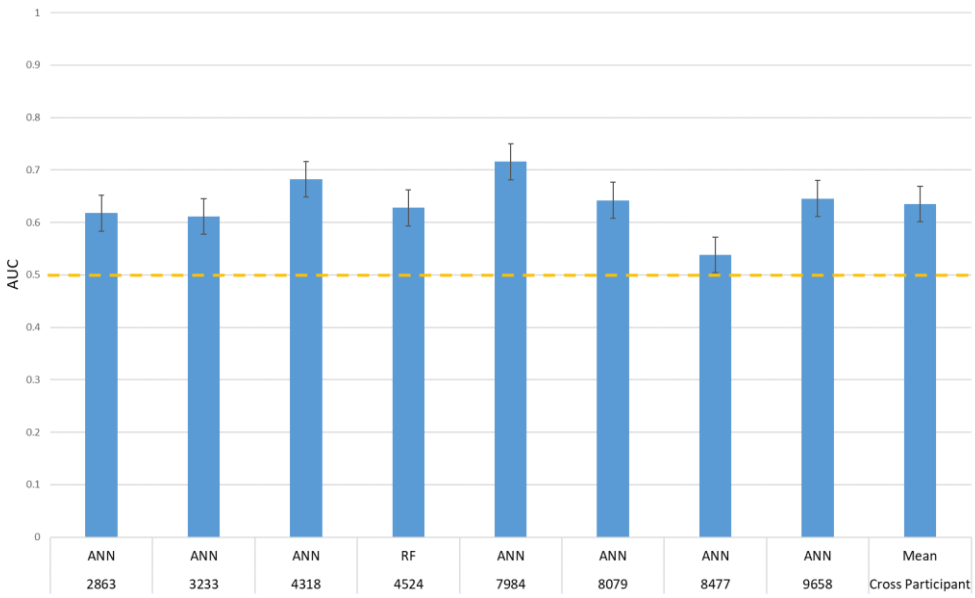

Figure 4.25: AUC for the Best Performing Models on the CIAT Task

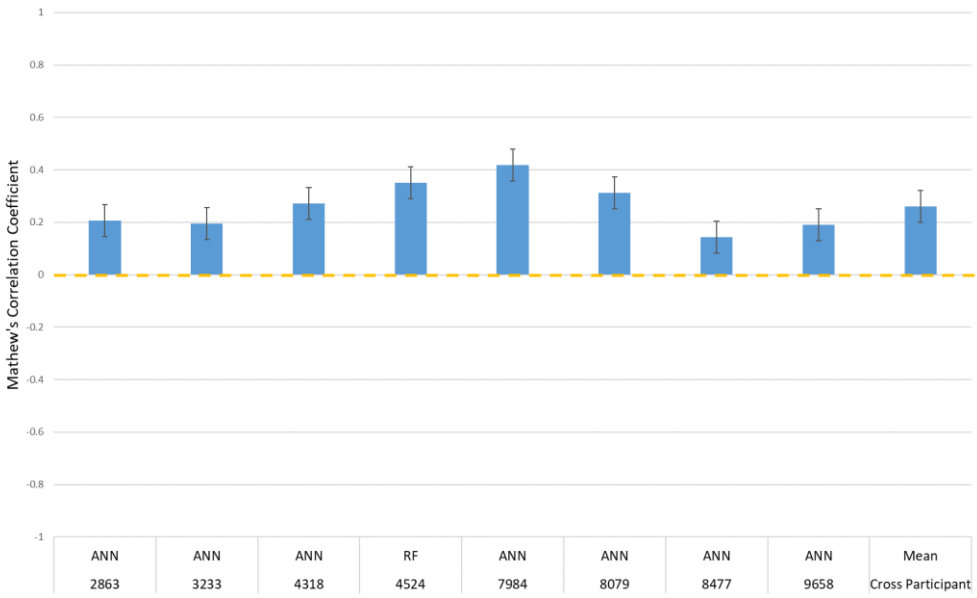

Figure 4.26: MCC for the Best Performing Models on the CIAT Task

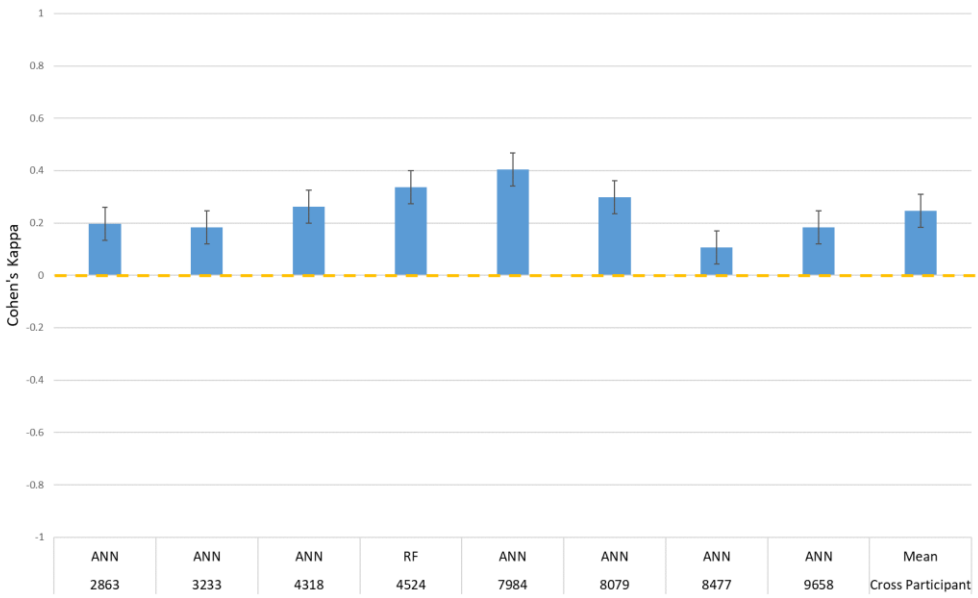

Figure 4.27: Cohen's Kappa for the Best Performing Models on the CIAT Task

To determine the utility of each frequency band towards classifying decision confidence, the same process used for the RDK task data where five single frequency band models were fit for each participant and compared against a paradigm where each participant's best performing model architecture was trained and evaluated using the frequency information from all but one band was used again. Table 4.12 displays each of the single band models compared to the leave-one-band-out models ranked by best performance and the largest decrease in performance, respectively. Similar to the RDK task, for all participants, models fit using features from all five frequency bands performed better than models fit using only individual bands or by leaving out any individual band. However, for five of the eight participants, there was no agreement between performance of the single band model and the leave-one-band-out model for two participants. This suggests that no frequency band provided significantly more utility than any other across participants.

Table 4.12: Comparison of CIAT Single Band Models (Column Header 1) to Leave one-band out (Column Header 4) Models with Respect to Highest Perfromance and Highest Perfromance Drop

	Participant															
	2863		**3233**		**4318**		**4524**		**7984**		**8079**		**8477**		**9658**	
Rank \ Bands	1	4	1	4	1	4	1	4	1	4	1	4	1	4	1	4
1	Δ	Θ	Δ	B	A	A	Γ	Γ	Θ	A	Θ	Θ	Δ	Γ	B	Δ
2	Θ	A	A	Θ	Δ	B	B	Θ	A	Θ	B	B	Γ	Δ	Δ	Θ
3	A	Γ	B	A	Θ	Γ	Δ	Δ	Δ	Δ	Γ	A	A	A	A	A
4	B	Δ	Θ	Δ	B	Δ	Θ	B	B	B	Δ	Δ	B	Θ	Θ	Γ
5	Γ	B	Γ	Γ	Γ	Θ	A	A	Γ	Γ	A	Γ	Θ	B	Γ	B

To further investigate the salient features for the CIAT task, feature importance was extracted using the same process as for the RDK data. Table 4.13 lists the features that were consistent across the logistic regression, LDA, and top 15 features ranked by the random forest models for each participant. For the three participants in which the single band model was in agreement with the leave-one-band-out model, features from the associated bands were included by the logistic regression and LDA models and also in the top 15 features ranked by the random forest model. However, the majority of channels selected were not from the expected regions of the brain. There is also no consistency across participants with respect to the channels selected.

Table 4.13: Salient Features Across LR, LDA, and RF Models for the CIAT Task

Participant							
2863	**3233**	**4318**	**4524**	**7984**	**8079**	**8477**	**9658**
CPz Δ	CP6 Δ	F7 A	P7 Γ	C2 Θ	T8 Θ	AF8 Δ	C1 Δ
Fz A	OZ Δ	FC3 B	C6 Γ	O1 Θ	Fp2 Θ	Fp2 Δ	TP7 B
CP4 Θ	C5 A		T8 Γ	F1 Θ		F7 A	AF4 Γ
			CP3 Δ	TP9 Δ		F2 Δ	
			FC6 B				

Analysis of the residuals revealed two patterns of misclassification across participants. First, models had difficulty inferring confident observations corresponding to the first query in an alert and unconfident observations corresponding to the last query in an alert. This is likely due to the small number of samples in which participants were initially confident in their decisions or unconfident in their final decisions, as confident observations corresponding to the first query in an alert comprise only 12.5% of the total number of confident observations and unconfident observations corresponding to the final query in an alert comprise only 5% of the total number of unconfident observations. Second, models had difficulty on observations in which the level of confidence was not the same as the previous decision. In other words, models had difficulty with confidence inference when there was a transition between levels of confidence. Similar to the previous observation, it is possible that this is due to the imbalance in the data with respect to decision transitions. Transitions in which confidence does not change represent 75% of the total number of decision transitions, whereas transitions in which confidence changes represents only 25%. It is also possible that important information encoding confidence, especially when there is a transition between levels of confidence, is captured during the evidence gathering portion of the task. However, since this information is not incorporated during the feature engineering process, the models are unable to learn these patterns. Ways to incorporate this information are discussed as future work.

The results of the behavioral and residual analysis imply that participant confidence is strongly tied to the alert query number. This suggests the need to compare to a new baseline which better controls for the effect of the query number. To make this comparison, two additional model types were fit per participant. The first model type was

a fully-connected ANN trained on the query number, which learned to always predict the majority class per query. The second model type was also a fully-connected ANN, but trained on both the query number and frequency domain EEG features. Both model types were tuned in the same manner described in Section 3.6.2.2.3. The performance of these models compared to the best performing EEG models are shown in Figures 4.28 to 4.31. For five of the eight participants, the query-only model performed noticeably better than the corresponding EEG-only model. The boost in performance for these five participants can be attributed to the class imbalance with respect to the first query. For these participants the data was much more skewed towards the unconfident class, and so these models were able to get more unconfident observations correct. For all participants, performance of the model trained on both query number and EEG features failed to outperform the query only model, performing strictly worse in seven of the eight participants. This indicates that the addition of the EEG features does not help improve model generalization. Possible reasons for this inability to generalize are similar to those discussed for the RDK task. First, it is possible that the hyperparameter search was too shallow and that a more careful tuning approach could result in better performance. Second, it is possible that there is an issue with the quality of the data. As discussed in Section 1.5, a major limitation of this research was that the equipment used to collect the electrophysiological data was known to be malfunctioning during the time of the experiment. The amount of extra noise introduced into the data due to this problem is unknown. Third, the assumption that prominent neural representations of confidence manifest at the time of a decision which formed the basis of the feature engineering

process may be incorrect. Lastly, it may be that there just isn't enough data to learn important patterns associated with confidence.

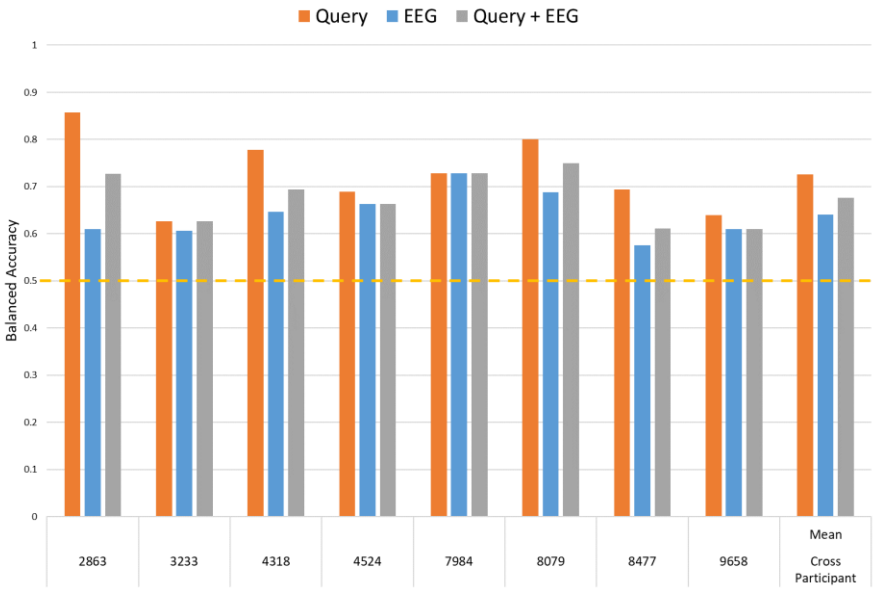

Figure 4.28: Comparison of BACC When Controlling for Query

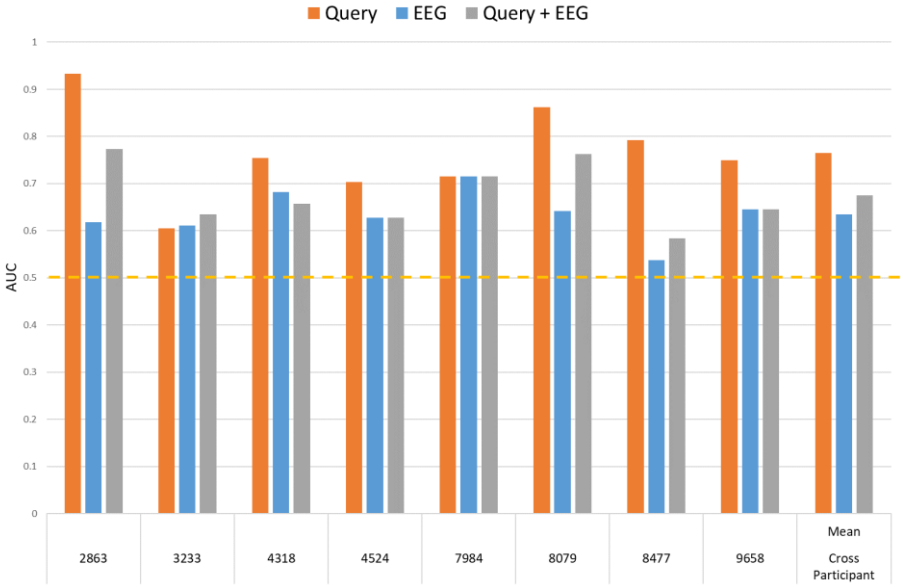

Figure 4.29: Comparison of AUC When Controlling for Query

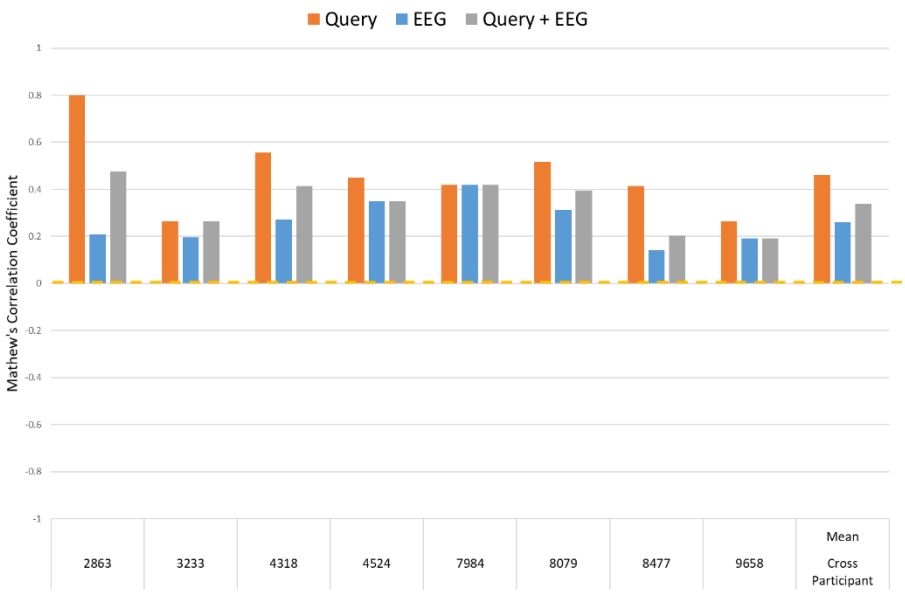

Figure 4.30: Comparison of MCC When Controlling for Query

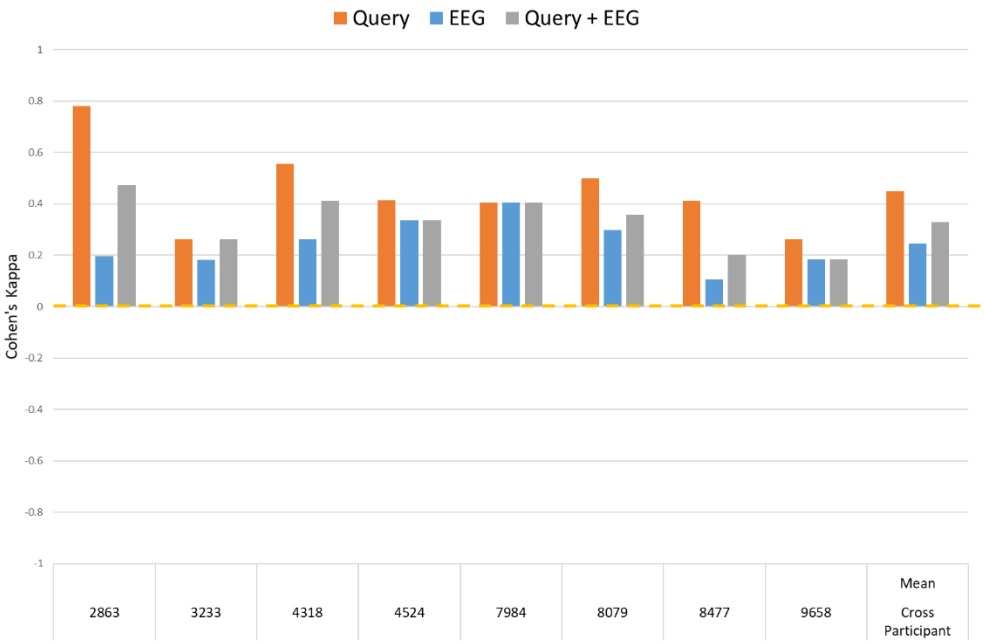

Figure 4.31: Comparison of Cohen's Kappa When Controlling for Query

As in the RDK data, classification using the time series features via a CRNN was also investigated. Each CRNN was trained using the process described in Section 3.6.2.2.4. Models were able to achieve 100% training accuracy after an average of 100

epochs. However, models corresponding to the lowest validation loss took an average of 30 epochs to train. The average time per epoch was 6.4 seconds. Performance of the CRNN models is shown in Figures 4.32 to 4.35. Mean results across participants are given in Table 4.14.

Table 4.14: Mean Performance of the CRNN Models for the CIAT Task

Metric	**Mean**	**95% Lower CL**	**95% Upper CL**
BACC	0.530	0.503	0.530
AUC	0.533	0.497	0.531
MCC	0.058	0.009	0.058
Cohen's Kappa	0.055	0.007	0.055

The best performing model had a BACC of 0.562, 95% CI [0.548, 0.578], AUC of 0.564, 95% CI [0.546, 0.582], MCC of 0.116, 95% CI [0.091, 0.140] and Cohen's kappa of 0.113, 95% CI [0.089, 0.139], which were the highest values for each metric across participants and were associated with participant 7984. No model performed at a level reasonably greater than random chance. The worst BACC and AUC were 0.503, and 0.499 with 95% CIs [0.490, 0.517] and [0.481, 0.517] respectively and were associated with participant 4318. The worst MCC and Cohen's kappa were 0.007 and 0.006 with 95% CIs [-0.017, 0.032] and [-0.017, 0.030] and were associated with participant 2863. For all participants the CRNN model performed substantially worse than their best performing model fit using frequency band information, with an average decrease in performance of 0.124 for BACC, 0.119 for AUC, 0.228 for MCC, and 0.215 for Cohen's kappa. Similar to the CRNN fit to the RDK time series data, analysis of the residuals did not reveal any patterns of misclassification other than the tendency to predict the confident class for the majority of observations. Once again, a more thorough hyperparameter search may

improve performance. However, it is more likely that number of samples available for training isn't large enough for the network to learn anything meaningful.

Figure 4.32: BACC for the CRNN fit on the CIAT Task Data

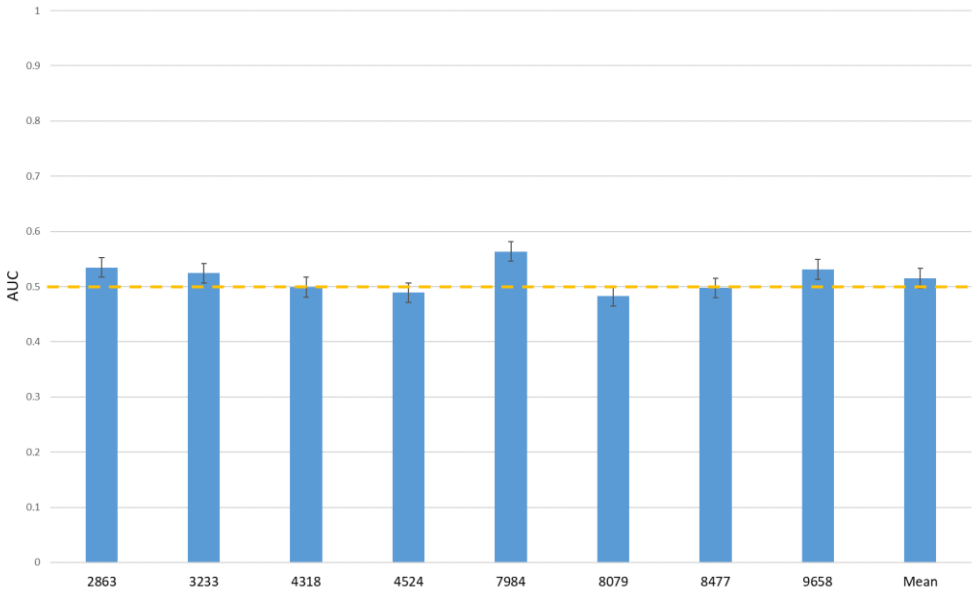

Figure 4.33: AUC for the CRNN fit on the RDK Task Data

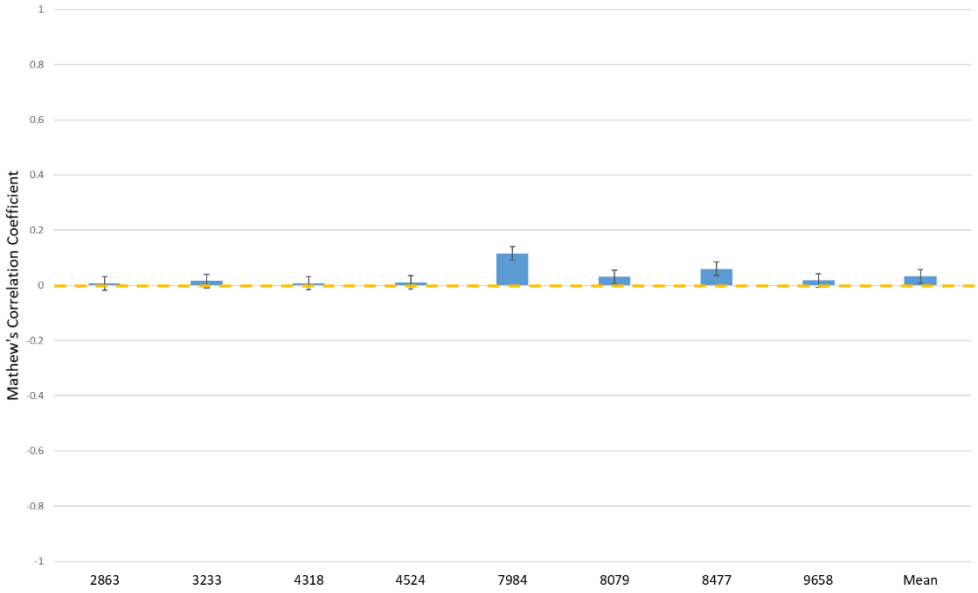

Figure 4.34: MCC for the CRNN fit on the CIAT Task Data

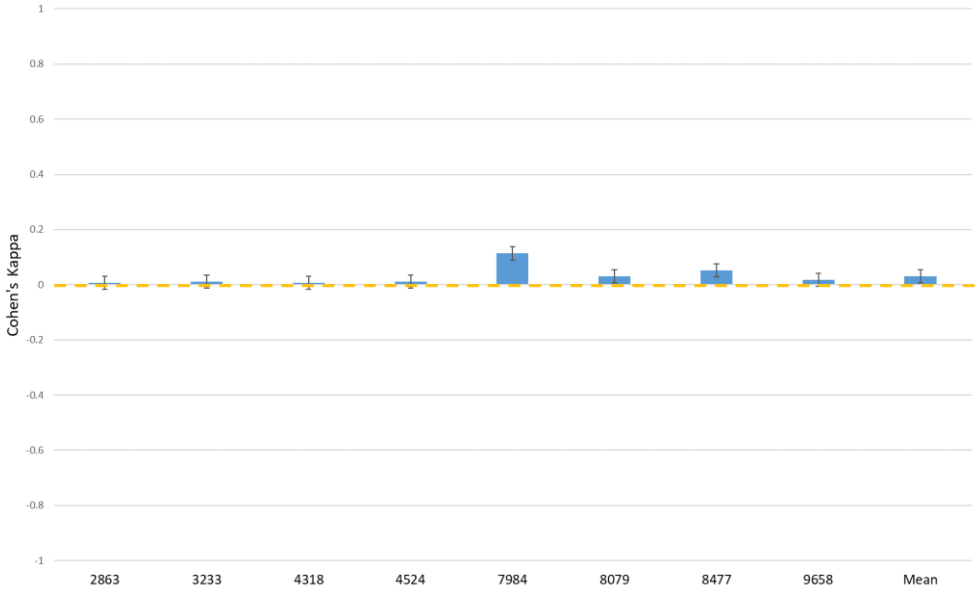

Figure 4.35: Cohen's Kappa for the CRNN fit on the CIAT Task Data

4.4 Summary

The objective of this study was to attempt to fill the current research gap of using neural and behavioral correlates of decision confidence as features for tackling the

problem of confidence inference in both simple and complex decisions using machine learning. The analysis and results showed that EEG could be used in combination with machine learning to infer confidence in a simple decision with a performance greater than chance, but that more research is necessary to evaluate the utility of using EEG to infer confidence in the types of decisions made by cyber operators in their operational environment. For the RDK task, the mean performance across participants of classification models fit using the collected EEG data exceeded random chance with respect to four performance metrics. In addition, mean power in the alpha band was identified as the most important feature in half the participants. For the CIAT task, it was expected that participant reaction time and information seeking behaviors would be related to confidence and thus be suitable features for machine learning. However, the statistical analysis showed that these behaviors were not significant when accounting for when a participant was queried for a decision and the alert difficulty. The addition of EEG features was also observed to provide little utility when compared to a naïve model which always predicts the majority class per query.

V. Conclusions and Recommendations

5.1 Conclusions of Research

This study was successful in its objective of investigating the use of neural and behavioral correlates of decision confidence in combination with machine learning techniques to infer confidence in a simple decision as well as investigating whether the results extended to more complex decisions similar to those made by cyber defense operators. In order to achieve this goal, a two-task human-subject experiment was designed in which electrophysiological and behavioral data was recorded and analyzed.

The first research question posed in this work investigated if electrophysiological features could be used in combination with machine learning techniques to infer decision confidence in simple decisions with a performance greater than chance. EEG data was collected from a motion discrimination task in which participants had to decide whether the global direction of dot motion for each RDK stimulus was to the left or to the right. As hypothesized, machine learning models were able to learn neural correlates of decision confidence from frequency domain representations of the EEG data. The best performing models achieved a performance greater than random chance with respect to four performance metrics for all participants. Fully-connected ANNs typically had the best performance, ranking as the top model for seven out of eight participants. Models exceeded the baseline BACC and AUC of 0.50 with a mean BACC of 0.704 and mean AUC of 0.697 and exceeded the baseline MCC and Cohen's kappa of 0 with a mean MCC of 0.399 and mean Cohen's kappa of 0.386.

The second research question sought to determine the important features for decision confidence classification in a simple decision. Results of the analysis in which five single frequency band models were fit for each participant and compared against a paradigm where each participant's best performing model architecture was trained and evaluated using the frequency information from all but one band suggest that the alpha band features were most important as the models were in agreement in half the participants. To investigate spatial importance with respect to the individual frequency bands, feature importance as determined by recursive feature elimination and random forest feature importance were examined. The features selected by these algorithms provide further support for the importance of individual frequency bands, however there was no consistency with respect to channels across participants, demonstrating that spatial importance varied with participant.

The third research question investigated the relationship between participant behaviors and decision confidence. It was hypothesized that reaction time and information seeking behaviors would be useful features for decision confidence classification. However, when accounting for the query number and difficulty, it was observed that across participants, no relationship existed between reaction time, tool transitions, and decision confidence. These results suggest there is no utility in using these behaviors as features for classification of decision confidence.

The final research question investigated whether the answers to the previous three questions extend to the complex decisions made by cyber defense operators in their operational environment. Once again, the best performing models achieved a performance greater than random chance with respect to four performance metrics for all participants,

though performance was typically worse than for the RDK task. However, when controlling for the effect of the query number by comparing against a naïve model which always predicts the majority class per query, the addition of EEG did not improve results. Additionally, whereas alpha band features were determined to be the most important for the RDK task, no frequency band provided significantly more utility than any other across participants for the CIAT task. Similar to the RDK task, no consistency was observed with respect to channels across participants, demonstrating that spatial importance varied with participant.

5.2 Significance of Research

Current research on decision confidence inference from electrophysiological data has focused solely on decisions which meet the assumptions of the drift-diffusion model [5]. However, in the cyber operational environment, the decisions made by cyber operators as they investigate potential threats do not meet these assumptions. This work augmented existing studies on confidence inference from EEG signals by exploring the use of more flexible machine learning models such as the random forest classifier and fully-connected ANN, and was the first to apply these techniques to decision confidence inference in a motion coherence task using RDKs. Though inconclusive, this work is also the first to investigate decision confidence inference using machine learning models trained on EEG signals collected from decisions similar to those made in the cyber operational environment. The performance evaluation of the machine learning models fit using the CIAT data serves as a reminder that a blind reliance on common performance metrics can inflate results.

5.3 Recommendations for Future Research

Many avenues exist in which the problem of decision confidence inference could be further explored and involve either additional analysis of the data collected during this study or modification of the experimental design to facilitate new data collection. Several of those avenues are recommended in this section.

5.3.1 CIAT Data Segmentation

In this study only one form of data segmentation was utilized to label and transform the raw EEG data into features suitable for machine learning. The method of data segmentation assumed that confidence is reflected in the EEG data between just before the decision query up until the decision submission, and that all time points falling in this window reflect the same confidence level. A major disadvantage to this method is that it ignores the data collected during the investigative portion of the task, which accounts for 80% of the total data. The reason for ignoring this data is that it is difficult to label it without additional reported confidence information obtained from participant responses. Any incorrect labelling effectively amounts to introducing noise into the model fitting process. However, it is likely that there are patterns associated with confidence in this data that were not represented in the data the was utilized. One avenue for incorporating the unused data is to segment the data using a non-stimulus aligned approach. A possible implementation would be to label data segments prior to a decision with the confidence level for that decision. Mislabeling of data could then be reduced by only retaining data segments in which confidence levels did not change between decisions.

5.3.2 Machine Learning Improvements

5.3.2.1 Dimensionality Reduction

As mentioned in IV, the problem of having high-dimensional data and a small number of observations, known as the curse of dimensionality, may have impacted classifier performance. Future work should investigate dimensionality reduction techniques to reduce the number of features used for model fitting. In particular, the use of Principal Component Analysis (PCA) should be explored. PCA finds a low-dimensional representation of a data set that contains as much of the variation as possible [25]. It does so by transforming the set of features into a set of linearly uncorrelated variables known as principle components. The first principle component accounts for the largest amount of variance in the training set, and each succeeding component accounts the largest amount of remaining variance. Thus using PCA, the set of 320 features used in this research can be reduced to the top N principle components, potentially lowering the impact of the curse of dimensionality along with model capacity.

5.3.2.2 Group Modelling

This research considered only single-participant models that were fit solely on data from the participant being modeled and not data from other participants. Since these models are tuned to the specific individual, a separate model must be trained for each new individual. This requirement is both resource intensive and computationally expensive, and may be impractical for inferring decision confidence in real-world operational environments. Future work should investigate the performance of group modelling, where data from a set of individuals is used for model training and the models are later used to

infer confidence in decisions made by those individuals or potentially new individuals. One possible implementation of a group model for the experiment data collected in this research would be to use a nested cross validation where the outer cross validation loop leaves one participant out and the inner loop trains a model using leave-one-out cross validation (LOOCV) on the remaining seven participants. This process would provide some insight towards the generalization performance of a group model, but would not be informative over hyperparameter selection.

5.3.2.3 Feature Importance Analysis

It was shown that fully-connected ANNs consistently produced the best results across participants for both tasks. However, ANNs offer little in terms of explanatory insight into the importance of features used during the prediction process. In this study, importance of the individual frequency bands was estimated by excluding individual frequency bands from the input, and then training and testing the ANN. The most important bands were taken as those that resulted in the biggest decline in classification performance when excluded. However, this method did not take the specific channels into account and so channel importance had to be investigated using models that were not directly comparable. Several methods exist which can be used to better estimate feature importance. In particular, the connection weights method [62] should be investigated and compared with the results of this study. This method calculates variable importance as the product of the raw input-hidden and hidden-output connection weights between each input and output neuron and sums the product across all hidden neurons. It has been shown to be

the best methodology for accurately quantifying variable importance when compared to other published methods [63].

5.3.3 ECG and EOG Analysis

While EOG and ECG data was recorded, these signals were not analyzed in this work. Future research should investigate the utility of using these signals as features as results obtained by Shih et al. suggest that incorporating them can improve classification performance when compared to models fit using only EEG data [27].

5.3.4 Experimental Design Changes

In order to increase the number of observations, the CIAT experiment was modified to query participants for a decision at regular intervals. Unfortunately, this query system had the unintended effect of introducing a large class imbalance with respect to the individual queries. Participants were typically unconfident at the time of the first query and confident at the time of the last. Future work should investigate ways to increase the number of observations without having to query participants for decisions. A potential solution would be to rework the alerts such that each alert could be accomplished in a shorter amount of time and then increase the number of alerts.

The experimental design in this study modelled decision confidence as a binary response variable. However, since decision confidence reflects an estimate of the probability that the decision is correct it can also be modelled as an ordinal variable with more than two levels or as a continuous response variable. By changing the way in which the confidence response variable is modelled, the problem of decision confidence inference could be explored as either a multiclass classification or regression analysis

respectively. In order to change the data type of the confidence response variable, some changes need to be made to both the RDK and CIAT experiment interfaces. First, the “I Don’t Know” option must be removed from the decision prompt for both tasks. Then, a new prompt which asks participants how confident they are in their decision should be inserted immediately after the last prompt. For the multiclass problem, this new prompt would have participants submit their confidence as one of several discrete levels such as “Not Confident”, “Confident”, and “Very Confident”. For the regression problem, a confidence slider such as the one used by Borneman [2] can be implemented.

Appendices

Appendix A: Pre-Experiment Questionnaire

ID: _____ Date: _____

Pre-Experiment Questionnaire (ONLY Experiment Day)

How many hours of sleep did you have last night?
Circle one choice: 0-4 hours, 5-6 hours, 7-9 hours, 9+ hours

How would you characterize your sleep last night?
Circle one choice: Very Poor, Poor, Fair, Good, Very Good

Did you consume any products with caffeine today?
Circle one choice: yes or no
If yes:
What product(s) did you consume?
_____
When did last consume this product?
_____
Approximately how much (mg / ounces / cups) of this product have you consumed today? _____

Do you have any reason(s) to believe that your ability to accomplish tasks during this study (including investigating cyber alerts and making decisions about them) today would be abnormal (for example: distracted, overly tired, hungry, stressed, injured)? _____
_____
If yes:
Do you still want to participate in the cyber study today? Circle one choice: Yes / No
If no:
Would you like to reschedule participation for another day? _____

Appendix B: Post-Experiment Questionnaire

ID: _____ Date: _____

Post-Experiment Questionnaire (ONLY Experiment Day)

In general, how difficult were the cyber investigations for you? (Circle one choice)

Very Easy	Easy	Moderate	Hard	Very Hard
1	2	3	4	5

Computer experience:

What sort of electronic devices do you use?

Circle all choices:

Personal computer/Desktop/Laptop
TV/Game Console
Smartphone/Tablet
Enterprise Server
Other, _____

How often do you use electronic devices?

Response items: Daily, A few times a week, Once a week, Never

Do you use electronic devices in your job?

Response items: Yes, No, Prefer not to answer

Do you have any cyber security experience?

Response items: Yes, No, Prefer not to answer

Have you earned any cybersecurity certifications?

Response items: Yes, No, Prefer not to answer

If yes:

Please list any cyber security certifications you have earned: _____
_____

Age: _____

Are you male or female? Male___ Female____ Prefer not to answer ____

What's your highest education level?

A. Lower than high school
B. Graduated from high school
C. Some college, no degree
D. Associate's Degree
E. Bachelor's Degree
F. Master's degree
G. Ph.D. degree

Appendix C: General Cyber Alert Investigation Workflow Handout

Experiment Information Only: Not for use in real-world operations.

Cyber Alert Classification Process Workflow

This is general guidance for use in the Cyber Alert Classification Experiment.

1) Look at the information in the alert and notice the alert name and time the alert was generated.
   - The Alert Name will suggest which signature/behavior to examine first in the **Alert Lookup** tool
   - You will want to determine whether the alert triggered before or after the captured behavior in the **PCap** and **Frame Info** tools
   - If you are unfamiliar with any terms or acronyms, consult the **Glossary** tool

2) Open the **Alert Lookup** tool.
   - Your goal is to understand what the signature is, and why it may have triggered on the **PCap** information (in the next step). Look at the description of the alert and identify triggering information to confirm later.

3) Open the **PCap** (Packet Capture) tool to look at (simulated) raw packet information. Devices that generate cyber alerts use rules based on PCap information, but the rules may not always work properly.
   - Your goal is to confirm the suspected threat occurred by comparing this raw data to the **Alert Lookup** details.
   - Compare the **PCap** info's IPs (Source/Destination), Protocol, and Info fields. You will want to see if these match the signature.

4) The **Frame Info** tool will often provide more detailed information corresponding to the rows from the **PCap** tool.
   - At a minimum, the information from the **Frame Info** tool will consist of a verbose form of the data from the PCap tool, and may provide more detailed network activity logs. Additional log information will be available in this tool, which can help validate the signature or lead to other search terms in the **Glossary** tool.

5) Next look at the **Network Info** tool. It contains info about whether certain IP addresses are known to be dangerous or safe. If it doesn't contain an IP then there is no info on that IP.
   - Check whether any of the IPs (Source/Destination) from the alert on the main screen are listed in this tool and whether they are known to be dangerous or safe.
   - Check the IPs displayed in the **PCap** tool to see if any of them are known to be dangerous or safe.

6) Review and complete your typed case notes, as they will help you remember important details found, details which are missing and details which are conflicting/inconsistent to help you make your decision.
7) |

Experiment Information Only: Not for use in real-world operations.

Bibliography

[1] G. Hockey, "Operator functional state: The prediction of breakdown in human performance," in *Measuring the Mind Speed, control, and age*, 2005, pp. 373–394.

[2] M. Borneman, "Estimating Defensive Cyber Operator Decision Confidence," Air Force Institute of Technology, https://scholar.afit.edu/etd/1796, 2018.

[3] C. S. Ho and D. E. Giaschi, "Low- and high-level first-order random-dot kinematograms: Evidence from fMRI," *Vision Res.*, vol. 49, no. 14, pp. 1814–1824, 2009.

[4] J. I. Gold and M. N. Shadlen, "The Neural Basis of Decision Making," *Annu. Rev. Neurosci.*, vol. 30, no. 1, pp. 535–574, 2007.

[5] R. Ratcliff and G. McKoon, "The Diffusion Decision Model: Theory and Data for Two-Choice Decision Tasks," *Neural Comput.*, vol. 20, no. 4, pp. 873–922, 2008.

[6] N. Yeung and C. Summerfield, "Metacognition in human decision-making: confidence and error monitoring," *Philos. Trans. R. Soc. Lond. B. Biol. Sci.*, vol. 367, no. 1594, pp. 1310–21, May 2012.

[7] M. Falkenstein, J. Hoormann, S. Christ, and J. Hohnsbein, "ERP components on reaction errors and their functional significance: A tutorial," *Biol. Psychol.*, vol. 51, no. 2–3, pp. 87–107, 2000.

[8] M. X. Cohen, "Where Does EEG Come From and What Does It Mean?," *Trends Neurosci.*, vol. 40, no. 4, pp. 208–218, 2017.

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

[10] B. J. Roach and D. H. Mathalon, "Event-related EEG time-frequency analysis: An overview of measures and an analysis of early gamma band phase locking in

schizophrenia," *Schizophr. Bull.*, vol. 34, no. 5, pp. 907–926, 2008.

[11] T. Picton, "The P300 Wave of the Human Event-Related Potential," *J. Clin. Neurophysiol.*, vol. 9, pp. 456–479, 1992.

[12] G. A. Kerkhof, "Decision Latency: The P3 Component In Auditory Signal Detection," *Neurosci. Lett.*, vol. 8, pp. 289–294, 1978.

[13] A. Selimbeyoglu, Y. Keskin-Ergen, and T. Demiralp, "What if you are not sure? Electroencephalographic correlates of subjective confidence level about a decision," *Clin. Neurophysiol.*, vol. 123, no. 6, pp. 1158–1167, 2012.

[14] A. Boldt and N. Yeung, "Shared Neural Markers of Decision Confidence and Error Detection," *J. Neurosci.*, vol. 35, no. 8, pp. 3478–3484, 2015.

[15] P. Stoica and R. Moses, "Spectral Analysis of Signals," *Prentice Hall*, 2005.

[16] A. Hramov, A. Koronovskii, V. Makarov, A. Pavlov, and E. Sitnikova, *Wavelets in Neuroscience*. 2014.

[17] J. S. P. Macdonald, S. Mathan, and N. Yeung, "Trial-by-trial variations in subjective attentional state are reflected in ongoing prestimulus EEG alpha oscillations," *Front. Psychol.*, vol. 2, no. MAY, pp. 1–16, 2011.

[18] V. Kolev, J. Yordanova, M. Schürmann, and E. Başar, "Increased frontal phase-locking of event-related alpha oscillations during task processing," *Int. J. Psychophysiol.*, vol. 39, no. 2, pp. 159–165, 2001.

[19] J. Kubanek, N. J. Hill, L. H. Snyder, and G. Schalk, "Cortical alpha activity predicts the confidence in an impending action," *Front. Neurosci.*, vol. 9, no. JUL, pp. 1–15, 2015.

[20] M. Graziano, L. C. Parra, and M. Sigman, "Neural Correlates of Perceived

Confidence in a Partial Report Paradigm," *J. Cogn. Neurosci.*, vol. 27, no. 6, pp. 1090–1103, 2015.

[21] J. Samaha, L. Iemi, and B. R. Postle, "Prestimulus alpha-band power biases visual discrimination confidence, but not accuracy," *Conscious. Cogn.*, vol. 54, pp. 47–55, 2017.

[22] U. Engelke, A. Maeder, and H.-J. Zepernick, "On confidence and response times of human observers in subjective image quality assessment." pp. 910–913, 2009.

[23] W. Robitza and H. Hlavacs, "Assessing the Validity of Subjective QoE Data through Rating Times and Self-Reported Confidence." 2014.

[24] K. Desender, A. Boldt, and N. Yeung, "Subjective Confidence Predicts Information Seeking in Decision Making," *Psychol. Sci.*, 2018.

[25] G. James, D. Witten, T. Hastie, and R. Tibshirani, *An introduction to statistical learning*, vol. 112. Springer, 2013.

[26] T. Hastie, R. Tibshirani, and J. Friedman, *The Elements of Statistical Learning*. New York, NY, USA: Springer New York Inc., 2001.

[27] V. Shih, L. Zhang, C. Kothe, S. Makeig, and P. Sajda, "Predicting decision accuracy and certainty in complex brain-machine interactions," *2016 IEEE Int. Conf. Syst. Man, Cybern. SMC 2016 - Conf. Proc.*, pp. 4076–4081, 2017.

[28] P. P. Paul, H. Leung, D. A. Peterson, T. J. Sejnowski, and H. Poizner, "Detecting neural decision patterns using SVM-based EEG classification," *2010 4th Int. Conf. Bioinforma. Biomed. Eng. iCBBE 2010*, 2010.

[29] L. Parra *et al.*, "Linear Spatial Integration for Single-Trial Detection in Encephalography," *Neuroimage*, vol. 17, no. 1, pp. 223–230, 2002.

[30] M. Steinhauser and N. Yeung, “Decision Processes in Human Performance Monitoring,” *J. Neurosci.*, vol. 30, no. 46, pp. 15643–15653, 2010.

[31] S. Gherman and M. G. Philiastides, “Neural representations of confidence emerge from the process of decision formation during perceptual choices,” *Neuroimage*, vol. 106, pp. 134–143, 2015.

[32] A. Gron, *Hands-On Machine Learning with Scikit-Learn and TensorFlow: Concepts, Tools, and Techniques to Build Intelligent Systems*, 1st ed. O’Reilly Media, Inc., 2017.

[33] V. Vijayakumar, M. Case, S. Shirinpour, and B. He, “Quantifying and Characterizing Tonic Thermal Pain Across Subjects From EEG Data Using Random Forest Models,” *IEEE Trans. Biomed. Eng.*, vol. 64, no. 12, pp. 2988–2996, 2017.

[34] F. Chollet, *Deep Learning with Python*, 1st ed. Greenwich, CT, USA: Manning Publications Co., 2017.

[35] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, no. 7553, pp. 436–444, 2015.

[36] I. G. Y. Bengio and A. Courville, “Deep Learning,” 2016.

[37] G. F. Wilson and C. a Russell, “Real-time assessment of mental workload using psychophysiological measures and artificial neural networks.,” *Hum. Factors*, vol. 45, no. 4, pp. 635–643, 2004.

[38] J. C. Christensen, J. R. Estepp, G. F. Wilson, and C. A. Russell, “The effects of day-to-day variability of physiological data on operator functional state classification,” *Neuroimage*, vol. 59, no. 1, pp. 57–63, 2012.

[39] R. G. Hefron, B. J. Borghetti, J. C. Christensen, and C. M. S. Kabban, "Deep long short-term memory structures model temporal dependencies improving cognitive workload estimation," *Pattern Recognit. Lett.*, vol. 94, pp. 96–104, 2017.

[40] S. Tripathi, S. Acharya, R. D. Sharma, S. Mittal, and S. Bhattacharya, "Using Deep and Convolutional Neural Networks for Accurate Emotion Classification on DEAP Dataset.," *Twenty-Ninth AAAI Conf.*, pp. 4746–4752, 2017.

[41] R. Bogacz, E. Brown, J. Moehlis, P. Holmes, and J. D. Cohen, "The Physics of Optimal Decision Making : A Formal Analysis of Models of Performance in Two-Alternative Forced-Choice Tasks," *Psychol. Rev.*, vol. 113, no. 4, pp. 700–765, 2006.

[42] J. Peirce, "Generating stimuli for neuroscience using PsychoPy," *Front. Neuroinform.*, vol. 2, p. 10, 2009.

[43] K. S. Pilz, L. Miller, and H. C. Agnew, "Motion coherence and direction discrimination in healthy aging," *J. Vis.*, vol. 17, no. 1, p. 31, 2017.

[44] Cognionics, "Mobile-72 Wireless EEG System." .

[45] A. Delorme and S. Makeig, "EEGLAB: An open source toolbox for analysis of single-trial EEG dynamics including independent component analysis," *J. Neurosci. Methods*, vol. 134, no. 1, pp. 9–21, Mar. 2004.

[46] N. Bigdely-Shamlo, T. Mullen, C. Kothe, K.-M. Su, and K. A. Robbins, "The PREP pipeline: standardized preprocessing for large-scale EEG analysis," *Front. Neuroinform.*, vol. 9, p. 16, 2015.

[47] C. Chang, S. Member, S. Hsu, and S. Member, "Evaluation of Artifact Subspace Reconstruction for Automatic EEG Artifact Evaluation of Artifact Subspace

Reconstruction for Automatic EEG Artifact Removal," no. June, 2018.

[48] M. B. Pontifex, V. Miskovic, and S. Laszlo, "Evaluating the efficacy of fully automated approaches for the selection of eye blink ICA components," *Psychophysiology*, vol. 54, no. 5, pp. 780–791, 2018.

[49] G. King and L. Zeng, "Logistic Regression in Rare Events Data," *Polit. Anal.*, vol. 9, no. 2, pp. 137–163, 2001.

[50] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*. Monterey, CA: Wadsworth and Brooks, 1984.

[51] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," pp. 1–15, 2014.

[52] K. H. Brodersen, C. S. Ong, K. E. Stephan, and J. M. Buhmann, "The balanced accuracy and its posterior distribution," *Proc. - Int. Conf. Pattern Recognit.*, pp. 3121–3124, 2010.

[53] M. L McHugh, "Interrater reliability: The kappa statistic," *Biochem. Medica*, vol. 22, pp. 276–282, 2012.

[54] D. C. Montgomery, E. A. Peck, and G. G. Vining, *Introduction to Linear Regression Analysis (4th ed.)*. Wiley & Sons, 2006.

[55] S. Seabold and J. Perktold, "Statsmodels : Econometric and Statistical Modeling with Python," *Proc. 9th Python Sci. Conf.*, pp. 57–61, 2010.

[56] E. Maris and R. Oostenveld, "Nonparametric statistical testing of EEG- and MEG-data," *J. Neurosci. Methods*, vol. 164, no. 1, pp. 177–190, 2007.

[57] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *arXiv*, pp. 1–23, 2017.

[58] V. Nacher, A. Ledberg, G. Deco, and R. Romo, “Coherent delta-band oscillations between cortical areas correlate with decision making,” *Proc. Natl. Acad. Sci.*, vol. 110, no. 37, pp. 15085–15090, 2013.

[59] J. Jacobs, G. Hwang, T. Curran, and M. J. Kahana, “EEG Oscillations and Recognition Memory: Theta Correlates of Memory Retrieval and Decision Making,” *Neuroimage*, vol. 32, no. 2, pp. 978–987, 2006.

[60] S. D. Muthukumaraswamy, “High-frequency brain activity and muscle artifacts in MEG/EEG: a review and recommendations,” *Front. Hum. Neurosci.*, vol. 7, no. April, pp. 1–11, 2013.

[61] L. Hintze and R. D. Nelson, “Violin Plots: A Box Plot-Density Trace Synergism,” *Am. Stat.*, vol. 52, no. 2, pp. 181–184, 2010.

[62] J. D. Olden and D. A. Jackson, “Illuminating the ‘“ black box ”’: a randomization approach for understanding variable contributions in artificial neural networks,” *Ecol. Modell.*, vol. 154, pp. 135–150, 2002.

[63] J. D. Olden, M. K. Joy, and R. G. Death, “An accurate comparison of methods for quantifying variable importance in artificial neural networks using simulated data,” *Ecol. Modell.*, vol. 178, pp. 389–397, 2004.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

The public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of the collection of information, including suggestions for reducing this burden to Department of Defense, Washington Headquarters Services, Directorate for Information Operations and Reports (0704-0188), 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302. Respondents should be aware that notwithstanding any other provision of law, no person shall be subject to an penalty for failing to comply with a collection of information if it does not display a currently valid OMB control number.
PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
22-03-2019	Master's Thesis	September 2017 – March 2019

TITLE AND SUBTITLE
Confidence Inference in Defensive Cyber Operator Decision Making

5a. CONTRACT NUMBER
F4FBGN7340J001

5b. GRANT NUMBER
JON#18G147B

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Ganitano, Graig S., Captain, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)
Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way, Building 640
WPAFB OH 45433-8865

8. PERFORMING ORGANIZATION REPORT NUMBER
AFIT-ENG-MS-19-M-028

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
711th Human Performance Wing
2610 Seventh St Bldg 441
WPAFB, OH 45433
937-255-8222
rajesh.naik@us.af.mil
ATTN: Dr. Rajesh Naik

10. SPONSOR/MONITOR'S ACRONYM(S)
711 HPW/CL

11. SPONSOR/MONITOR'S REPORT NUMBER(S)

12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRUBTION STATEMENT A. APPROVED FOR PUBLIC RELEASE; DISTRIBUTION UNLIMITED.

13. SUPPLEMENTARY NOTES
This material is declared a work of the U.S. Government and is not subject to copyright protection in the United States.

14. ABSTRACT

Cyber defense analysts face the challenge of validating machine generated alerts regarding network-based security threats. Operations tempo and systematic manpower issues have increased the importance of these individual analyst decisions, since they typically are not reviewed or changed. Analysts may not always be confident in their decisions. If confidence can be accurately assessed, then analyst decisions made under low-confidence can be independently reviewed and analysts can be offered decision assistance or additional training. This work investigates the utility of using neurophysiological and behavioral correlates of decision confidence to train machine learning models to infer confidence in analyst decisions. Electroencephalography (EEG) and behavioral data was collected from eight participants in a two-task human-subject experiment and used to fit several popular classifiers. Results suggest that for simple decisions, it is possible to classify analyst decision confidence using EEG signals. However, more work is required to evaluate the utility of EEG signals for classification of decision confidence in complex decisions.

15. SUBJECT TERMS
decision confidence, Cyber Intruder Alert Testbed (CIAT), machine learning

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Dr. Brett J. Borghetti, AFIT/ENG
U	U	U	UU	142	19b. TELEPHONE NUMBER *(Include area code)* (937) 255-6565, ext 4612 brett.borghetti@afit.edu

Standard Form 298 (Rev. 8-98)
Prescribed by ANSI Std. Z39-18