CROSS-SUBJECT CONTINUOUS ANALYTIC WORKLOAD PROFILING USING STOCHASTIC DISCRETE EVENT SIMULATION

THESIS

Joseph J. Giametta, Captain, USAF

AFIT-ENG-MS-16-M-018

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENG-MS-16-M-018

CROSS-SUBJECT CONTINUOUS ANALYTIC WORKLOAD PROFILING USING STOCHASTIC DISCRETE EVENT SIMULATION

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 Computer Science

Joseph J. Giametta, BS

Captain, USAF

March 2016

AFIT-ENG-MS-16-M-018

CROSS-SUBJECT CONTINUOUS ANALYTIC WORKLOAD PROFILING USING STOCHASTIC DISCRETE EVENT SIMULATION

Joseph J. Giametta, BS

Captain, USAF

Committee Membership:

Dr. B. J. Borghetti
Chair

Maj C. F. Rusnock, PhD
Member

Maj B. G. Woolley, PhD
Member

AFIT-ENG-MS-16-M-018

Abstract

Operator functional state (OFS) in remotely piloted aircraft (RPA) simulations is modeled using electroencephalograph (EEG) physiological data and continuous analytic workload profiles (CAWPs). A framework is proposed that provides solutions to the limitations that stem from lengthy training data collection and labeling techniques associated with generating CAWPs for multiple operators/trials. The framework focuses on the creation of scalable machine learning models using two generalization methods: 1) the stochastic generation of CAWPs and 2) the use of cross-subject physiological training data to calibrate machine learning models. Cross-subject workload models are used to infer OFS on new subjects, reducing the need to collect truth data or train individualized workload models for unseen operators. Additionally, stochastic techniques are used to generate representative workload profiles using a limited number of training observations. Both methods are found to reduce data collection requirements at the cost of machine learning prediction quality. The costs in quality are considered acceptable due to drastic reductions in machine learning model calibration time for future operators.

Acknowledgments

I would like to express my sincere appreciation to my wife and children for dealing with the many days and nights that were lost in order complete this thesis. I love you all…

Jay/Dad

Table of Contents

Page

List of Figures

Page

List of Tables

Page

CROSS-SUBJECT CONTINUOUS ANALYTIC WORKLOAD PROFILING USING DISCRETE EVENT SIMULATION

I. Introduction

Modern military forces rely heavily on the support of remotely piloted aircraft (RPA) for successful completion of combat operations worldwide. These aircraft provide critical intelligence, surveillance, and reconnaissance (ISR) capabilities to battlefield commanders without the need for a pilot in the cockpit. Human operators man flight controls and monitor aircraft vitals from distances that can potentially span the entire globe. The growing demand for unmanned assets coupled with current military manning constraints has created an environment in which RPA operators are being pushed to perform beyond their individual capabilities. This environment has led to research thrusts that focus on augmenting operator performance using computerized aiding techniques. Unfortunately computerized aiding cannot be implemented without several considerations. The physical separation of pilot and aircraft, coupled with an increasing reliance on automated navigation systems has the potential to lead to "automation deficits", i.e. decreases in: 1) situation awareness; 2) system awareness; and 3) manual flight skills in human operators (Parasuraman, Bahri, Deaton, Morrison, & Barnes, 1992). It has been proposed that an adaptive aiding approach, in which assistance is dynamically provided based on operator need (Rouse, 1988), has the potential to reduce these negative effects, while increasing the benefits offered by automation. This document describes a physiological-based assessment technique that can be used to estimate operator functional state (OFS), then, if necessary, trigger computerized augmentation to avoid mission degradation caused by operator overload (or underload).

1.1. Problem Statement

Current physiological-based OFS estimation techniques rely heavily on the calibration of supervised machine learning models. Model calibration, or training, is often operator specific, and requires labeled activity data to link physiological responses (e.g. electroencephalogram (EEG), electrocardiogram (ECG), and electrooculogram (EOG)) to unique functional states. A commonly used labeling technique, as noted by Rusnock, Borghetti, & McQuaid (2015), is the use of cumulative, subjective workload measures to assign broad, task difficulty values to extended periods of operator activity (Hart & Staveland, 1988). The use of these cumulative, subjective measures has limited the amount of detail that can be provided by supervised machine learning methods, especially when attempting to infer OFS across multiple subjects. Recently, Rusnock et al. (2015) described a method for creating continuous analytic workload profiles (CAWPs) using Discrete Event Simulation (DES) that allows us to study the effects of second-by-second workload changes on physiological state throughout complex multi-objective tasks. Unfortunately, labeling activity data in complex tasks such as RPA operation is often time consuming, requiring extensive human analysis from a subject matter expert (SME). Considering the fact that supervised learning models often require many training observations prior to successful calibration, this method becomes infeasible when models for several operators are required.

1.2. Research Questions

This research effort focuses on creating scalable machine learning models using two generalization methods. The first method utilizes physiological data from multiple

previously observed operators to estimate OFS for unseen operators (cross-subject models). The second uses generalized, distribution-based representations of operator behavior rather than exact second-by-second data to train machine learning models (stochastic models), reducing the number of observations needed for model calibration. The following questions explore the effects of each generalization method:

Q1. Is there a significant performance difference between machine learning models fitted using cross-subject, rather than within-subject physiological data?

Q2. Is there a significant performance difference between machine learning models fitted using stochastic, rather than deterministic CAWPs?

1.3. Assumptions/Limitations

Mental workload is highly sensitive to individual differences in operator skill level, cognitive capabilities, and individual effort. The assignment of objective mental workload values based on subject observations is expected to limit the accuracy of the models under study. Data for this research was provided by a human subject experiment performed by an external organization. It is assumed that the human subjects involved in this research activity were trained to a stable skill level prior to data collection and that learning effects were minimal across the trials. Furthermore, it is expected that each subject gave maximum effort during the completion of each of his or her assigned tasks. It is also expected that actual workload transition times and those recorded by researchers

may vary up to one second per observation due to limited video recording capabilities. Additionally, it is assumed that deviations in recording times did not cause a significant decrease in model accuracy.

1.4. Contributions

The proposed methods provided potential solutions to the limitations that stem from lengthy training data collection and labeling techniques associated with generating CAWPs for multiple operators/trials. Measuring machine learning model performance on unseen data allowed us to compare the effectiveness of models fit using different physiological readings (cross-subject or within-subject) and differently-generated CAWPs (stochastic or deterministic). It was assumed that a lack of statistical difference in performance between cross-subject and within-subject machine learning models would imply that once "trained", group workload models could be used to infer OFS on new subjects, reducing the need to collect truth data or train individualized workload models for new subjects. It was also assumed that a lack of statistical difference in performance between machine learning models utilizing stochastic and deterministic modeling techniques would suggest that stochastic techniques could be used to create representative workload profiles using a limited number of training observations.

Both cross-subject machine learning techniques and stochastic workload profiling methods were found to significantly reduce machine learning model performance. Post-hoc analysis showed that even though both techniques resulted in poorer quality machine learning models, they still produced meaningful estimations of OFS. This

demonstrated a reduced need to collect new training observations for future subjects performing identical tasks at the cost of an acceptable decrease in model fidelity.

1.5. Overview

This document is composed of five chapters. Chapter II presents a review of current research focused on inferring mental workload from EEG using machine learning techniques. Chapter III describes the data collection process; the production of CAWPs using DES; and the testing and training of machine learning models. Chapter IV details the performance of both, cross and within subject, as well as, stochastic and deterministic machine learning models. Lastly, Chapter V provides discussion, conclusions, and the potential for future work related to this research.

II. Literature Review

In a highly influential article, Byrne & Parasuraman (1996) described two pillars in adaptive automation research: 1) providing information about the effects of different forms of automation and 2) providing information about physiological measures that can be used to measure operator mental state, and in turn, regulate automation levels. The researchers believed that real-time assessment capabilities provided by physiological measures were a distinct advantage when compared to other methods such as subjective workload ratings. They described a theoretical framework for regulating the delivery of automation, based on these continuous assessments that could be used to optimize human-machine interactions. The realization of their framework was dependent on the identification of valid and reliable physiological workload measures by future researchers. This chapter provides a review of core concepts in mental workload theory, subjective workload measurement, and electroencephalograph (EEG) based workload estimation, then concludes with an analysis of current research pertaining to continuous workload modeling techniques.

2.1. Mental Workload Theory

Mental workload is defined as "the relation between the (quantitative) demand for resources imposed by a task and the ability to supply those resources by the operator" (Wickens, 2002). Mental workload in operators has been suggested to have a non-linear, inverted U-shaped relationship with task performance (Cassenti & Kelley, 2006), where either too little or too much workload results in decreases in performance. Over the years, mental workload theory has attempted to explain the relationship between

workload and performance in terms of resource accessibility. Welford (1967) hypothesized that mental resources were accessed serially, and that bottlenecks caused by previously queued decision processes led to performance decrement. Wickens (2002) explained operator workload using a multi-dimensional model and believed that specific mental resources could be used in parallel, but that overuse of shared processing stages, perceptual modalities, visual channels, or processing codes could lead to resource interference and decreases in task performance. It would then appear to follow that by decreasing resource demand, performance could be increased. While this seems to be the case in situations involving operator overload, decreases in operator workload that result in underload have also been shown to negatively impact task performance (Young & Stanton, 2002). Their findings reinforce the need for reliable measures of operator functional state (OFS) when employing adaptive aiding techniques, due to differences in cognitive capabilities between individual operators.

2.2. Subjective Workload Measurement

Subjective, self-assessment methods could, quite possibly, be the most reliable of all workload measures, because they are "scored" directly by the subject under study. These assessments often take the form of self-report questionnaires such as the NASA task load index (NASA-TLX) (Hart & Staveland, 1988) and Subjective Workload Assessment Technique (SWAT) (Reid & Nygren, 1988).

The NASA-TLX was the most commonly used subjective measure in the literature that was surveyed for this research activity. (Hart & Staveland, 1988) created the multi-dimensional rating scale over several years of laboratory studies involving

simple manual control tasks, complex supervisory control tasks, and aircraft simulations. The NASA-TLX requires subjects to rate task demands on six scales ranging from 0 to 100 in increments of five, then to prioritize each of the scales from greatest to least importance, based on which the subject felt were most applicable to the given task. The task demand ratings selected for each scale allow researchers to understand how difficult a task is perceived to be, and the prioritization of scales also gives them insight into which resources are most important to the rater.

The Air Force Research Laboratory (AFRL) (Funke et al., 2013) recently conducted an evaluation of five subject workload techniques: the NASA-TLX, Workload Profile (WP) (Tsange & Velazquez, 1996), Multiple Resources Questionnaire (MRQ) (Boles & Adair, 2001), and Subjective Workload Dominance (SWORD) Technique (Vidulich, Ward, & Schueren, 1991). The group used a space based video game similar to Atari's Asteroids to measure differences among subjective measures when compared to subject task performance in three levels of task difficulty (10, 15, and 20 asteroids). NASA-TLX, MRQ, and WP ratings showed moderate correlations with one another ranging from 0.251 to 0.437, but of the three, WP was found to be the best indicator of subject performance with a correlation of -0.195.

WP is similar to another subjective measure, Workload Index (W/INDEX) (North & Riley, 1989), in that both are rooted in Multiple Resource Theory (MRT) (Wickens, 2002). Both approaches rate mental workload based on shared processing stages, perceptual modalities, visual channels, or processing codes. The WP differs from W/INDEX in that it is meant to be used as a post-hoc assessment of workload provided by the research subject, rather than a predictive tool used by the researcher. The post-hoc

nature of WP seems to explain the findings of Funke et al. (2013), due to subjects experiencing all task difficulty levels prior to assigning ratings to each one. While the technique appears to be valid for small sample sizes (3 trials per subject), one could question the validity of cumulative scores that require recalling task difficulty from hours or days earlier in extended studies. Unfortunately, this cumulative nature, as well as the intrusiveness of completing these questionnaires often makes subjective measures unsuitable for continuous measurement of OFS.

2.3. EEG Based Workload Estimation

As previously discussed, Byrne & Parasuraman (1996) believed that physiological features had the potential to measure OFS continuously, in real-time. Physiological features are often measured through the use of electro-biological methods, such as the EEG. The EEG measures electrical potentials in the scalp that are generated when masses of neurons in the brain are activated (Teplan, 2002). Through massive amplification, these potentials can be observed at the microvolt level, providing insight into underlying brain activity. The international 10-20 electrode system (Jasper, 1958) shown in Figure 2.1 is often used to ensure standardized placement of measurement equipment. Electrodes are placed over the (F)rontal, (T)emporal, (C)entral, (P)arietal, and (O)ccipital lobes and are expected to record electrical activity originating from each specific area of the brain, but the spatial resolution of the EEG is known to be limited by the depth of the originating electrical signal (Dale & Sereno, 1993).

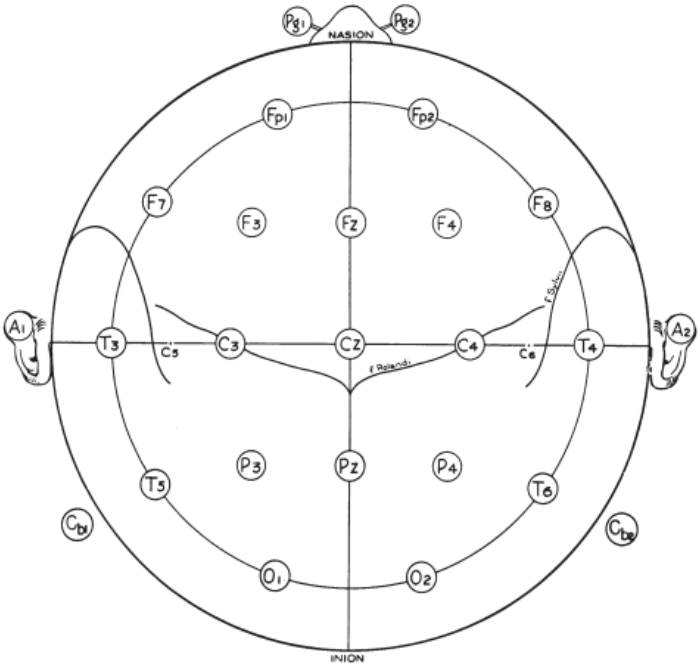

Figure 2.1. The 10-20 International Electrode System (Klem, Lüders, Jasper, & Elger, 1999)

The EEG makes up for limited spatial resolution with excellent resolution in the time domain. Time resolution in EEG recordings is often limited only by the sampling frequency of recording equipment. The timing of EEG oscillations have been linked to mental workload since Berger (1929) created the first recording over half of a century ago. Common interpretations of oscillation frequency are often linked to specific bands of interest shown in Table 2.1.

Table 2.1. EEG Frequency Bands (Ochoa, 2002)

Band	***Range***	***Common Associations***
Delta	0.5-4 Hz	Deep sleep; Eye and muscle related artifacts
Theta	4-7 Hz	Emotional Stress; Creative Inspiration; Meditation
Alpha	8-13 Hz	Empty mind; Closed eyes
Beta	13-30 Hz	Active thinking; Attention; Problem solving
Gamma	35 Hz and higher	Blending of multiple brain functions; Muscle related artifacts

A common trend in recent studies is the use of machine learning algorithms to link EEG spatial/spectral features to operator performance and task difficulty. Wilson & Russell (2007) provided adaptive aiding in real-time using physiological measures during a complex aerial attack simulation. Operators were required to monitor the status of four autonomous vehicles as they flew pre-planned bombing missions. As vehicles progressed along bombing routes, radar images of attack sites became available for download. Attack targets were required to be marked on images prior to UAV arrival at the corresponding locations. Unmarked/incorrectly marked targets reduced mission effectiveness.

Electrocardiogram (ECG), EEG, and electrooculogram (EOG) were collected throughout these missions and were broken into ten-second epochs. Artificial Neural Network (ANN) classifiers were then trained on labeled physiological data from pre-accomplished training missions for each operator. Individual difficulty thresholds as well as group (high performer/low performer) difficulty thresholds were calculated. Adaptive aiding was then provided based on task difficulty as determined by the classifiers utilizing these thresholds. Aiding efforts were accomplished by reducing UAV speed and displaying informative vehicle status information to operators. Reported classifier accuracy, when distinguishing between high and low difficulty task conditions, was 83.6% for tasks using individual vehicle speed thresholds and 75.5% for those using group vehicle speed thresholds.

In another study, Hogervorst, Brouwer, & Erp (2014) sought to examine classification accuracy of three "sensor groups" of physiological variables that were similar to those used by Wilson & Russell (2007). The sensor groups were defined as:

EEG (event related potentials (ERP), spectral power features), Physiology (skin conductance level (SCL), respiration rate, ECG), and Eye (pupil size, blink rate).

Physiological data was measured while research participants completed n-back memory tasks (0, 1, and 2 back) (Kirchner, 1958). Data was then combined into sensor groups; partitioned into epochs following each task; broken into test and training sets; then used to train two classification models: a Support Vector Machine (SVM) and an elastic net with logistic regression. Classifiers trained on each sensor group were benchmarked according to their accuracy in classifying 0-back versus 2-back epochs.

Results showed that classifiers trained on EEG data alone reached accuracy rates of nearly 86%, while those from the Physiological only and Eye only groups fell within the range of 70% to 75%. The authors went on to show that combining features from each group did not result in significant gains in accuracy over EEG alone (the highest combined accuracy being EEG and Eye at 89%). When sensor groups were partitioned further and ERP data taken from only the Pz electrode of the EEG was used, classification accuracy from the elastic net was reported at 88%. These findings strongly implied that larger feature sets do not necessarily result in better workload estimation. This concept was a driving factor in the decision to use only EEG physiological features for the current study.

2.4. Continuous Workload Modeling

Analytical workload estimation tools, such as the Army Research Laboratory's Improved Performance Research Integration Tool (IMPRINT) (Archer & Adkins, 1999) have been proposed as a continuous alternative to physiological and subjective workload

measures, which are limited by intermittent updates (Rusnock et al., 2015). IMPRINT allows researchers to model workload using both distribution-based (stochastic) and scripted (deterministic) discrete event simulation (DES) techniques. The simulations generate second-by-second workload profiles using researcher defined activities and completion times. During model development researchers assign Visual, Auditory, Cognitive, and Psychomotor (VACP) (Aldrich, Szabo, & Bierbaum, 1989) difficulty ratings to each activity. Task times and activity branching logic are then determined using either stochastic probabilities and distributions or statically defined deterministic variables.

In another study, Rusnock & Geiger (2014) used stochastic IMPRINT DES to evaluate the effects of task difficulty on alternative interface designs in unmanned ground vehicle surveillance simulations. The authors were able to generate representative task times and branching logic for four alternative interface designs, under three levels of difficulty, using distributions created from 150 test subjects. After running each simulation ten times, the effects of task difficulty on each of the interface designs were able to be compared. The continuous profiles enabled the researchers to determine mean differences in VACP workload among interface designs, and also allowed for a better understanding of workload variability within each difficulty/interface pair.

Later, Smith, Borghetti, & Rusnock (2015) utilized IMPRINT in their effort to compare the cross-applicability of physiological based regression tree and random forest machine learning models. The group analyzed workload changes in two different tasks involving remotely piloted aircraft (RPA) simulations. Unlike previously discussed machine learning research, the group used deterministic DES to generate workload truth

data for each of the simulations. Rather than training machine learning models to classify task difficulty based on predefined labels (e.g. low, medium, high), IMPRINT-generated CAWPs were used to calibrate regression models that estimated second-by-second user activity on a continuous VACP scale.

The researchers emphasized the need for models that could be reused for multiple tasks or individuals in real world operational scenarios. They stated that in these scenarios, when models do not generalize well across multiple task conditions or individuals, exhaustive sets of models must be generated that cover all operational task condition/subject combinations. To address the importance of cross-applicability, full regression tree, pruned regression tree, and random forest models were compared in their ability to estimate workload across tasks and subjects. After comparing each model using cross-validated root mean squared error (RMSE), the group concluded that random forest models provided the best performance across each of the tested contexts.

2.5. Summary

This chapter has reviewed core concepts in adaptive aiding, mental workload theory, subjective workload measures, and EEG recording. Recent studies utilizing continuous workload measures were also discussed. The review showed that subjective workload measures, such as the NASA-TLX, provide reliable estimates of operator task performance, but fail to meet the continuous assessment requirements needed for adaptive aiding. Alternatively, EEG-based machine learning models were shown to accurately provide continuous estimates of task difficulty, but the models lacked the detail provided by subjective assessments. Research presented by Smith et al. (2015)

attempted to increase the detail provided by EEG based machine learning models by fitting models using deterministic CAWPs, but did not explore the usefulness of stochastic CAWPs demonstrated by Rusnock & Geiger (2014). These findings, along with the core concepts described earlier in the chapter led to my belief that training machine learning models using stochastic CAWPs and cross-subject physiological data would drastically reduce data collection requirements for EEG-based OFS estimation without significantly reducing model performance. The findings and core concepts also created the basis for the research methodology described next.

III. Methodology

The main objective of this research was to effectively model the operator functional state (OFS) of remotely piloted aircraft (RPA) operators. The study described in this chapter utilized a dynamic aerial surveillance environment to simulate real-world RPA operations. Supervised machine learning models were trained using electroencephalograph (EEG) physiological data and continuous analytic workload profiles (CAWPs). The viability of stochastic/deterministic CAWPs as well as cross-subject/within-subject physiological training data were compared, based on machine-learning model performance.

In this section the following research questions are explored: Q1.) Is there a significant performance difference between machine learning models fitted using cross-subject, rather than within-subject physiological data? and Q2.) Is there a significant performance difference between machine learning models fitted using stochastic, rather than deterministic CAWPs? An additional verification question will also be explored in order to ensure that DES was successful: Do CAWPs created using deterministic DES correlate with cumulative subjective task load ratings and follow similar distributions?

Two alternative research hypotheses are also tested:

Hypothesis 3.1. Based on the bias vs. variance tradeoff described by James, Witten, Hastie, & Tibshirani (2013), it is believed that "smoothed" stochastic CAWPs will reduce variance in cross-subject machine learning models and provide superior generalization when compared to deterministic CAWPs.

Hypothesis 3.2. Based on the overwhelming success of non-parametric, non-linear machine learning models (i.e. artificial neural networks and support vector machines) in related EEG-based classification, it is expected that random forest (RDF) regression will outperform linear regression (LM) models when used to infer OFS.

These questions and hypotheses directly contribute to the research thrust described in Rusnock et al. (2015): to enhance physiological computing and neuroergonomic research, through the use of CAWPs. Mapping relationships between these continuous profiles and operators' physiological states using machine learning enables "indirect estimations of workload in real-time" that are necessary for useable adaptive aiding.

3.1. Domain of Study

This experiment used existing data which was the result of a separate study conducted by Air Force Research Laboratory's (AFRL) 711th Human Performance Wing. Their study aimed to replicate a high-stress, dynamic, military surveillance scenario in which individual performance and mental workload could vary in real-time based on operator capabilities. This task environment represented a significant step towards simulating the complexities of real-world activities by mirroring the highly dynamic nature of realistic military operations. Physiological data and video footage from the experiments were used to craft CAWPs and evaluate the concepts described in the previous sections.

3.1.1. Participants

Twelve individuals volunteered to participate in the study. Participants were between 22 and 46 years old (mean age 25.66), four female and eight male. Unfortunately, complete video footage was only available for six subjects for analysis in this thesis. All participants were right-handed and reported normal or corrected-to-normal vision. Each participant was compensated $15 per hour for their involvement in the study. The research activity was approved and conducted in accordance with AFRL Institutional Review Board guidelines.

3.1.2. Task Environment

The Air Force Vigilant Spirit Control Station (VSCS) environment was used during the AFRL study to simulate the control of multiple, semi-autonomous RPAs performing intelligence, surveillance, and reconnaissance (ISR) over-flight. The primary task, surveillance, was a visual search based task, in which subjects were required to pan and zoom a RPA camera in order to locate pedestrian targets that matched a predefined set of characteristics, while searching a medium sized geographical area. A high-value target from the surveillance task is shown in Figure 3.1.

Figure 3.1. A high-value target in the primary surveillance task

Additional trials were completed using an alternative primary task that is not included in the current research effort. A short duration secondary task was presented periodically during the continuous execution of each primary task. The secondary task simulated audio communication over a multi-modal communication tool (MMC) via live radio call and text messaging. In order to complete the secondary task, participants were required to answer distance, speed, and altitude related questions involving basic multiplication, division, and addition operations. Table 3.1 details the timing data for system event during surveillance tasks in the VSCS environment.

Table 3.1. Surveillance Scenario Experimental Script

State	*Description*	*Start Time*	*Finish Time*	*Total (sec)*
Trial Start	Timing begins	0	0	0
Target Idle 1	The time period before HVT 1 appears	0	9	9
Radio Idle 1	The time period before Radio Call 1 is heard	0	30	30
HVT 1	HVT 1 is on screen	9	59	50
Radio Call 1	Radio Call 1 is heard	30	35	5
Radio Idle 2	The time period before Radio Call 2 is heard	35	90	55
Target Idle 2	The time period before HVT 2 appears	59	69	10
HVT 2	HVT 2 is on screen	69	119	50
Radio Call 2	Radio Call 2 is heard	90	95	5
Radio Idle 3	The time period before Radio Call 3 is heard	95	150	55
Target Idle 3	The time period before HVT 3 appears	119	129	10
HVT 3	HVT 3 is on screen	129	179	50
Radio Call 3	Radio Call 3 is heard	150	155	5
Radio Idle 4	The time period before Radio Call 4 is heard	155	225	70
Target Idle 4	The time period before HVT 4 appears	179	204	25
HVT 4	HVT 4 is on screen	204	254	50
Radio Call 4	Radio Call 4 is heard	225	230	5
Radio Idle 5	The time period after the last Radio Call is heard	230	254	24

Task difficulty in the surveillance tasks was varied using two binary conditions: fuzz and distractors. Fuzz was toggled on or off, and affected the clarity of the RPA video feed being observed by each operator. Distractors were set to high or low and determined the number of non-HVT pedestrians that were present in the operators' assigned search area.

3.1.3. Study Design

Participants completed four blocks of 15 minute trials over the course of four sessions, resulting in 16 total trials for each primary task (four repetitions of each of the four condition combinations). At the beginning of each block, subjects performed four minutes of a trial activity followed by a three minute NASA-TLX questionnaire period. Afterwards, another five minutes of trial activity along with an additional three minute NASA-TLX period were completed. Rest periods were given between each trial (five minutes between trials 1-2 and 3-4, and 15 minutes between trials 2-3). Including simulation and physiological equipment setup, participants spent roughly 120 minutes in the lab each session.

3.1.4. Data Collection

Video footage and physiological data were collected continuously for each trial throughout AFRL's study. Their physiological data eventually served as the set of independent variables (observations) used by machine learning models in this research activity to infer OFS. Video footage was used to develop IMPRINT models which were executed to output CAWPs that represented operator workload during each trial.

3.1.4.1. Physiological Data

EEG and EOG signals were sampled at 480 Hz using a CleveMed BioRadio 150. Electrodes placed above and below the right eye collected vertical EOG (VEOG) data. Horizontal EOG (HEOG) data was measured by electrodes placed to the left of the left eye and right of the right eye. EEG was collected by means of a BioSemi ActiveTwo electrode skullcap with Ag-AgCl electrodes placed at F7, F8, Fz, O2, Pz, T3, and T4 according to the international 10-20 electrode system (Jasper, 1958). All EEG signals were referenced to the right mastoid. The left mastoid was used as a ground source to prevent line noise on the BioRadio 150 User Units.

3.1.4.2. Video Footage

Video footage was recorded from six different sources during each trial. The footage captured time stamped aircraft information, RPA camera output, and subject activity. Descriptions for each video are shown in Table 3.2.

Table 3.2. AFRL Video Details

Video	***Description***
1	RPA Screen #1 (Aircraft Information)
2	RPA Screen #2 (Camera View of Enemy Territory)
3	Message Console (Communications Requests/Responses)
4	Simulation Status (Physio Sensor Status/Trial Activity)
5	Subject Assessment (Estimated Workload/Performance)
6	Facial Monitoring (Displays Subject Focal Point/Movement)

3.2. Physiological Feature Extraction

EEG data were filtered at 0.2 Hz high-pass and 40 Hz low-pass using a third order Butterworth filter. A 40 Hz low-pass filter was chosen to avoid muscle related noise

stemming from unconstrained subject movement. Data segments used in this study were limited to the first 171 seconds for each trial. The remaining 83 seconds were removed due to EEG disruption caused from biomarker collection (i.e. oral swabbing) at the end of each trial. Eye related artifacts were removed from EEG signals using EOG signals and the Independent Component Analysis (ICA) method described by Jung et al. (2000). The short-time Fourier transform (STFT) was used to extract time-frequency features from five frequency bands (delta (1-3 Hz), theta (4-7 Hz), alpha (8-12 Hz), beta (13-30 Hz), and gamma (31-40 Hz)) at each of the seven EEG electrodes resulting in a total of 35 spectral features for each trial. The STFT was performed using the `stft()` function from the e1071 0.4-7 R package. A Hanning window of ten seconds with a one second increment was used. 240 Fourier coefficients (one coefficient per Hz, up to the Nyquist frequency of 240 Hz) were calculated for each transform. Lastly, the mean power in each frequency band was log transformed and converted to decibels.

3.3. Continuous Analytic Workload Profiling

Video footage recorded during the AFRL study was encoded for post-hoc DES in order to create CAWPs. Both stochastic and deterministic CAWPs were created from the encoded video footage using the Army Research Laboratory's IMPRINT DES software (Archer & Adkins, 1999). Prior to working with IMPRINT, a cognitive task analysis (CTA) was completed for the VSCS surveillance task to ensure proper modeling of subject activity.

3.3.1. Cognitive Task Analysis

Applied CTA (Militello & Hutton, 1998) was used to describe the cognitive elements of the VSCS surveillance task through the use of task diagrams, knowledge auditing, and simulated expert interviews. Figure 3.2 illustrates primary and secondary task sequences that were created to capture the major activities encountered during VSCS surveillance. After decomposing the task into major activities, a knowledge audit (Appendix A) was performed to understand the expertise needed for task completion. Next, a simulation interview, in which a subject matter expert (SME) was asked to explain their mental process when performing the task, was completed to ensure that all major objectives were covered (Appendix B). Finally, a cognitive demands table (Table 3.3) was created to describe common difficulties, cues, and strategies encountered during the surveillance task.

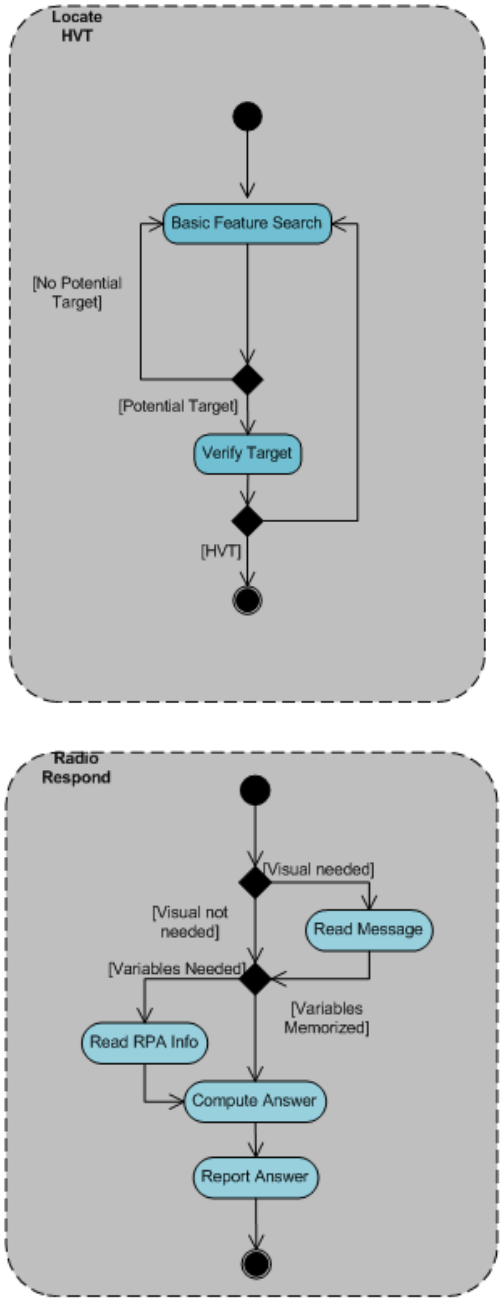

Figure 3.2. Surveillance Task Diagram

Table 3.3. VSCS Surveillance Task Cognitive Demands Table

Cognitive Element	*Reason for Difficulty*	*Common Errors*	*Cues and Strategies*
Basic Feature Searching	Transitioning between zones that do not contain identifiable landmarks becomes difficult due to a continuously changing camera perspective	Following an inefficient scan path that does not allow for easy transitions between zones when scanning	Break the area of interest up into zones that can be easily identified
	Targets are only visible for a short amount of time. Prioritizing high traffic areas first, gives better odds of locating a target early	Moving slowly and missing potential targets that leave the scene	Prioritize high traffic areas (targets rarely remain stationary)
	Scanning with an improper zoom level leads to either a slow scan or missed details	Moving quickly and not recognizing potential targets	Scan at an appropriate zoom level
	Slow scanning increases the chances of a target entering an area that was previously scanned	Incorrectly identifying potential targets	Scan thoroughly, but quickly
	Many potential targets wear similar clothing, and carry items easily mistaken as weapons		Prioritize by target clothing, posture, and potential weapons
			Mobile individuals, carrying large objects are of greatest interest
Target Verification	Excessive focus on incorrect targets reduces scene awareness and increases the chances of losing position along scan route	Many potential targets may carry large tools instead of weapons	Zoom only as far as necessary to verify potential targets
		Spending too long focusing on an incorrect target	Individuals carrying rifles are high value targets
		Losing scene awareness due to improper use of zoom	
Target Tracking	Targets unexpectedly change directions or temporarily move out of sight	Target loss due to unforeseen blind spot	Pay close attention to target movement in crowded areas and estimate potential movements
			Be aware of camera movement that will result in target visibility being lost
Computing Radio Responses	Diverting attention from target location/tracking in order to view message traffic or aircraft information increases the risk of target loss	Loss of target due to use of text messaging or information lookup	Closely monitor radio traffic
			Memorize aircraft velocity and altitude
			Information is not requested if a target is not present

3.3.2. Video Encoding

Video footage was broken into eleven events that were annotated throughout each trial. Three separate videos were used to collect the required data. Event times were recorded using system time (rounded to the nearest second) for the target video source. The eleven events are shown in Table 3.4.

Table 3.4. Video Encoding Events

Input Data Name	Description	Video
potentialTarget	The time in seconds at which each potential target was identified.	2
verifyTarget	The time in seconds at which the identity of a potential target was verified.	2
targetFound	An indicator of whether or not a HVT was located.	2
targetLost	The time in seconds at which a HVT leaves view of the camera.	2
reportComm	The time at which a subject begins processing the answer to a radio communication.	3
readMsgNeed	An indicator of whether or not a subject needs to read the current radio message.	6
readMsg	The time at which a subject finishes reading a communications message.	6
readRPAInfoNeed	An indicator of whether or not a subject needs to read the current RPA information	6
readRPAInfo	The time at which a subject finishes reading the current RPA information.	6
computeAnswer	The time at which a subject finishes computing the answer to a radio response.	6
reportAnswer	The time at which a subject reports a radio response.	3

3.3.3. Discrete Event Simulation

DES in IMPRINT enabled both deterministic and stochastic modeling of operator workload. All workload models utilized a task network that was based on activity diagrams developed during the CTA. Individual tasks were then assigned Visual, Auditory, Cognitive, and/or Psychomotor workload values (Aldrich et al., 1989; Bierbaum, Szabo, & Aldrich, 1990) using task details provided in the cognitive demands table. The individual tasks and their assigned values are shown in Table 3.5.

Table 3.5. Assigned VACP Workload Values

State	*Visual*	*Auditory*	*Cognitive*	*Psychomotor*	*Overall*
Feature Search	7	0	3.7	2.6	13.3
Verify Target	6.8	0	4	5.8	16.6
Track HVT	4.4	0	1	4.6	10
Monitor Radio	0	1	0	0	1
Process Question	0	4.9	5.3	0	10.2
Read Message	5.9	0	5.3	0	11.2
Read RPA Info	5.9	0	5.3	0	11.2
Compute Answer	0	0	7	0	7
Report Answer	0	0	5.3	3.2	8.5

3.3.3.3. Deterministic Workload Profiles

Deterministic IMPRINT profiles utilized exact task times and branching logic from encoded video footage. Prior to use in IMPRINT, data was formatted using a four step process: 1) absolute time values were converted to relative time offsets based on trial start time; 2) vectors containing timing values were created for each event type; 3) Boolean values created for model branching logic; and 4) zero values were input for events that were not observed. Each of the eleven events were then represented by variables in IMPRINT, then read from arrays containing input values for each subject/trial. Table 3.6 shows an example IMPRINT input file created from video footage. The first four columns of each input file include event timing for potential targets. Remaining columns contain details pertaining to operator communication requests. Each row represents a single target or communications request. Zeros were used to annotate tasks that were not observed. Figure 3.3 shows the estimated VACP workload produced by IMPRINT for the same input file.

Table 3.6. A Deterministic IMPRINT Input File

Potential Target	Verify Target	Target Found	Target Lost	Report Comm	ReadMsg Need	Read Msg	ReadRPA InfoNeed	ReadRPA Info	Compute Answer	Report Answer
9	11	FALSE	0	30	FALSE	0	FALSE	0	37	41
24	27	FALSE	0	90	FALSE	0	FALSE	0	97	103
36	39	FALSE	0	150	FALSE	0	FALSE	0	156	162
47	53	FALSE	0	225	TRUE	233	FALSE	0	234	240
60	61	FALSE	0							
63	65	FALSE	0							
73	75	FALSE	0							
85	91	FALSE	0							
103	104	TRUE	115							
125	126	FALSE	0							
136	137	FALSE	0							
142	144	FALSE	0							
155	156	TRUE	178							
194	195	FALSE	0							
206	207	FALSE	0							
217	221	FALSE	0							
245	247	TRUE	252							

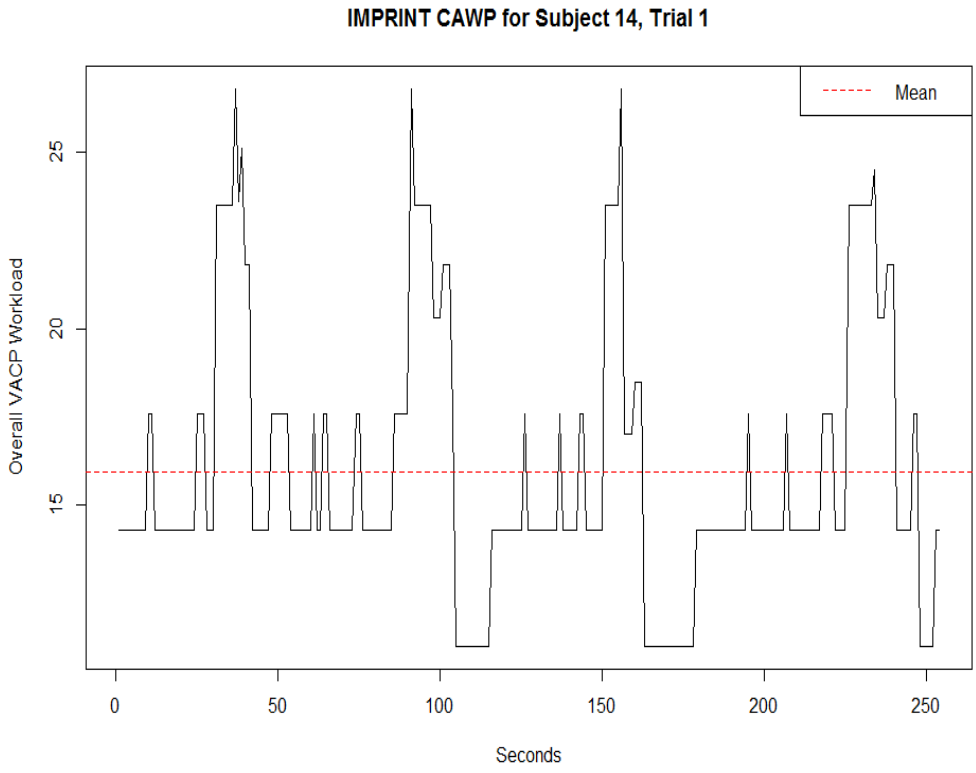

Figure 3.3. A Deterministic Workload Profile

3.3.3.4. Stochastic Workload Profiles

Stochastic IMPRINT profiles utilized the same eleven variables that were needed for deterministic profiles. Rather than reading exact timing from input files, the variables were sampled randomly from distributions fitted to represent the observations collected during the trials. Rockwell's Arena Input Analyzer software was used to fit each of the distributions based on the residual sum of squares (RSS) (Equation 3.1) between histograms of recorded observations and a set of ten commonly used distributions (Appendix C). Figure 3.4 shows a histogram of 300 observations for communications computation time. The red line in the figure shows the distribution of best fit (Beta), based on RSS.

$$RSS = \sum_{i=1}^{n} (y_n - \hat{y}_n) \tag{3.1}$$

where:

n is the number of data points

y_n is the desired output at observation n

$\hat{y}_n$ is the predicated output at observation n

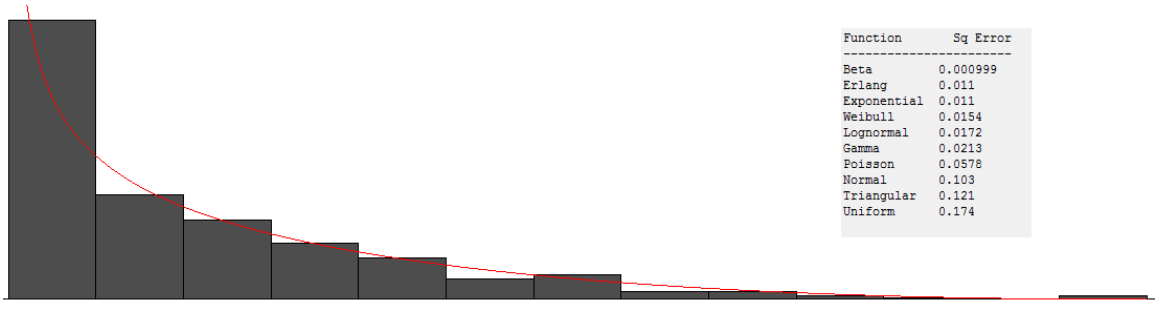

Figure 3.4. Input Analyzer Output for Communications Computation Time

After fitting each of the random variables, ten pilot runs were completed in order to estimate the variance of each model. Time-weighted VACP workload averages over entire 254 second trials were used as single samples. After variance was estimated, the necessary sample size for each model was determined using Equation 3.2 (NIST/SEMATECH e-Handbook of Statistical Methods, 2013). For all subjects, 40 runs were found to meet or exceed the necessary sample size.

$$N = \left(\frac{1.28}{0.1}\right)^2 \sigma^2 \tag{3.2}$$

where:

N is the necessary sample size

1.28 is the corresponding Z score for the chosen confidence interval

0.1 is the chosen margin of error

In each of the 40 runs a randomized workload profile was created and saved. After all simulations were complete, a single representative workload profile was created by averaging VACP workload values at each second of the 40 runs. This averaging process acted similarly to a low-pass signal filter where specific events were washed out, and larger trends emerged. A "smoothed" stochastic CAWP is shown in Figure 3.5. The generation of these stochastic workload profiles led to Hypothesis 3.1: Based on the bias vs. variance tradeoff described by James, Witten, Hastie, & Tibshirani (2013), it is believed that "smoothed" stochastic CAWPs will reduce variance in cross-subject machine learning models and provide superior generalization when compared to deterministic CAWPs.

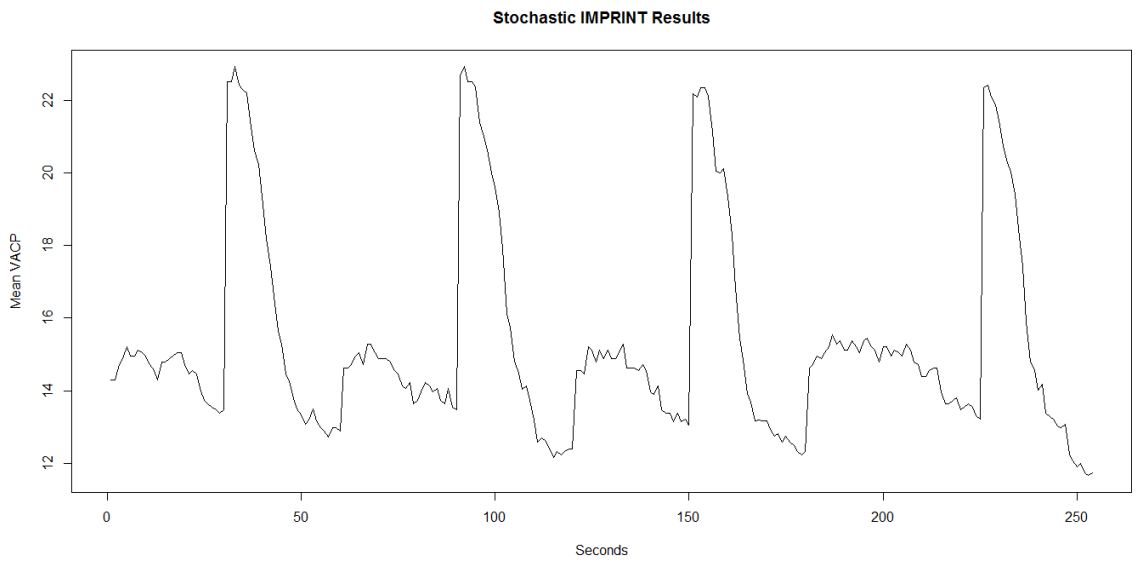

Figure 3.5. A Stochastic Workload Profile

Following the creation of deterministic and stochastic workload profiles for each subject, machine learning models were trained. EEG recordings were synchronized with the CAWP data and used as independent variables (IVs) to estimate workload. The CAWPs served as the dependent variable (DV) for each model.

3.4. Machine Learning Algorithms

Random forest (RDF) (Breiman, 2001) and linear regression (LM) machine learning models were used to estimate VACP workload based on EEG feature vectors. In order to determine cross-subject model viability, both within-subject testing (i.e. models trained using individual training data were tested on data from the same individual) and cross-subject testing (i.e. models trained using training data from all but one subject were tested on data from the left out subject) were accomplished for each model. Machine learning models were also trained on both stochastic and deterministic CAWPs to

measure the viability of the stochastic profiles. Naïve predictors were created for each subject to serve as baselines of comparison for LM and RDF models.

3.4.1. Random Forests

RDFs were used due to their reported resistance to over-fitting, their ease of use (two tunable parameters), and their ability to model non-linear data. The number of trees, *ntree*, used in each model was held constant at 100 to avoid excessive computation time. The number of features randomly sampled as candidates at each split, *mtry*, was set to *f/3* where *f* was the number of features available in the given model, based on results from a previous pilot study that utilized only one subject. RDF models were trained and tested using the `randomForest()` and `predict.randomForest()` functions available in the randomForest 4.6-10 R package.

3.4.2. Linear Regression

Simple LMs were used to provide a less complex machine learning alternative to RDFs. LMs were trained and tested using the `lm()` and `predict.lm()` functions available in the base stats package for R 3.2.1. QR decomposition was used to fit each linear model. The tendency of LM models to "over-fit" training data prompted the use of best subset feature selection for each of the fitted models. During best subset selection, models of all possible subsets of features were compared, and the best was chosen based on goodness of fit, calculated using Mallow's C_p (Equation 3.3) (James et al., 2013).

$$C_p = \frac{1}{n}(RSS + 2d\hat{\sigma}^2) \tag{3.3}$$

where:

n is the number of data points

d is the number of predictors

$\hat{\sigma}^2$ is an estimate of the error associated with each response

3.4.3. Naïve Predictor

Naïve predictors for each subject were created based solely on previously observed VACP truth data from other 15 trials (data from all trials except for the one being predicted). The mean of all previous VACP data was output for each experimental trial. Physiological data was not used for these predictors.

3.5. Physiological Training Data

Machine learning models were trained using either within-subject or cross-subject physiological data. Within-subject machine learning models were fitted for each subject by pairing physiological data from that individual with deterministic or stochastic workload profiles for each experimental trial. Cross-subject models were fitted for each subject using a similar process, but only physiological data from other subjects was paired with workload profile data (e.g. cross-subject models for subject 2 were trained using physiological data from subjects 5,6,7,11, and 14, but not subject 2).

3.6. Cross-Validation

Leave one out cross validation (LOOCV) was used for all machine learning models in order to approximate model generalization. For each model, root mean squared error (RMSE) (Equation 3.4) and the coefficient of determination (R^2) (Equation 3.5) were calculated for multiple "folds" of observations. Finally, RMSE and R^2 were averaged across each of the folds.

$$RMSE = \sqrt{\frac{RSS}{n}} \tag{3.4}$$

where n is the number of data points

$$R^2 = 1 - \frac{RSS}{TSS} \tag{3.5}$$

where TSS is the total sum of squares for the observed data points

For within-subject validation, each of the 16 completed surveillance trials was considered a fold. Individual machine learning models were calibrated for each of the completed trials using the remaining 15 trials as training data. The calibrated machine learning models were then tested on the held out trial. In cross-subject validation, all observations from an individual subject were considered a fold. Machine learning models were trained and tested in a similar fashion to the within-subject process for each of the six subjects. Table 3.7 shows the number of training and testing observations per fold for both of the validation methods.

Table 3.7. Cross-Validation Observations Per Fold

Validation Method	*# of Folds*	*Test Observations Per Fold*	*Training Observations Per Fold*
Within-subject	16	171	2736
Cross-subject	6	2736	13680

3.7. Summary

This chapter described the creation of multiple EEG-based machine learning models. Workload profiling method (stochastic vs. deterministic), model generalization (cross-subject vs. within-subject), and algorithm choice (RDF vs. LM) were varied to explore the effect of each of the factors as well as their interactions. Model performance and observed effects will be reported in the following chapter.

IV. Analysis and Results

In this chapter the effects of varying continuous analytic workload profile (CAWP) method (stochastic vs. deterministic), physiological training data (cross-subject vs. within-subject), and algorithm choice (random forest (RDF) vs. linear regression (LM)) in EEG-based mental workload models are analyzed. Analysis of variance (ANOVA) is used to determine significant sources of variation in cross-validated root mean squared error (RMSE) and coefficient of determination (R^2) of machine learning models. Additional post-hoc analysis is completed using Tukey's Honest Significant Difference (Tukey's HSD) to answer the research questions: Q1.) Is there a significant performance difference between machine learning models fitted using cross-subject, rather than within-subject physiological data? and Q2.) Is there a significant performance difference between machine learning models fitted using stochastic, rather than deterministic CAWPs?

4.1. Correlation and Distribution Analysis

Pearson's correlation coefficients and Quantile-Quantile (Q-Q) plots were used to answer the verification question: Do CAWPs created using deterministic DES correlate with cumulative subjective task load ratings and follow similar distributions? Correlation between time-weighted Visual, Auditory, Cognitive, and Psychomotor (VACP) workload from deterministic CAWPs and NASA Task-Load Index (NASA-TLX) ratings for the six subjects ranged from 0.4273 to 0.7825. The mean correlation across all subjects was 0.6719 with a standard deviation of 0.1404. Correlation values for each subject are shown

in Figure 4.1. The 90% lower confidence bound rested above zero, with the lowest, 0.0004, belonging to Subject 5.

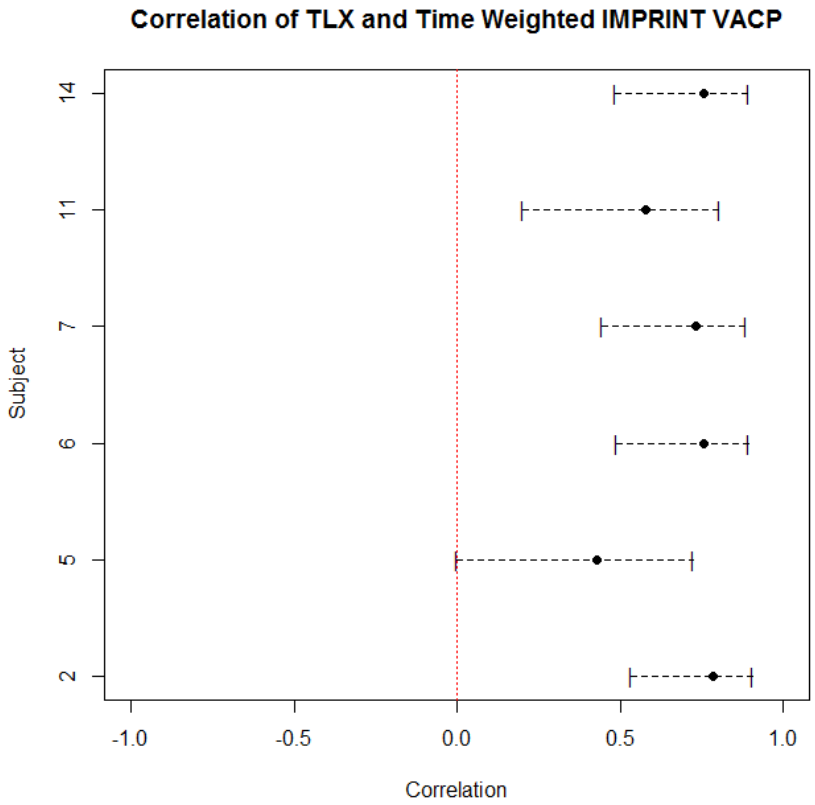

Figure 4.1. Correlation of TLX and Time Weighted IMPRINT VACP

A Q-Q plot was used to visually compare Improved Performance Research Integration Tool (IMPRINT) and NASA-TLX distributions. Prior to creating the plots, IMPRINT and NASA-TLX observations from all subjects/trials were z-scored by subject (to remove scaling differences between subjects). The values were then ordered and plotted against one another. The resulting plot is shown in Figure 4.2. The distributions appear to have slightly different values in the right tails, but overall the plot follows a $y=x$ line.

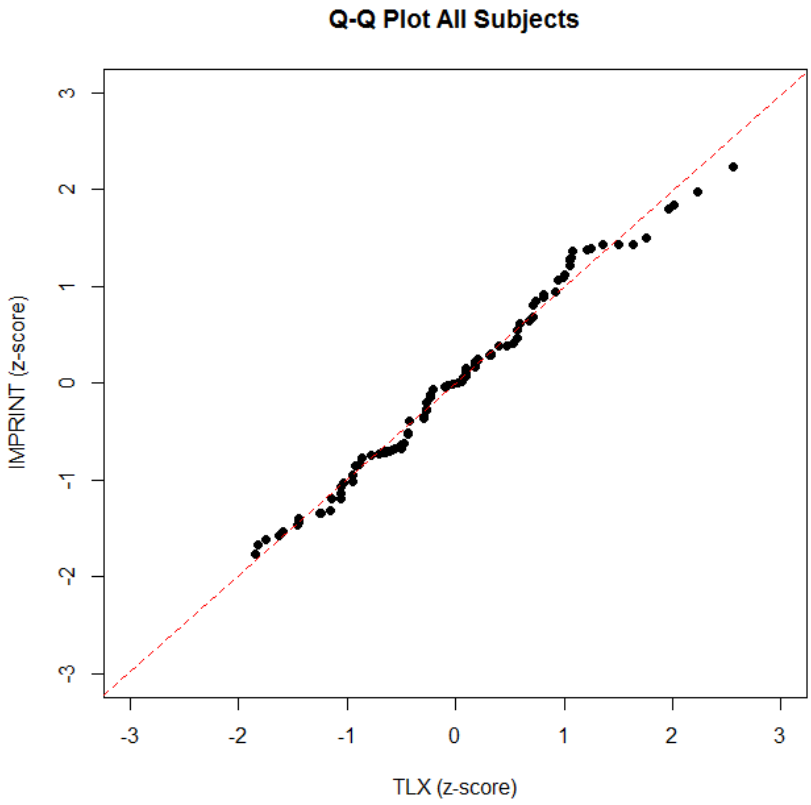

Figure 4.2. Q-Q Plot of IMPRINT and NASA-TLX Observations

The range of the reported confidence intervals paired with the y=x relationship shown in the Q-Q plot suggests that the deterministic workload profiling method described in Chapter III was a success. This verified our decision to utilize the deterministic CAWPs as ground truth data when evaluating machine learning model performance.

4.2. Frequency of Truth Data

Truth data that was used to calibrate and test the performance of machine learning models was heavily skewed. Figure 4.3 illustrates the frequency of the VACP values observed in the truth data. The histogram highlights large imbalances between values,

with 6545 observations assigned a VACP value of 14.3, but only 9 observations assigned a value of 27.8. It follows that machine learning models had the smallest magnitude residuals at values near 14.3 and larger residuals at less frequent values near 27.8. It is believed that a more uniform distribution of observations would have improved machine learning performance.

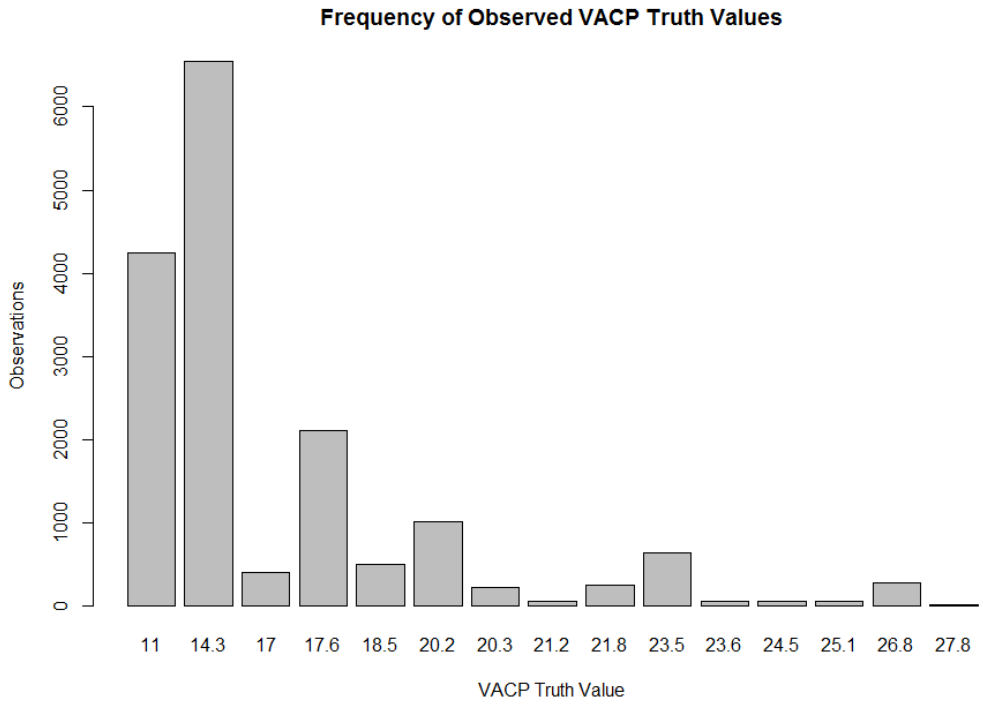

Figure 4.3. Frequency of Observed VACP Truth Values

4.3. Analysis of Variance

Separate two-way ANOVAs were accomplished using R^2 and RMSE as dependent variables to identify statistically significant factors in machine learning model performance. The factors and levels used for each ANOVA are shown in Table 4.1. Both main and interaction effects were analyzed. Appendix D details the full list of observations used for this analysis.

Table 4.1. ANOVA Factors and Levels

Factor	*Level 1*	*Level 2*
Workload Model	Deterministic	Stochastic
Machine Learning Algorithm	LM	RDF
Training Data Source	Within-subject	Cross-subject

Both of the ANOVAs revealed significant main effects of 'Physiological Data' (R^2: $p < 0.001$, RMSE: $p < 0.01$) on model performance. Comparing R^2 values identified a larger number of significant factors. The R^2 based ANOVA also identified a significant main effect of 'Workload Model' ($p < 0.05$) and a significant interaction between 'Workload Model' and 'Training Data Source' ($p < 0.01$). Table 4.2 and Table 4.3 show the detailed results of each ANOVA.

Table 4.2. R^2 based ANOVA

Factor	*Df*	*Sum Sq*	*Mean Sq*	*F value*	*Pr(>F)*	
ML.Algorithm	1	0.00119	0.00119	0.847	0.363	
WL.Model	1	0.00923	0.00923	6.54	0.0144	*
Training.Data	1	0.07305	0.07305	51.779	1.00E-08	***
ML.Algorithm:WL.Model	1	0.00016	0.00016	0.116	0.7351	
ML.Algorithm: Training.Data	1	0.00369	0.00369	2.613	0.1139	
WL.Model: Training.Data	1	0.00893	0.00893	6.332	0.016	*
ML.Algorithm:WL.Model:Training.Data	1	0.00011	0.00011	0.077	0.7828	
Residuals	40	0.05643	0.00141			

Table 4.3. RMSE based ANOVA

Factor	*Df*	*Sum Sq*	*Mean Sq*	*F value*	*Pr(>F)*	
ML.Algorithm	1	0.0055	0.00548	0.16	0.69156	
WL.Model	1	0.0379	0.03786	1.103	0.2999	
Training.Data	1	0.3017	0.30165	8.789	0.00509	**
ML.Algorithm:WL.Model	1	0.0007	0.00074	0.021	0.88428	
ML.Algorithm: Training.Data	1	0.0152	0.01516	0.442	0.51009	
WL.Model: Training.Data	1	0.0402	0.04023	1.172	0.28548	
ML.Algorithm:WL.Model:Training.Data	1	0.0005	0.00054	0.016	0.90096	
Residuals	40	1.3729	0.03432			

4.4. Comparison of Workload Model Performance

The workload models factor had two levels: stochastic and deterministic. The levels represented the two workload profiling methods using to model operator workload. Figure 4.4 shows the effect of the workload model factor on machine learning performance. The large performance difference between workload model levels for subject 2 appears to have had the largest effect on R^2 values. A potential explanation for this is that subject 2's true workload profile was much different than the rest of the subjects under study. Tukey's HSD revealed that deterministic workload profiles had higher R^2 (diff: 0.0058 to 0.0496, 95% CI) across all subjects. 0 provides details of deterministic and stochastic workload model performance across the 15 possible VACP truth values. Perfect predictions would have resulted in values lying on the dashed line. With the exception of the VACP values ranging between 17 and 20.2, the median of deterministic models rested closer than the stochastic models to the dashed line at residual value zero. A potential explanation for increased performance of stochastic models at these VACP values is their proximity to the mean VACP value of 15.3.

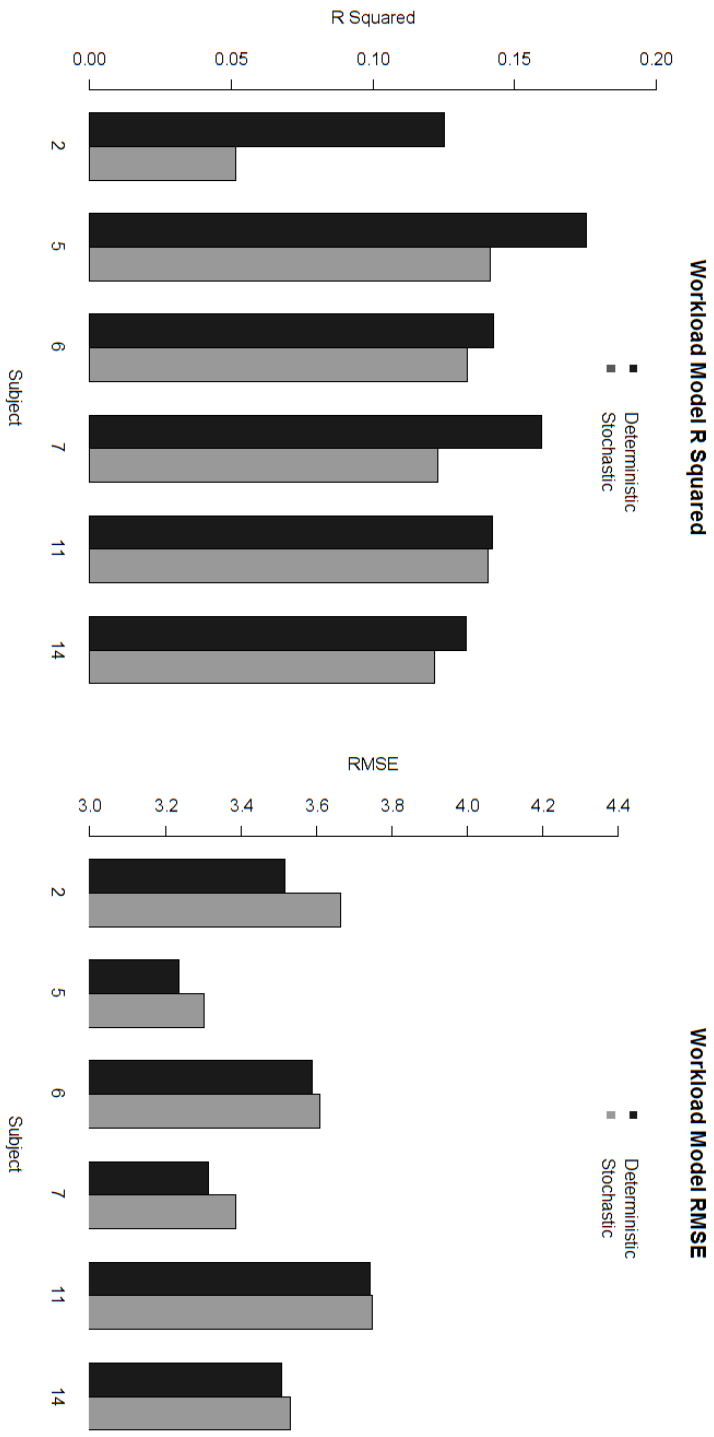

Figure 4.4. Effect of Workload Model Factor on Performance

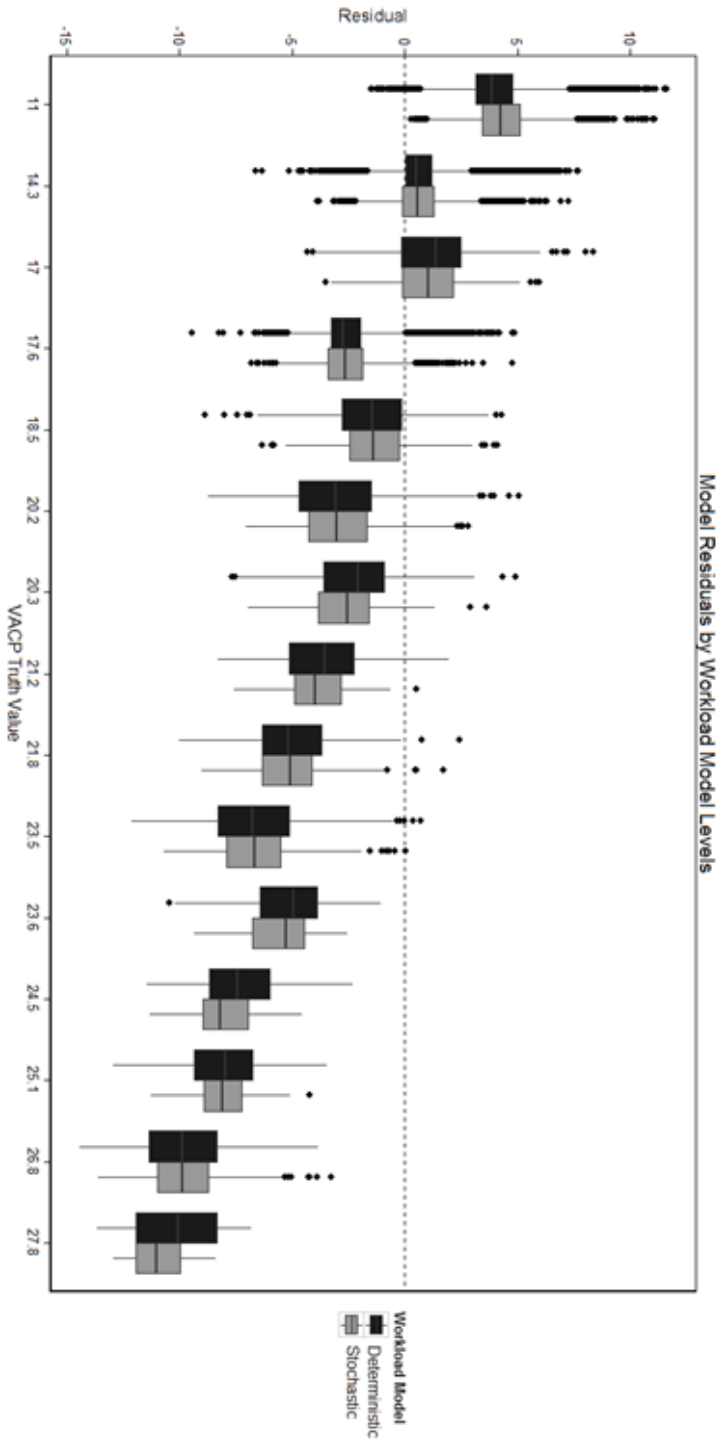

Figure 4.5. Model Residuals by Workload Model Levels

4.5. Comparison of Machine Learning Algorithm Performance

The machine learning algorithm factor had three levels: naïve, LM, and RDF. The levels represented the three algorithms used to estimate VACP workload. Figure 4.6 shows the effect of the machine learning algorithm factor on machine learning performance. The negative R^2 values for the naïve predictor imply that its predictions were worse than simply choosing the mean across all truth values. Tukey's HSD revealed no significant difference when comparing the performance of LM and RDF algorithms. However, both algorithms performed significantly better than the naïve predictor. The 95% CI on the differences in performance between RDF and naïve models were 0.1223 to 0.1648 for R^2 and -0.4089 to -0.1561 for RMSE. The 95% CI on the differences in performance between LM and naïve models were 0.1123 to 0.1548 for R^2 and -0.3876 to -0.1347 for RMSE. Figure 4.7 details the performance of each of the three algorithms across the 15 possible VACP truth values. LM models appeared to provide a performance gain when making predictions near the mean VACP value of 15.3156. The benefits of the RDF algorithm can be observed at values greater than 18 where fewer observations were available for model calibration.

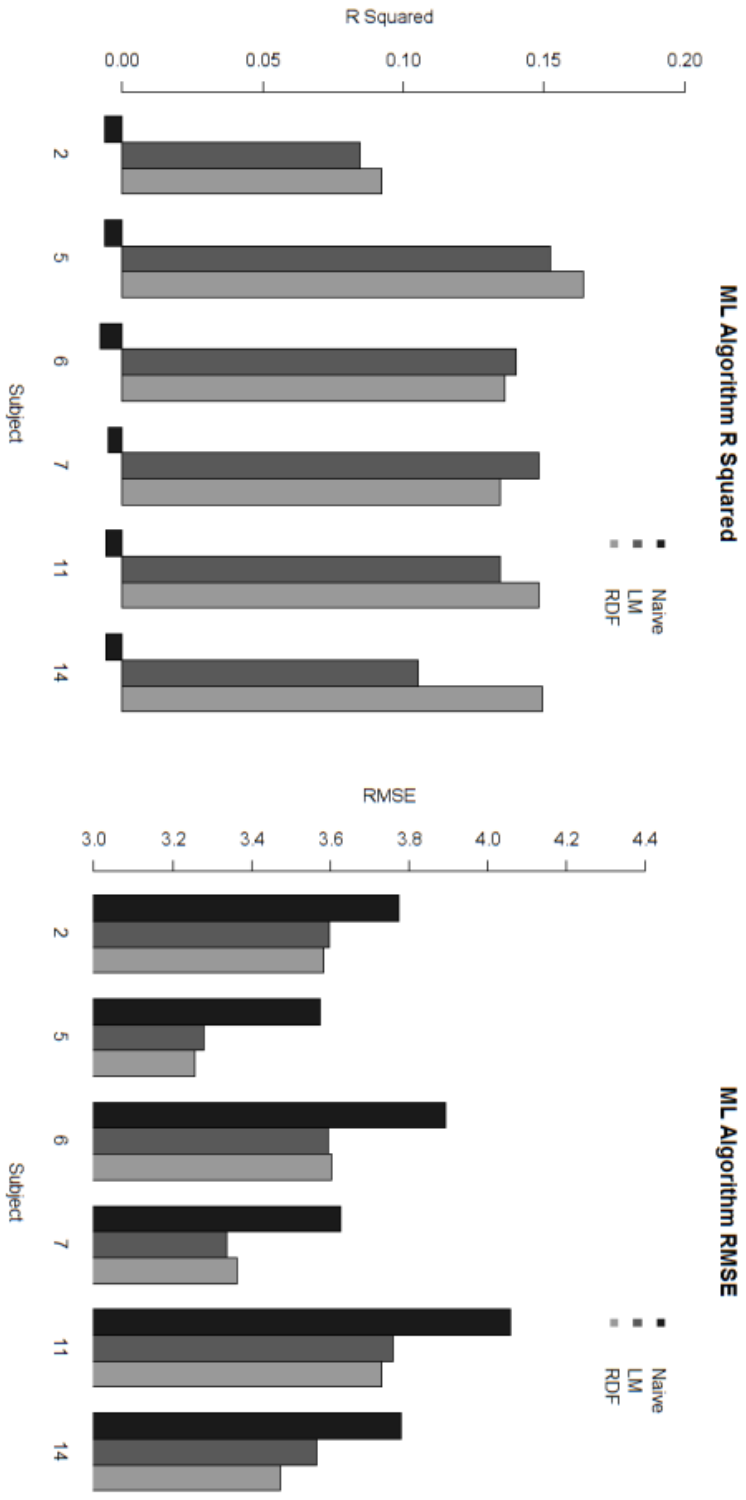

Figure 4.6. Effect of Machine Learning Algorithm Factor on Performance

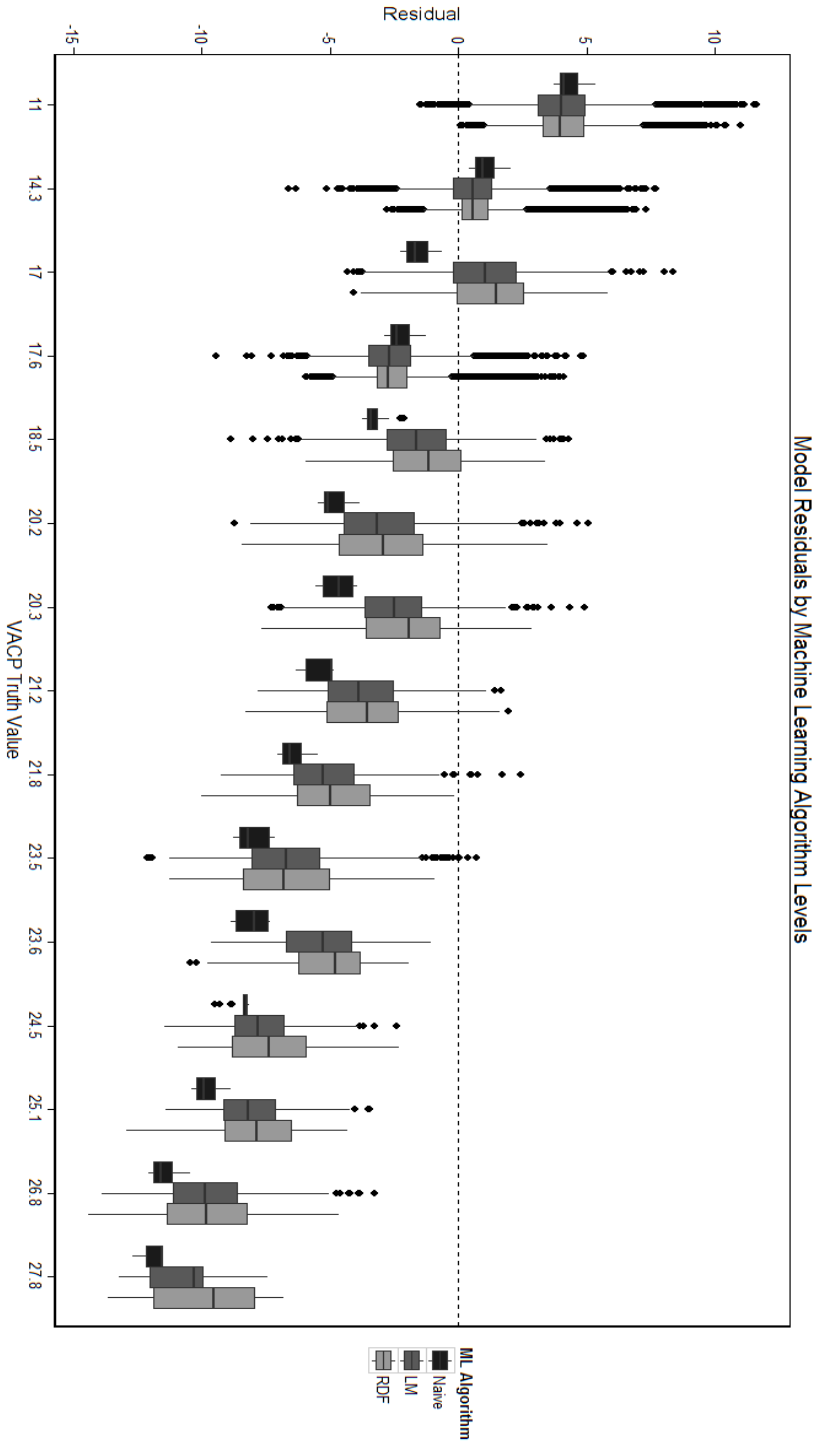

Figure 4.7. Model Residuals by Machine Learning Algorithm Levels

4.6. Comparison of Physiological Training Data Source

The training data source factor had two levels: within-subject and cross-subject. The levels represented the source of the EEG data that was paired with VACP truth data to calibrate, or train, machine learning models. Figure 4.8 shows the effect of the physiological training data factor on machine learning performance. A clear boundary between within-subject and cross-subject models can be observed in the R^2 plot. While not as pronounced as the R^2 plot, differences between within-subject and cross-subject models can also be seen in the RMSE plot. Tukey's HSD showed that within-subject models had higher R^2 (diff: 0.0561 to 0.0999, 95% CI) and lower RMSE (diff: -0.2666 to -0.0505, 95% CI) across all subjects. Figure 4.9 provides details of within-subject and cross-subject models across the 15 possible VACP truth values. Cross-subject model performance was highest at VACP values of 24.3, 17, and 17.6. Again, proximity to the mean VACP value of 15.3 was expected to have been a factor.

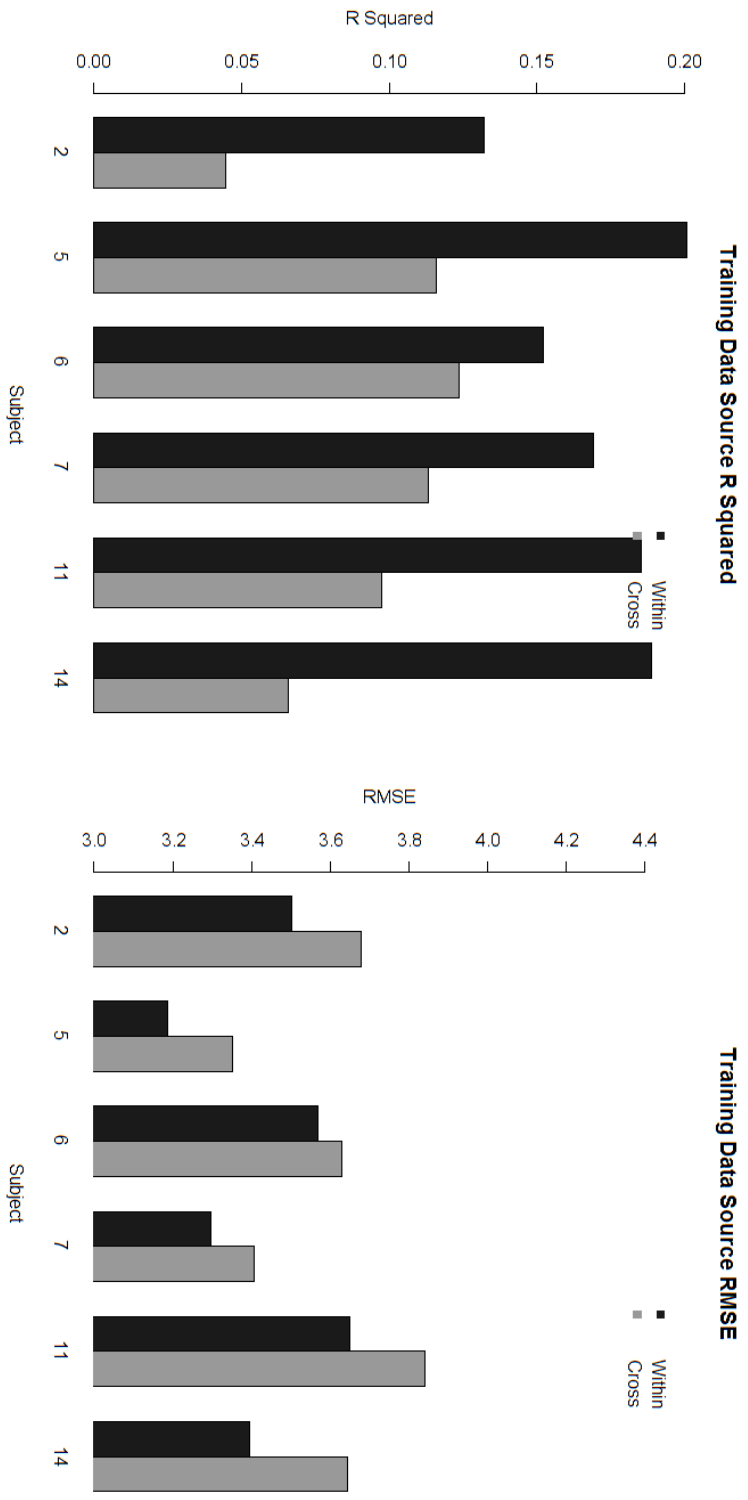

Figure 4.8. Effect of Physiological Training Data Factor on Performance

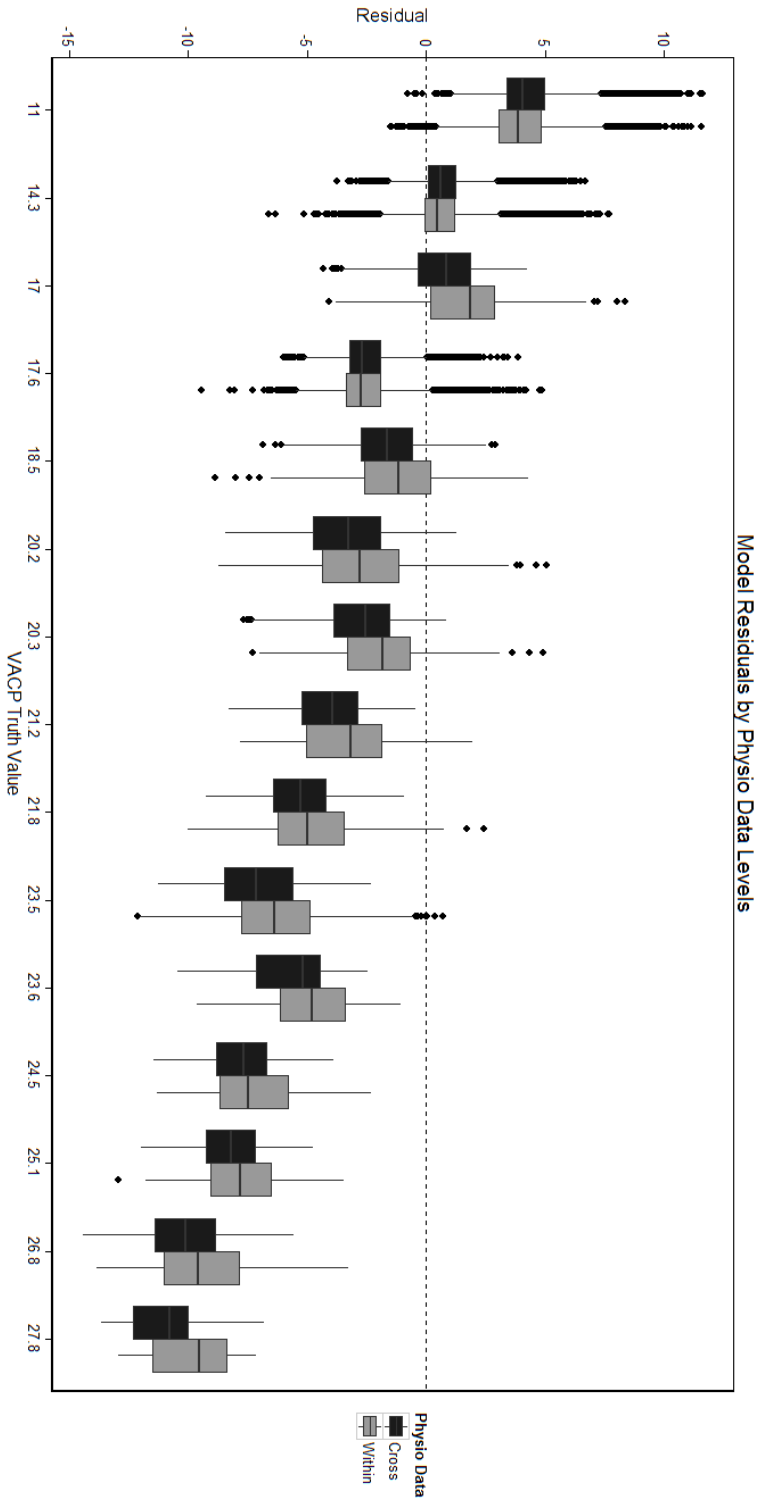

Figure 4.9. Model Residuals by Physiological Training Data Levels

4.7. Summary

In this chapter, experimental results and analysis from the methods described in Chapter III were presented. Results failed to support the hypotheses presented in Chapter III. The investigative questions and hypotheses along with the answers supported in this chapter are summarized below:

Q1. Is there a significant performance difference between machine learning models fitted using cross-subject, rather than within-subject physiological data? Post-hoc testing on an R^2 and RMSE-based ANOVAs revealed statistically significant decreases in performance when using cross-subject physiological training data, rather than within-subject data. On average, cross-subject data decreased machine learning R^2 values from 0.1715 to 0.0935.

Q2. Is there a significant performance difference between machine learning models fitted using stochastic, rather than deterministic CAWPs? Post-hoc testing on an R^2 based ANOVA revealed a statistically significant decrease in performance when using stochastic, rather than deterministic CAWPs, but no significant difference was found between stochastic and deterministic workload models when testing only cross-subject data. On average, stochastic CAWPs decreased machine learning R^2 values from 0.1463 to 0.1186.

Hypothesis 3.1. Based on the bias vs. variance tradeoff described by James et al. (2013), it is believed that "smoothed" stochastic CAWPs will reduce variance in cross-subject

machine learning models and provide superior generalization when compared to deterministic CAWPs. As mentioned in the discussion regarding Q2, stochastic CAWPs did not improve performance when compared to deterministic CAWPs. As expected, the use of stochastic CAWPs washed away much of the variance in the machine learning models, but the amount of bias introduced was higher than expected. The result was models that tended to make predictions that were too close to the mean observed VACP truth value.

Hypothesis 3.2. Based on the overwhelming success of non-parametric, non-linear machine learning models (i.e. artificial neural networks and support vector machines) in related EEG-based classification, it is expected that RDF regression will outperform LM models when used to infer OFS. Post-hoc testing on R^2 and RMSE based ANOVAs revealed no significant difference when comparing the performance of LM and RDF algorithms. The expected performance gains from RDF models appear to have been overshadowed by the performance of LM models near the mean observed VACP truth value.

V. Conclusions and Recommendations

Research focused on creating scalable machine learning models that were capable of estimating operator functional state (OFS) in a remotely piloted aircraft (RPA) simulation, using two generalization methods. The first method utilized physiological data from multiple previously observed operators to estimate OFS for unseen operators, reducing the need to collect truth data or train individualized workload models for new operators. The second used stochastic, distribution-based representations of operator behavior rather than exact second-by-second data to train machine learning models reducing the number of observations needed for model calibration.

A full factorial design was used to create machine learning models that utilized cross-subject or within-subject physiological data as well as stochastic or deterministic CAWPs to estimate OFS. The performance of each model was then calculated using two measures: coefficient of determination (R^2) and root mean squared error (RMSE). Analysis of variance (ANOVA) was accomplished using both the R^2 and RMSE values to determine the effects of physiological training data (cross-subject vs. within subject), workload model (stochastic vs. deterministic profiling), and machine learning algorithm (naïve vs. linear regression (LM) vs. random forest (RDF)) on machine learning model performance.

5.1. Research Findings

Post-hoc testing on R^2 and RMSE based ANOVAs revealed significant decreases in performance when using cross-subject, rather than within-subject physiological training data. The post-hoc testing on R^2 ANOVAs also revealed performance decreases

when using stochastic, rather than deterministic CAWPs. Cross-subject models performed worse than within-subject models when comparing R^2 performance, dropping from 0.1715 to 0.0935. Stochastic models decreased performance less, reducing R^2 performance from only 0.1463 to 0.1186 when compared to deterministic models.

5.2. Future Research

While completing research related to this thesis, additional research activities were identified. The most obvious of these activities is the extension of this work to an operational environment. Operational environments pose several challenges to the methods used in this research effort: Will physiological reading devices like the electroencephalogram (EEG) provide adequate signal to noise (SNR) in order to estimate OFS outside of a laboratory environment? Can we effectively model the complexities of an operational environment using the described CAWP creation process? Unfortunately, we cannot know the extent of these challenges without moving away from the control of laboratory settings.

Another potential line of research is the use of machine learning clustering methods to analyze which periods of operator activity are similar based solely on physiological data. The analysis could then be used to develop more meaningful operator states when performing discrete event simulation (DES). This could lead to more accurate truth data and improve performance of supervised machine learning models when estimating OFS.

Lastly, automated encoding of user activity data would be extremely useful when generating CAWPs. The amount of time needed to manually encode user activity from

video footage was a major limiting factor in this thesis. In addition to reducing data collection time for future researchers, automated encoding would also reduce potential encoding errors (e.g. incorrect time recording or undocumented user activity).

5.3. Significant Contributions

The proposed methods provided solutions to the limitations that stem from lengthy training data collection and labeling techniques associated with generating CAWPs for multiple operators/trials. It was shown that group workload models could be used to infer OFS on new subjects, reducing the need to collect truth data or train individualized workload models for new subjects. Performance decreases related to cross-subject modeling were steep, reducing R^2 values by nearly three times the amount reduced by using stochastic models. Stochastic techniques that were used to generate representative workload profiles using a limited number of training observations were shown to be a more viable solution.

The findings presented in this research required successful completion of tasks that spanned several disciplines. Digital signal processing concepts, i.e. the short-time Fourier transform (STFT), were required to extract time-frequency data for EEG signals. Cognitive task analyses (CTA) and the creation of DES networks in the Improved Research Integration Tool (IMPRINT) incorporated ideas from Ergonomics and Systems Engineering. Lastly, data wrangling and cross-validation techniques were used to fit machine learning models to the observed data. By combining all of these techniques, I provided a framework to map relationships between physiological recordings and OFS in previously accomplished human performance studies.

Appendix A. CTA Knowledge Audit

1. Basic Feature Searching (Noticing/Job Smarts)

   a. *Cues and Strategies*

      1) Break the area of interest up into zones that can be easily identified

      2) Prioritize high traffic areas (targets rarely remain stationary)

      3) Scan at an appropriate zoom level

      4) Scan thoroughly, but quickly

      5) Prioritize by target clothing, posture, and potential weapons

   b. *Reasons for Difficulty*

      1) Transitioning between zones that do not contain identifiable landmarks becomes difficult due to a continuously changing camera perspective

      2) Targets are only visible for a short amount of time. Prioritizing high traffic areas first, gives better odds of locating a target early

      3) Scanning with an improper zoom level leads to either a slow scan or missed details

      4) Slow scanning increases the chances of a target entering an area that was previously scanned

      5) Many potential targets wear similar clothing, and carry items easily mistaken as weapons

2. Target Verification (Noticing/Big Picture)

   a. *Cues and Strategies:* Zoom only as far as necessary to verify potential targets

   b. *Reasons for Difficulty:* Excessive zooming on incorrect targets reduces scene awareness and increases the chances of losing position along scan route

3. Target Tracking (Past and Future/Noticing)

    a. *Cues and Strategies:* Pay close attention to target movement in crowded areas and estimate potential movements

    b. *Reasons for Difficulty:* Targets unexpectedly change directions or temporarily move out of sight

4. Computing Radio Responses (Job Smarts)

    a. *Cues and Strategies*

        1) Closely monitor radio traffic

        2) Memorize aircraft velocity and altitude

    b. *Reasons for Difficulty:* Diverting attention from target location/tracking in order to view message traffic or aircraft information increases the risk of target loss

Appendix B. Simulation Interview

1. Basic Feature Searching

    a. *Actions*

        1) Determine the target surveillance area

        2) Move the camera around the target area

        3) Set zoom for appropriate feature searching

        4) Scan scene for potential targets

    b. *Assessment:*

        1) Need to understand landscape and traffic patterns.

        2) Potential targets near scene boundaries may exit prior to proper scanning

    c. *Critical Cues*

        1) High traffic areas

        2) Easily identified landmarks

        3) Mobile individuals

        4) Individuals carrying large objects

    d. *Potential Errors*

        1) Following an inefficient scan path that does not allow for easy transitions between zones when scanning

        2) Moving slowly and missing potential targets that leave the scene

        3) Moving quickly and not recognizing potential targets

        4) Incorrectly identifying potential targets

2. Target Verification

    a. *Actions*: Analyze potential target

b. *Assessment:* Pay close attention to objects that the target is carrying

c. *Critical Cues:* Individuals carrying rifles are high value targets

d. *Potential Errors*

1) Many potential targets may carry large tools instead of weapons

2) Spending too long focusing on an incorrect target

3) Losing scene awareness due to improper use of zoom

3. Target Tracking

a. *Actions*: Anticipate target movement

b. *Assessment*

1) Look for potential blind spots

2) Be aware of uncontrollable camera rotation

c. *Critical Cues:* Camera moving such that target visibility will be lost

d. *Potential Errors:* Target loss due to unforeseen blind spot

4. Computing Radio Responses

a. *Actions*

1) Compute response

2) Respond to radio information request

b. *Assessment:* Determine if radio information is an information request

c. *Critical Cues:* Information is not requested if a target is not present

d. *Potential Errors*

1) Loss of target due to use of text messaging or information lookup

Appendix C. Stochastic Variable Distributions

		Subject 2		Subject 5	
		Exp	p	Exp	p
Primary	**Search**	0.5 + LOGN(8.7, 9.34)	0.005	0.5 + LOGN(7.81, 8.27)	0.048
	Verify	0.5 + 24 * BETA(0.596, 5.97)	0.005	0.5 + 24 * BETA(0.623, 6.42)	0.005
	Found Target	P(0.2129)		P(0.2010)	
Secondary	**Process Question**	1.5 + LOGN(5.43, 2.05)	0.005	1.5 + LOGN(5.32, 2.06)	0.005
	Needed Console	P(0.225)		P(0.2219)	
	Read Message	0.5 + 9 * BETA(0.521, 1.78)	0.0427	0.5 + 9 * BETA(0.515, 1.75)	0.041
	Answered Comm	P(0.9469)		P(0.9469)	
	Compute Answer	0.5 + 13 * BETA(0.624, 3.03)	0.75	0.5 + 13 * BETA(0.541, 2.82)	0.38

		Subject 6		Subject 7	
		Exp	p	Exp	p
Primary	**Search**	0.5 + LOGN(8.49, 9.11)	0.0121	0.5 + LOGN(8.21, 8.72)	0.0466
	Verify	0.5 + 24 * BETA(0.623, 6.42)	0.005	0.5 + 24 * BETA(0.623, 6.42)	0.005
	Found Target	P(0.2201)		P(0.2109)	
Secondary	**Process Question**	1.5 + LOGN(5.19, 1.93)	0.005	1.5 + LOGN(5.49, 2.01)	0.005
	Needed Console	P(0.1938)		P(0.2156)	
	Read Message	0.5 + 9 * BETA(0.521, 1.63)	0.0368	0.5 + 9 * BETA(0.479, 1.7)	0.05
	Answered Comm	P(0.975)		P(0.9469)	
	Compute Answer	1.5 + 23 * BETA(0.521, 3.18)	0.742	0.5 + 13 * BETA(0.65, 3.09)	0.75

		Subject 11		Subject 14	
		Exp	p	Exp	p
Primary	**Search**	0.5 + LOGN(7.89, 8.12)	0.0183	0.5 + LOGN(7.81, 8.24)	0.2
	Verify	0.5 + 24 * BETA(0.623, 6.42)	0.005	0.5 + 24 * BETA(0.623, 6.42)	0.005
	Found Target	P(0.2215)		P(0.2102)	
Secondary	**Process Question**	3.5 + ERLA(0.599, 5)	0.005	1.5 + LOGN(5.27, 2.04)	0.005
	Needed Console	P(0.1188)		P(0.1969)	
	Read Message	0.5 + 7 * BETA(0.515, 2.4)	0.005	0.5 + 9 * BETA(0.522, 1.73)	0.0117
	Answered Comm	P(0.9719)		P(0.9469)	
	Compute Answer	0.5 + 11 * BETA(0.657, 4.18)	0.23	0.5 + 13 * BETA(0.516, 2.76)	0.381

Appendix D. Detailed Machine Learning Performance

Subject	*ML.Algorithm*	*WL.Model*	*Training.Data Source*	*R2*	*RMSE*
2	LM	Sto	Cross	0.0290	3.7084
2	LM	Det	Cross	0.0513	3.6657
2	RDF	Sto	Cross	0.0350	3.6970
2	RDF	Det	Cross	0.0633	3.6424
2	Naïve	Sto	Cross	-0.0061	3.7749
2	Naïve	Det	Cross	-0.0061	3.7749
2	LM	Det	Within	0.1818	3.4043
2	LM	Sto	Within	0.0757	3.6182
2	RDF	Det	Within	0.2045	3.3567
2	RDF	Sto	Within	0.0668	3.6356
2	Naïve	Det	Within	-0.0061	3.7749
2	Naïve	Sto	Within	-0.0061	3.7749
5	LM	Sto	Cross	0.1301	3.3242
5	LM	Det	Cross	0.1255	3.3331
5	RDF	Sto	Cross	0.0853	3.4087
5	RDF	Det	Cross	0.1228	3.3382
5	Naïve	Sto	Cross	-0.0063	3.5754
5	Naïve	Det	Cross	-0.0063	3.5754
5	LM	Det	Within	0.2012	3.1856
5	LM	Sto	Within	0.1529	3.2804
5	RDF	Det	Within	0.2518	3.0830
5	RDF	Sto	Within	0.1969	3.1941
5	Naïve	Det	Within	-0.0063	3.5754
5	Naïve	Sto	Within	-0.0063	3.5754
6	LM	Sto	Cross	0.1316	3.6139
6	LM	Det	Cross	0.1349	3.6070
6	RDF	Sto	Cross	0.1248	3.6280
6	RDF	Det	Cross	0.1030	3.6728
6	Naïve	Sto	Cross	-0.0077	3.8930
6	Naïve	Det	Cross	-0.0077	3.8930
6	LM	Det	Within	0.1628	3.5483
6	LM	Sto	Within	0.1311	3.6150
6	RDF	Det	Within	0.1700	3.5332
6	RDF	Sto	Within	0.1457	3.5844
6	Naïve	Det	Within	-0.0077	3.8930
6	Naïve	Sto	Within	-0.0077	3.8930
7	LM	Sto	Cross	0.1119	3.4091
7	LM	Det	Cross	0.1555	3.3244
7	RDF	Sto	Cross	0.0871	3.4564
7	RDF	Det	Cross	0.0989	3.4340
7	Naïve	Sto	Cross	-0.0048	3.6262
7	Naïve	Det	Cross	-0.0048	3.6262
7	LM	Det	Within	0.1856	3.2645
7	LM	Sto	Within	0.1398	3.3551
7	RDF	Det	Within	0.1988	3.2380
7	RDF	Sto	Within	0.1529	3.3294
7	Naïve	Det	Within	-0.0048	3.6262
7	Naïve	Sto	Within	-0.0048	3.6262
11	LM	Sto	Cross	0.1286	3.7759

11	LM	Det	Cross	0.0445	3.9538
11	RDF	Sto	Cross	0.1364	3.7589
11	RDF	Det	Cross	0.0806	3.8785
11	Naïve	Sto	Cross	-0.0059	4.0567
11	Naïve	Det	Cross	-0.0059	4.0567
11	LM	Det	Within	0.2157	3.5821
11	LM	Sto	Within	0.1497	3.7297
11	RDF	Det	Within	0.2286	3.5526
11	RDF	Sto	Within	0.1480	3.7335
11	Naïve	Det	Within	-0.0059	4.0567
11	Naïve	Sto	Within	-0.0059	4.0567
14	LM	Sto	Cross	0.0529	3.6693
14	LM	Det	Cross	0.0710	3.6341
14	RDF	Sto	Cross	0.0662	3.6436
14	RDF	Det	Cross	0.0730	3.6303
14	Naïve	Sto	Cross	-0.0056	3.7811
14	Naïve	Det	Cross	-0.0056	3.7811
14	LM	Det	Within	0.1443	3.4879
14	LM	Sto	Within	0.1523	3.4716
14	RDF	Det	Within	0.2429	3.2808
14	RDF	Sto	Within	0.2159	3.3387
14	Naïve	Det	Within	-0.0056	3.7811
14	Naïve	Sto	Within	-0.0056	3.7811

Bibliography

Aldrich, T. B., Szabo, S. M., & Bierbaum, C. R. (1989). The Development and Application of Models to Predict Operator Workload During System Design. In *Applications of Human Performance Models to System Design* (pp. 65–80). Springer.

Archer, S., & Adkins, R. (1999). IMPRINT User's Guide Prepared for US Army Research Laboratory. Human Research and Engineering Directorate.

Berger, H. (1929). Über das Elektrenkephalogramm des Menschen. *European Archives of Psychiatry and Clinical Neuroscience*, *87*(1), 527–570.

Bierbaum, C. R., Szabo, S. M., & Aldrich, T. B. (1990). *Task Analysis and Workload Prediction Model of the MH-60K Mission and a Comparison with UH-60A Workload Predictions*. Fort Rucker, AL.

Boles, D. B., & Adair, L. P. (2001). The Multiple Resources Questionnaire (MRQ). In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 45, pp. 1790–1794).

Breiman, L. (2001). Random Forests. *Machine Learning*, *45*(1), 5–32.

Byrne, E. A., & Parasuraman, R. (1996). Psychophysiology and Adaptive Automation. *Biological Psychology*, *42*(3), 249–268.

Cassenti, D. N., & Kelley, T. D. (2006). Towards the Shape of Mental Workload. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Vol. 50, pp. 1147–1151).

Dale, A. M., & Sereno, M. I. (1993). Improved Localization of Cortical Activity by Combining EEG and MEG with MRI Cortical Surface Reconstruction: A Linear Approach. *Journal of Cognitive Neuroscience*, *5*(2), 162–176.

Funke, G., Knott, B., Mancuso, V. F., Strang, A., Estepp, J., Menke, L., … Miller, B. (2013). Evaluation of Subjective and EEG-Based Measures of Mental Workload. *Communications in Computer and Information Science*, *373*, 412–416.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of Empirical and Theoretical Research. *Advances in Psychology*, *52*(C), 139–183.

Hogervorst, M. A., Brouwer, A.-M., & Erp, J. (2014). Combining and Comparing EEG , Peripheral Physiology and Eye-Related Measures for the Assessment of Mental Workload. *Frontiers in Neuroscience*, *8*(October), 1–14.

James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical Learning*. New York: Springer.

Jasper, H. H. (1958). Report of the Committee on Methods of Clinical Examination in Electroencephalography: 1957. *Electroencephalography and Clinical Neurophysiology*, *10*(2), 370–375.

Jung, T. P., Makeig, S., Humphries, C., Lee, T. W., McKeown, M. J., Iragui, V., & Sejnowski, T. J. (2000). Removing Electroencephalographic Artifacts by Blind Source Separation. *Psychophysiology*, *37*(2), 163–178.

Kirchner, W. K. (1958). Age Differences in Short-Term Retention of Rapidly Changing Information. *Journal of Experimental Psychology*, *55*(4), 352.

Klem, G. H., Lüders, H. O., Jasper, H. H., & Elger, C. (1999). The Ten-Twenty Electrode System of the International Federation. In G. Deuschl & A. Eisen (Eds.), *Recommendations for the Practice of Clinical Neurophysiology: Guidelines of the International Federation of Clinical Physiology (EEG Suppl. 52)* (Vol. 52, p. 3). Elsevier Science B.V.

Militello, L. G., & Hutton, R. J. B. (1998). Applied Cognitive Task Analysis (ACTA): A Practitioner's Toolkit for Understanding Cognitive Task Demands. *Ergonomics*, *41*(11), 1618–1641.

NIST/SEMATECH e-Handbook of Statistical Methods. (2013). Retrieved January 1, 2015, from http://www.itl.nist.gov/div898/handbook/

North, R. A., & Riley, V. A. (1989). W/INDEX: A Predictive Model of Operator Workload. In *Applications of Human Derformance Models to System Design* (pp. 81–89). Springer.

Ochoa, J. B. (2002). *EEG Signal Classification for Brain Computer Interface Applications*.

Parasuraman, R., Bahri, T., Deaton, J. E., Morrison, J. G., & Barnes, M. (1992). *Theory and Design of Adaptive Automation in Aviation Systems*. *Report No. NAWCADWAR-92033-60*. Warminster, PA: Naval Air Warfare Center, Aircraft Division.

Reid, G. B., & Nygren, T. E. (1988). The Subjective Workload Assessment Technique: A Scaling Procedure for Measuring Mental Workload. *Advances in Psychology*, *52*(C), 185–218.

Rouse, W. B. (1988). Adaptive Aiding for Human/Computer Control. *Human Factors*, *30*(4), 431–443.

Rusnock, C. F., Borghetti, B. J., & McQuaid, I. (2015). Objective-Analytical Measures of Workload – the Third Pillar of Workload Triangulation? In D. D. Schmorrow & C. M. Fidopiastis (Eds.), *Foundations of Augmented Cognition* (Vol. 9183, pp. 124–135). Springer International Publishing.

Rusnock, C. F., & Geiger, C. D. (2014). Simulation-Based Assessment of Performance-Workload Tradeoffs for System Design Evaluation. In Y. Guan & H. Liao (Eds.), *Proceedings of the 2014 Industrial and Systems Engineering Research Conference*.

Smith, A., Borghetti, B. J., & Rusnock, C. F. (2015). Improving Model Cross-Applicability for Operator Workload Estimation. In *Proceedings of the Human Factors and Ergonomics Society 59th Annual Meeting*.

Teplan, M. (2002). Fundamentals of EEG Measurement. *Measurement Science Review*, *2*(2), 1–11.

Tsange, P. S., & Velazquez, V. L. (1996). Diagnosticity and Multidimensional Subjective Workload Ratings. *Ergonomics*, *39*(3), 358–381.

Vidulich, M. A., Ward, G. F., & Schueren, J. (1991). Using the Subjective Workload Dominance (SWORD) Technique for Projective Workload Assessment. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *33*(6), 677–691.

Welford, A. T. (1967). Single-Channel Operation in the Brain. *Acta Psychologica*, *27*, 5–22.

Wickens, C. D. (2002). Multiple Resources and Performance Prediction. *Theoretical Issues in Ergonomics Science*, *3*(2), 159–177.

Wilson, G. F., & Russell, C. a. (2007). Performance Enhancement in an Uninhabited Air Vehicle Task Using Psychophysiologically Determined Adaptive Aiding. *Human Factors*, *49*, 1005–1018.

Young, M. S., & Stanton, N. A. (2002). Attention and Automation: New Perspectives on Mental Underload and Performance. *Theoretical Issues in Ergonomics Science*, *3*(2), 178–194.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From – To)*
24-03-2015	Master's Thesis	September 2014 – March 2016

4. TITLE AND SUBTITLE
Cross-Subject Continuous Analytic Workload Profiling Using Stochastic Discrete Event Simulation

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

5. AUTHOR(S)
Giametta, Joseph J., Capt, USAF

5d. PROJECT NUMBER
JON 15G129

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-16-M-018

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
AFRL / 711th Human Performance Wing
Attn: Scott Galster
2510 Fifth Street, Bldg. 840
Wright-Patterson Air Force Base, Ohio 45433
(937)-798-3632 scott.galster@us.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)
AFRL/RHCPA

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

Operator functional state (OFS) in remotely piloted aircraft (RPA) simulations is modeled using electroencephalograph (EEG) physiological data and continuous analytic workload profiles (CAWPs). A framework is proposed that provides solutions to the limitations that stem from lengthy training data collection and labeling techniques associated with generating CAWPs for multiple operators/trials. The framework focuses on the creation of scalable machine learning models using two generalization methods: 1) the stochastic generation of CAWPs and 2) the use of cross-subject physiological training data to calibrate machine learning models. Cross-subject workload models are used to infer OFS on new subjects, reducing the need to collect truth data or train individualized workload models for unseen operators. Additionally, stochastic techniques are used to generate representative workload profiles using a limited number of training observations. Both methods are found to reduce data collection requirements at the cost of machine learning prediction quality. The costs in quality are considered acceptable due to drastic reductions in machine learning model calibration time for future operators.

15. SUBJECT TERMS
Human Performance, Machine Learning, Continuous Analytic Workload Profiling, Mental Workload

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 Borghetti, AFIT/ENG
U	U	U	UU	75	19b. TELEPHONE NUMBER *(Include area code)* (937) 785-6565, ext 4612 brett.borghetti@afit.edu

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