3-26-2015

Robust Models for Operator Workload Estimation

Andrew M. Smith

ROBUST MODELS FOR OPERATOR WORKLOAD ESTIMATION

THESIS

Andrew M. Smith, Second Lieutenant, USAF

AFIT-ENG-MS-15-M-064

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENG-MS-15-M-064

ROBUST MODELS FOR OPERATOR WORKLOAD ESTIMATION

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 Engineering

Andrew M Smith, BS

Second Lieutenant, USAF

March 2015

AFIT-ENG-MS-15-M-064

ROBUST MODELS FOR OPERATOR WORKLOAD ESTIMATION

Andrew M. Smith, BS

Second Lieutenant, USAF

Committee Membership:

Brett J. Borghetti, PhD
Chair

Brian G. Woolley, Maj, USAF, PhD
Member

Gilbert L. Peterson, PhD
Member

AFIT-ENG-MS-15-M-064

Abstract

As modern technology continues to advance, how can we prevent the human from becoming the weakest component of the human-machine system? When operators are overwhelmed, judicious employment of automation can be beneficial. Ideally, a system which can accurately estimate current operator workload can make better choices when to employ automation. Supervised machine learning models can be trained to estimate workload in real time from operator physiological data. Unfortunately, estimating operator workload using trained models is limited: using a model trained in one context can yield poor estimation of workload in another. This research examines the utility of three algorithms (linear regression, regression trees, and Artificial Neural Networks) in terms of cross-application workload prediction. The study is conducted for a remotely piloted aircraft simulation under several context-switch scenarios – across two tasks, four task conditions, and seven human operators.

Regression tree models were able to cross-predict both task conditions of one task type within a reasonable level of error, and could accurately predict workload for one operator when trained on data from the other six. Six physiological input subsets were identified based on method of measurement, and were shown to produce superior cross-application models compared to models utilizing all input features in certain instances. Models utilizing only EEG features show the most potential for decreasing cross-application error for certain contexts. These findings will contribute to the future development of robust workload estimators for use in on-line adaptive aiding systems.

Acknowledgments

I would like to express my sincere appreciation to my faculty advisor, Dr. Brett Borghetti, for his guidance and support throughout the course of this thesis effort.

Andrew M. Smith

Table of Contents

Page

List of Figures

Page

List of Tables

Page

ROBUST MODELS FOR OPERATOR WORKLOAD ESTIMATION

I. Introduction

General Issue

As modern technology continues to advance, how can we prevent the human from becoming the weakest component of the human-machine system? Recent advances in automation technologies make it conceivable for a single human operator to control multiple remotely piloted aircraft (RPA) simultaneously, a vast improvement in resource utilization compared to existing operations that require several operators per vehicle. The limiting factor of multi-aircraft-control (MAC) is the ability of the operator to expediently interpret information and attend to the increased number of time-critical subtasks. Automation can help alleviate operator task load, but it must be applied judiciously. Adaptive aiding is a strategy that uses an estimate of current operator workload to decide how and when to best apply automation. Workload is a representation of "the cost incurred by a human operator to achieve a particular level of performance" which "emerges from the interaction between the requirements of a task, the circumstances under which it is performed, and the skills, behaviors, and perceptions of the operator" (Hart & Staveland, 1988). One area of research attempts to infer operator workload from operator physiological data, such as electroencephalography (EEG) and electrocardiography (ECG). An informed learning model which accurately estimates workload can be fitted using these physiological parameters (G. F. Wilson & Russell, 2004).

Problem Statement

Model generalizability is an important aspect of any potential adaptive aiding system: ideally, a continuous workload prediction model would be able to make accurate workload estimates over a wide array of mission contexts and personnel. One potential disadvantage of machine learning-based models is that they may perform well in the specific contexts in which they were trained, but poorly in other contexts. If a model is not robust, it would require that exhaustive training data be gathered for every possible operational scenario and from every new human operator. From each of these contexts, a separate model would need to be fitted and maintained. Clearly, having to maintain these separate models is a non-scalable solution for missions composed of a wide variety of operations and operators.

This research effort examines the cross-applicability of machine learning-based models using data from a simulated RPA human performance study. As a proof of concept, we compare the cross-applicability of models created using linear regression, regression trees, and Artificial Neural Networks to relate workload to human physiological data. The models are evaluated in terms of cross-application error under several context-switch scenarios – across two tasks, four task conditions (two per task), and seven human operators.

Research Objectives

The main objective of this research is to identify which of the three algorithms is the best candidate for developing robust models for workload estimation for use in a single operator RPA adaptive aiding application. This research effort also compares the

cross-application error of specific contexts, and examines the effect of reducing the set of physiological inputs on cross-application error.

Investigative Questions

The following questions outline the investigative trajectory of this research effort. The main research question is:

Can a machine learning-based workload prediction model achieve a reasonable standard of cross-application error when applied to a diverse range of experimental contexts?

Specific investigative questions include:

1. *Which algorithm is the best at cross-application workload estimation?*
2. *Which algorithm is the best at estimating workload within a specific context?*
3. *Are some contexts more generalizable than others?*
4. *Can reducing the number of input features significantly decrease cross-application error?*

Methodology

This research effort was partially inspired by *A Comparison of Artificial Neural Networks, Logistic Regressions, and Classification Trees for Modeling Mental Workload in Real-Time* (Fong, Sibley, Cole, Baldwin, & Coyne, 2010). Fong, et al. found that ANNs and classification trees were significantly better at estimating workload (based on three levels of task difficulty) than logistic regression for a numeric recall task. Previous research has indicated that supervised learning workload classifiers that detect periods of high workload based on physiological input and trigger the automation of key subtasks

can significantly improve operator performance in a simulated RPA environment (Christensen & Estepp, 2013). More recent efforts have attempted to perform workload regression – estimating workload as a numerical parameter from its mathematical relationship to operator physiological state parameters (Heger, Putze, & Schultz, 2010). Considering workload numerically instead of categorically allows for workload prediction models that can potentially estimate workload values from previously unseen physiological input data. By representing workload numerically instead of categorically, finer-grain automation decisions are possible. Furthermore, continuously variable (rather than categorical) automation can be employed which may better match the needs of the operator. A smoother employment of automation may prove even more effective at improving performance than previous efforts. This research effort utilizes numeric workload profiles generated using the Improved Performance Research Integration Tool (IMPRINT), a discrete event network modeling tool (Allender, Kelley, Archer, & Adkins, 1997). These profiles were generated with a 1-Hz sample rate from tailored IMPRINT models developed for each subject and each task the subject performed.

Assumptions/Limitations

This research relies heavily on the assumption that the ascribed workload values are an accurate representation of the task load imposed upon subjects at any given time during the task period. This research assumes that due care was taken during the collection and aggregation of the physiological data, and that subjects performed to the best of their ability during all phases of the experiment.

Implications

This research is part of an overall effort on behalf of Air Force Research Laboratory's Human Effectiveness Directorate to develop robust, accurate workload estimation models as part of their Sense-Assess-Augment taxonomy for human-centered research (Galster & Johnson, 2013). This research will directly contribute to the "Assess" portion of the taxonomy, and its success will help pave the way for future implementation of adaptive aiding strategies as part of the "Augment" portion.

Preview

This chapter outlined the necessity of developing an accurate and robust workload estimation model in order to effectively implement adaptive aiding strategies that would allow a single operator to control one or more RPA. Chapter II provides a background on previous human performance research using supervised learning methods to estimate workload from human physiological data. Chapter III outlines specific hypotheses and experiments for evaluating the algorithms' ability to create robust models. Chapter IV summarizes and interprets the results of the applied methodology from Chapter III. Chapter V evaluates the success of the research objectives, summarizes the results and their ensuing conclusions, and makes recommendations towards future avenues of inquiry.

II. Literature Review

This chapter provides a background on previous human performance research using machine learning to estimate workload from human physiological data. The first section describes commonly accepted methods for modeling and measuring workload. The second section identifies classes of human physiological response data commonly used for workload prediction, and the relative utility of each. The third section discusses the mathematical basis of the machine learning algorithms, as well as advantages and disadvantages of each. The fourth section addresses adaptive aiding strategies and successful implementations.

Workload Modeling and Measurement

One definition of workload is a "hypothetical construct that represents the cost incurred by a human operator to achieve a particular level of performance" which "emerges from the interaction between the requirements of a task, the circumstances under which it is performed, and the skills, behaviors, and perceptions of the operator" (Hart & Staveland, 1988). Since workload is the product of multiple situational factors, a good workload measure must generalize across multiple dimensions and not be specific to any particular task environment. Workload can be considered as a single categorical or numerical value, or as the summation of multiple sub-values. Multiple resource theory divides workload into distinct cognitive channels (visual, auditory, spatial, etc.) based on our ability to multitask effectively as long as no one channel is overloaded – e.g. walking and talking versus listening to two conversations (Wickens, 1984).

Many studies use inherent task difficulty to delineate between workload levels. For example, recalling two numbers from working memory may be associated with "low" workload, four numbers with "medium" workload, and six numbers with "high" workload (Fong et al., 2010). Using this method to train a physiological input-driven workload prediction model, the workload level chosen by the model is compared to the "true" workload based on predetermined task difficulty. Some studies establish a workload baseline by recording physiological data while the operator is looking at the task environment but not engaging with it (G. F. Wilson & Russell, 2003).

Another common method of workload measurement is subjective self-evaluation on behalf of the human subject. This method is useful in that it captures the subject's unique perception of task difficulty and reaction to imposed workload. The drawback is that surveying the subject usually interferes with the task being performed, so an evaluation at the end of the task period is meant to represent the average workload over the entire task period. This after-the-fact evaluation is subject to memory bias; the peak-end rule holds that people judge an experience based largely on its most intense point and its end (Kahneman, Fredrickson, Schreiber, & Redelmeier, 1993). The most basic assessment of subjective workload is simply asking operators to rate how difficult the task was on a numeric scale (Besson et al., 2013; Smith, Gevins, Brown, Karnik, & Du, 2001). One of the most commonly used workload surveys is the NASA Task Load Index (TLX), which incorporates multiple sources of workload in accordance with multiple resource theory (Hart & Staveland, 1988). The NASA TLX presents six workload subscales (Mental Demand, Physical Demand, Temporal Demand, Performance, Effort,

and Frustration) that the operator rates, then weights the results based on perceived importance to produce an overall workload rating on a 0 to 100 scale.

Another multiple resource theory-based workload representation is the Visual, Auditory, Cognitive, and Psychomotor (VACP) scale (McCracken & Aldrich, 1984). Component rating scales were developed by surveying a range of human factors experts using lists of matched verbal anchor pairs and having them indicate, for each pairing, which verb required a higher level of effort. The pair comparison frequencies were then used to develop interval scale values for each VACP component. The VACP scale was later expanded to 7 components, including speech and tactile components, and divided the psychomotor component into separate fine motor and gross motor components (Little et al., 1993). A table depicting the exapanded VACP rating scales with anchoring statement text descriptions can be found in Appendix A. Higher scale values indicate a greater degree of use of the resource component.

Classes of Physiological Inputs

Physiological measures are a useful indicator of workload because they are less interruptive than secondary task measures or subjective surveys, do not require the measurement of overt performance, and are inherently multidimensional and therefore can be expected to provide multiple views of operator workload (Kramer, 1990). Physiological metrics also offer continuous monitoring and may respond quickly to shifts in workload. The most widely used physiological data inputs typically fall into one of four categories: neurological activity, ocular activity, cardiopulmonary activity, and other (skin conductance, body temperature, passive drool, etc.). Many studies collect and

integrate multiple categories of data in order to enhance the accuracy of their workload prediction models.

Electrical activity in the brain is typically measured via electroencephalography (EEG) using a series of electrodes placed on twenty-one cranial locations and separated into six frequency bands between 0 and 40 Hz. Kamzanova, Kustubayeva and Matthews' research indicates that alpha-1 band (8-10.9 Hz) EEG activity increases as operator attention decreases during a vigilance task (long periods of observing an unchanging environment with infrequent target stimuli) (Kamzanova, Kustubayeva, & Matthews, 2012). Furthermore, the alpha-2 band (11-13.9 Hz) remains relatively constant in tasks that only require identifying target stimuli and not recalling information from working (short term) memory. Yin and Zhang identified theta (4-7 Hz) and gamma (32+ Hz) EEG bands as most salient to changes in workload (Yin & Zhang, 2014).

A variety of ocular data metrics can be calculated in real-time, such as percentage pupil closure (PERCLOS), raw eyelid closure, fixation duration, saccade duration, saccade velocity, saccade frequency, blink frequency, blink duration, and pupil diameter. There is no definitive scientific agreement as to which eye metric works best for workload assessment (Halverson, Estepp, Christensen, & Monnin, 2012). The majority of studies suggest that pupil size is the most effective ocular metric in an experimental setting, although it may prove less effective in real-world contexts where lighting can vary dramatically. Marshal demonstrated that eye metrics (blink duration, saccade frequency, and divergence) can be used to effectively discriminate between different cognitive states (relaxed versus engaged, focused versus distracted, and rested versus fatigued) with an average accuracy upwards of 70 percent across trials. (Marshall, 2007).

Van Orden, Limbert, Makeig, and Jung examined changes in eye metrics (blink frequency and duration, fixation frequency and dwell time, saccadic extent, and mean pupil diameter) as a function of task workload in a target identification memory task (Van Orden, Limbert, Makeig, & Jung, 2001). Nonlinear regression analysis found blink frequency, fixation frequency, and pupil diameter to be the most salient variables. Halverson, et al. also found pupil diameter and PERCLOS to be highly correlated with workload (Halverson et al., 2012). Fong, Sibley, Cole, Baldwin, and Coyne identified pupil divergence as the most salient input factor and more indicative of workload than either pupil diameter or fixation frequency (Fong et al., 2010). Wilson and Russell observed a strong correlation between electrooculography (EOG, the vertical or horizontal measurement of corneo-retinal standing potential) and workload in their Air Force Multi-Attribute Task Battery study (G. F. Wilson & Russell, 2003).

Cardiopulmonary measures are the least often utilized physiological metric, although they are often collected and integrated with EEG and EOG for completeness. Nikolova suggested that heart rate variability is a sensitive measure for examining mental effort, work stress, and operator functional state, and is actually more indicative of workload than actual or intrinsic heart rate (Nikolova, 2002). There is also evidence to suggest that heart rate could be highly representative of the gradual accumulation of fatigue (Yin & Zhang, 2014).

Regression Algorithms

This research effort compares the robustness of models created using three different supervised learning algorithms – linear regression, regression trees, and artificial

neural networks. Linear regression is an estimation of the linear function relating an independent input variable (or variables) to a dependent numeric output that minimizes the difference between observed and predicted output (Alpaydin, 2010). The relation can be expressed as

$$\boldsymbol{Y} = \boldsymbol{XW} + \boldsymbol{\epsilon} \tag{1}$$

Where

Y = the vector of output values of length p

X = the $n \times p$ matrix of input values

W = the input weight vector of length p

ε = the vector of zero mean Gaussian noise parameters of length n

n = the number of samples

p = the number of inputs

The expression is then solved for the W that minimizes the sum of the squared residuals. This is can easily be achieved by QR decomposition. Linear regression is an extremely simple means of performing a workload regression from physiological data. It is less time and memory intensive than the other two algorithms, and inferences about feature saliency can be easily drawn from the coefficient values of the resulting linear equation. The disadvantage of linear regression is that it oversimplifies the complex, often nonlinear relationship between workload and physiological data, and is less adept at predicting workload for new data.

A more versatile method for modeling regression as a nonlinear relation is the regression tree. Regression trees are constructed by analyzing all the input features and

determining which binary division of a single input feature best reduces the mean squared error of output prediction (Lawrence & Wright, 2001). The process is repeated for each portion of the data resulting from the first split until a stopping condition (such as minimum residual or samples per terminal node) is met, resulting in a hierarchical tree of uniquely defined nodes. Regression trees are more time and memory intensive than linear regression, but more adept at handling outlier data and predicting workload for new data samples. The disadvantage of regression trees is that they can be sensitive to small changes in the training data; eliminating even a few samples can result in radical changes in tree size and branch conditions. Furthermore, regression tree branch splits only consider one input feature at a time, and can overlook inherent relationships or dependencies that exist between features.

An Artificial Neural Network (ANN) is a directed graph that utilizes nonlinear transformation functions contained in "hidden" nodes connecting the input layer to the output layer that model the action potentials displayed by neurons in the brain. Using one or more hidden layers detects the salience of the input features by performing a nonlinear transformation on the input data into a feature space where the output data may become more easily separable (Haykin, 2009). An ANN learns the relation between input and output by iteratively adjusting the network edge weights according to the difference between the predicted output of a training sample and its true output value (Hagan & Menhaj, 1994). ANN training occurs in two phases – in the forward phase, the input signal is propagated through the network, layer by layer, until it reaches the output, which is a function of the edge weights and the neuron transfer functions. In the backward phase, an error signal is calculated by comparing the generated output to the desired

value, and then propagating the error backwards through the network, adjusting the weights appropriately. ANNs are adept at learning complex, nonlinear relationships and have a high tolerance for noisy data. However, neural networks rely on a substantial amount of training data for solution convergence, are time intensive, and obfuscate how individual features relate to workload. Penaranda and Baldwin demonstrated that neural networks can be used for robust workload prediction across both task and temporal shifts (Penaranda & Baldwin, 2012).

Adaptive Aiding

The ultimate goal of accurate workload estimation is to sense suboptimal workload levels and effectively adjust the current level of automation before performance is negatively affected (G. F. Wilson & Russell, 2003). Adaptive aiding implies adjusting the level of automation when current workload is "not at its optimal or desired levels by implementing proper adaptation strategies that can accommodate the differences" (Yoo, 2012). Adaptive aiding is achieved by reallocating tasks using three different allocation strategies. In complete allocation, functions are either completely controlled by the human operator or completely automated. In partial allocation, a function may be partially automated by applying fixed levels of automation, which should be appropriately selected to match the need of situational demands made on human capabilities. In gradual allocation, the level of automation is increased or decreased gradually until the demands are sufficiently satisfied. Considering workload numerically instead of categorically allows for workload prediction models that are better suited for applying gradual allocation, which may better match the needs of the operator. A

smoother employment of automation adjustment may prove even more effective at improving performance than previous efforts.

Christensen and Estepp examined the efficacy of applied adaptive aiding in a multiple remotely piloted aircraft (RPA) simulation environment (Christensen & Estepp, 2013). The study featured 10 participants monitoring the progress of RPA on two abutted computer screens as they flew a preplanned mission, with the performance metric equal to the percentage of targets successfully engaged. Automation consisted of changing the user interface to emphasize priority as well as automatically linking RPA to targets and displaying target images. The effectiveness of activating automated assistance via physiological feedback was tested against manual activation by the user and no automation. The adaptive aiding produced significantly improved performance (90 percent of targets successfully engaged) compared to manual aiding (84 percent) and no aiding (82 percent). Wilson and Russell utilized an ANN online workload classifier to remove the monitoring and communication tasks from the Air Force Mutli-Attribute Task Battery when the classifier detected high workload, resulting in a 44 percent reduction in tracking task error compared to the nonadaptive condition (G. F. Wilson & Russell, 2003). In a similar study, Wilson and Russell used an ANN workload classifier to trigger adaptive aiding in a simulated RPA environment, which improved individual performance by 50 percent compared to automation that was randomly asserted. (G. F. Wilson & Russell, 2007). Although past research suggests that adaptive aiding can significantly improve operator performance, it has also been shown that the incorrect application of automation can degrade performance (Kaber, Wright, Prinzel, & Clamann, 2005); (Dixon, Wickens, & Chang, 2004). Therefore, great care must be taken to ensure

that adaptive aiding systems are designed to adjust automation when and how it is most beneficial to the operator.

Summary

There are several accepted methods for representing and estimating workload, each with its own advantages and disadvantages. Past research indicates that there is no single physiological metric that is most effective for real-time workload estimation in all cases – a broad range of physiological data across multiple spectrums offers the most possible information about current operator state. It has also been shown that using physiologically-driven workload classifiers for adaptive aiding can significantly improve operator performance. Using machine learning for regression to create accurate and robust workload estimation models that can apply more nuanced adaptive aiding policies may prove even more effective at improving operator performance.

III. Methodology

The primary goal of this research effort is to investigate the robustness of machine learning-based workload prediction models in terms of cross-application error for a variety of experimental contexts. A more in depth view of the experimental dataset is shown, including a detailed description of the study, how the physiological data was aggregated, and how the workload time-series were generated. Specific hypotheses about model cross-application performance are posed, as well as descriptions of the experiments and statistical tests used to investigate those hypotheses. It should be noted that this research effort has the benefit of approaching the problem of workload estimation in on off-line environment, after the physiological data has been analyzed and aggregated and the associated workload profiles carefully generated by respective subject matter experts. However, examining how well machine learning-based models trained in one context can estimate workload in another context should yield valuable information towards the future development of robust on-line workload estimators for use in adaptive aiding systems.

Data Set

This research effort examined physiological data from an Air Force Research Laboratory (AFRL) human performance study, which monitored fourteen participants completing a series of simulated remotely piloted aircraft (RPA) operation tasks (Courtice et al., 2012). This study is part of an overall effort on behalf of the Human Effectiveness Directorate to develop accurate predictive workload models as part of the Sense-Assess-Augment taxonomy for human-centered research (Galster & Johnson,

2013). The goal of the study was to quantify cognitive states of RPA operators using a variety of physiological measures. The primary task environment was an RPA simulation environment utilizing two inter-coordinated software tools; the Vigilant Spirit Control Station (VSCS), which simulates the instrument and display panels used to manipulate the RPA's optical camera throughout a given mission, and the Multi-Modal Communication (MMC) tool used for sending audio prompts and receiving responses as a secondary task measure. Figure 1 depicts a screenshot of a typical VSCS task environment.

Figure 1 Operator's View of Task Environment (Courtice et al., 2012)

In the experimental scenarios, operators performed a simulated RPA mission that involved operating the RPA's optical camera within the simulated airspace and performing 'Surveillance' or 'Tracking' tasks. The object of the Surveillance task was to monitor a marketplace and attempt to locate four high value targets (HVTs). Each HVT carries an AK-47 rifle, as opposed to non-target distractors that carry a handgun, shovel, or nothing. Operators search the market by clicking where they desire the camera to center, and zooming in and out with the mouse scroll wheel to determine whether a

person is the HVT or a distractor. Once found, the HVT is tracked until he walks under one of twenty tents in the market, at which point the operator begins looking for the next HVT. Independent variables in the Surveillance task include the number of distractors (high or low), and visual sensor fuzz (either absent or present).

When the Surveillance task ended, operators had three minutes to complete the NASA-TLX questionnaire before the Tracking task began. Thirty seconds into the Tracking task, the first HVT walks out from underneath a tent and begins walking to a different tent where he gets on a motorcycle. The operator attempts to track the HVT as it leaves the market on the motorcycle and rides to a new location. In half of the trials, a second HVT leaves in a similar manner, thirty seconds after the first, and must also be tracked. If a HVT is lost, operators are instructed to zoom out and search the surrounding area in order to reacquire the HVT. In half of the trials the HVTs travel along urban roads and in the other half they travel along rural roads. Independent variables in the tracking task include number of HVTs (one or two) and route (urban or rural). Each operator completed 4 sessions of testing, with 4 trials per session, such that each subject experienced every combination of task conditions 4 times. Over the course of each trial, subjects completed a recurring secondary task by responding verbally to a question requiring them to perform simple mental arithmetic.

During the aforementioned scenarios, physiological data was collected from each of the operators using two monitoring systems; the CleveMed BioRadio 150 and the Smart Eye Pro. The BioRadio received electroencephalography (EEG) and electrooculography (EOG) inputs from the BioSemi ActiveTwo electrode skullcap, and electrocardiography (ECG) and respiration data from sensor electrodes placed on the

chest. The Smart Eye Pro utilized four infrared cameras and two infrared illuminators to capture highly detailed pupilometery data. This research effort examines 66 physiological features, including 56 EEG measures, 2 EOG measures, 2 ECG measures, 2 respiration measures, and 4 pupilometery measures. A table annotating the physiological features can be found in Appendix C.

The raw time-series physiological data was recorded at varying sensor rates. EEG and EOG data were recorded at 480 Hz. The EEG data underwent spectral analysis by AFRL, during which it was downsampled to 1Hz power spectral density values. The EOG data was analyzed for blinks and saccades and post-processed into a common 1 Hz sample rate. Fixation and blinkrate values were determined by counting the number of blinks or saccades in a 60-second rolling window. Respiration data was collected at approximately 18Hz, and fit with cubic splines in order to downsample to 1Hz. Pupillometry and ECG data were collected at approximately 60Hz and 1.5Hz, respectively, and aggregated using the same process as the respiration data.

Time-series workload profiles were developed using the Improved Performance Research Integration Tool (IMPRINT). Analysts determined start and end points for each subject's actions during the experiment, then used IMPRINT to model each trial as a series of discrete events at a sampling rate of 1 Hz. Each discrete event was then assigned expanded Visual Auditory Cognitive Psychomotor (VACP) workload values according to the level of work that each workload channel experienced at that second. Workload values were assigned according to a version of the expanded VACP scale adapted specifically for the study, which can be seen in Appendix B. Overall workload, the sum of the visual, auditory, cognitive, fine motor, and speech channel values, is used as the

predictive variable for training and testing the machine learning algorithms. Figure 2 shows a workload profile for a surveillance task trial.

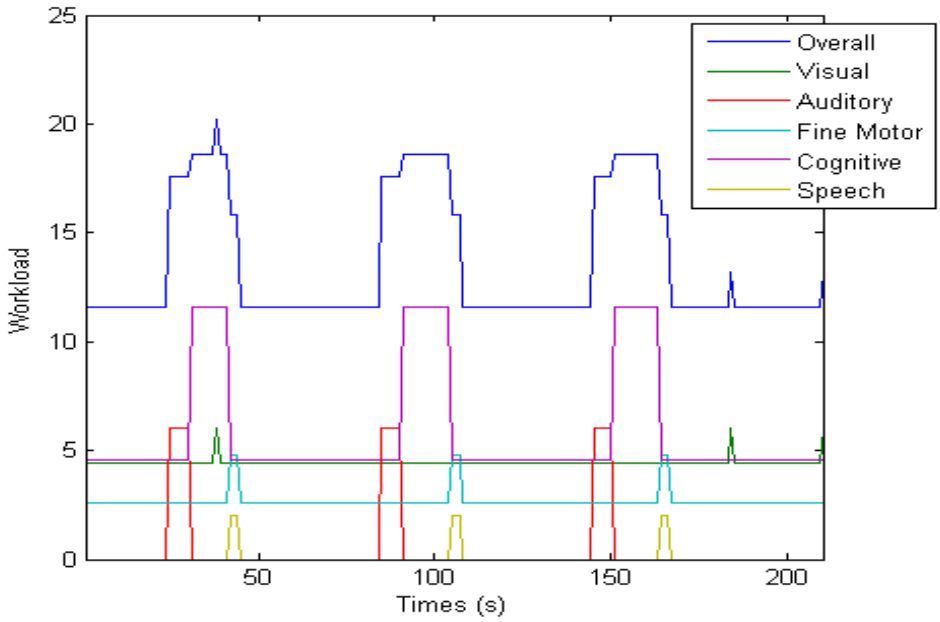

Figure 2 Workload Profile Example

The auditory workload spikes correspond to the secondary task questions, which occurred at regular intervals. The cognitive, fine motor, and speech workload spikes vary between trials according to how long it took subjects to respond to the prompts. As is often the case with human subject testing, some of the subject data was discarded due to subject or measurement issues; only data from subjects 2, 4, 5, 7, 8, 12, and 13 is considered in this analysis.

Testable Hypotheses

The primary focus of this research effort is to investigate the robustness of workload estimation models created using three supervised learning algorithms for regression. Models are evaluated in terms of root mean squared error (RMSE) to give an impression of the average error of a particular configuration in units of workload. Every

data sample from the study falls into a particular context: the task (Surveillance or Tracking), the task conditions (fuzz or no fuzz and high or low distractors for Surveillance, 1 or 2 HVT and urban or rural egress route for Tracking), and the subject (one of seven human operators). Ideally, a workload estimation model would be flexible enough so that it could estimate workload for samples outside of its training context just as well as it could estimate workload for samples inside its training context. Suppose that data from one particular context *A* is used to train a workload estimation model. The model is then used to estimate workload for previously unseen data from context *A* resulting in some root mean squared error $E_{A\to A}$. Then the model is used to estimate workload for data from the opposing context *B*, resulting in some root mean squared error $E_{A\to B}$. The process is then repeated, this time training a model on context *B*, resulting in a roots mean squared error $E_{B\to B}$ and $E_{B\to A}$. This research will test the hypotheses

$$\boldsymbol{H_1}: \begin{matrix} \boldsymbol{E_{A\to A} < E_{A\to B}} \\ \boldsymbol{E_{B\to B} < E_{B\to A}} \end{matrix} \tag{2}$$

against the null hypotheses

$$\boldsymbol{H_0}: \begin{matrix} \boldsymbol{E_{A\to A} = E_{B\to A}} \\ \boldsymbol{E_{B\to B} = E_{A\to B}} \end{matrix} \tag{3}$$

This hypothesis will be tested by comparing the distributions of RMSEs generated over 40-fold cross-validation (see Figure 3). The sample distribution of RMSEs can be assumed to be normal (by the central limit theorem), a requirement for the Student's t-test. In this research effort, null hypotheses will be rejected on a 95 percent confidence level, that is, if the p-value of a t-test is < 0.05, meaning a less than 5 percent chance of achieving those results by coincidence.

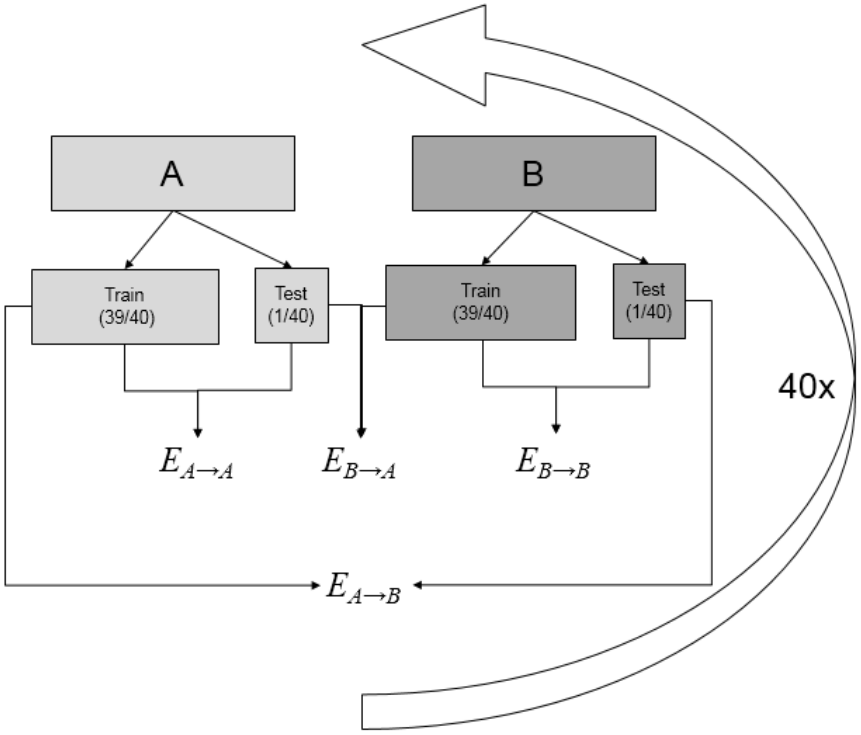

Figure 3 40-Fold Cross-validation

This research effort will test cross-application for three context categories – task type (Surveillance versus Tracking), task conditions (fuzz versus no fuzz, high distractors versus low distractors, 1 HVT versus 2 HVT, urban route vs rural route), and human subjects (6 subjects versus 1 individual subject). The desired outcome of these tests is failure to reject the null hypothesis on a 95 percent confidence level, indicating that a machine learning-based workload estimation model trained under one particular context can estimate workload for samples from the same context and the opposing context with approximately equal accuracy.

The cross-application RMSEs will also be compared against E_{μ}, the RMSE of a trivial predictor that ignores the physiological data completely and simply guesses the mean workload value of the training set for every sample of the testing set, specific to one context.

$$\boldsymbol{E_{\mu}} = \sqrt{\frac{\sum_{i=1}^{n}(W_i - W_{\mu})^2}{n}} \tag{4}$$

Where

E_{μ} = the root mean squared error of the trivial predictor

W_{μ} = the mean workload value of the training set

W_i = the *i*th workload value of the testing set

n = the number of samples in the testing set

Even if a model performs poorly at cross-application, it would be expected to at least outperform such a trivial predictor. This research will test the hypothesis

$$\boldsymbol{H_1}: \begin{matrix} \boldsymbol{E_{A\rightarrow B}} < \boldsymbol{E_{\mu(B)\rightarrow B}} \\ \boldsymbol{E_{B\rightarrow A}} < \boldsymbol{E_{\mu(A)\rightarrow A}} \end{matrix} \tag{5}$$

against the null hypothesis

$$H_0: \begin{matrix} E_{A\rightarrow B} = E_{\mu(B)\rightarrow B} \\ E_{B\rightarrow A} = E_{\mu(A)\rightarrow A} \end{matrix} \tag{6}$$

The desired outcome of these tests would be the rejection of the null hypothesis on a 95 percent confidence level, indicating that a machine learning-based workload estimation model trained under a specific context can estimate workload for samples from the opposing context with significantly less error than simply guessing the mean workload value for opposing context samples.

A secondary aim of this research effort is to investigate the effect of reducing the set of physiological inputs on cross-application error. Limiting the size and scope of the necessary physiological monitoring device helps to minimize both interference to the

operator and overall system lifecycle cost. The physiological data can be logically separated into four subsets based on how they were collected; EEG (from the electrode skullcap), EOG (from sensors extending from the skullcap placed around the eyes), Cardiopulminary (heart rate and respiration data from the chest sensors), and Pupilomtery (from the Smart Eye camera system). These subsets may be further grouped together into Skullcap (EEG and EOG), and BioRadio (Skullcap and Cardiopulminary), as the inclusion of one subset makes the other readily available. Although it might be assumed that having more input information would result in better model cross-application performance, it is also possible that discarding some features might enhance model robustness by relaxing the learned policy. This research will test the hypothesis

$$\boldsymbol{H_1}: \boldsymbol{E_s} < \boldsymbol{E_S} \tag{7}$$

against the null hypothesis

$$H_0: E_s = E_S \tag{8}$$

where E is the aggregation of both cross-application RMSEs $E_{A\rightarrow B}$ and $E_{B\rightarrow A}$, S is the a model utilizing all 66 available input features, and s is a model utilizing only the features included in one of the six identified subsets. The desired outcome of these tests would be the rejection of the null hypothesis on a 95 percent confidence level, indicating that cross-application error for a particular set of contexts can be significantly reduced using a subset of the input features. Although finding the optimal subset of all input features that minimizes cross-application error is nontrivial, it is beyond the scope of inquiry for this research effort.

Algorithm Specifications

Specific design choices were made in regards to the regression tree and ANN models with intent to enhance their robustness. Constructing a single regression tree using all the training data will usually over-fit the model, hampering its robustness (Lawrence & Wright, 2001). The risk of over-fitting a tree is reduced by employing a pruning scheme during 10-fold cross validation (generating ten trees from different subsets of 90 percent of the training data and validating on the remaining 10 percent). The bottom-most tree nodes are then iteratively removed until a minimal mean validation error across all ten trees is reached. The optimal pruning level is then applied to the original full tree.

ANN models are trained according to the Levenberg-Marquardt algorithm (see Appendix D). During each training epoch for an ANN model, 10 percent of the training data is randomly selected as a holdout validation set. Training continues until a maximum of 100 training epochs is reached, or the validation error increases for six consecutive epochs. This research effort utilizes a single hidden layer of 10 nodes for all ANN models. Although the optimization of the size and number of hidden layers and other training parameters is nontrivial, it is beyond the scope of enquiry for this research effort.

Summary

This research effort seeks to utilize supervised learning regression algorithms to relate a broad array of physiological data to operator workload for a simulated RPA reconnaissance task. It is hypothesized that a robust workload estimation model cannot estimate workload for samples outside of its training context as well as it can estimate

workload for samples inside its training context. It is hypothesized that a robust workload estimation model can estimate workload for samples outside of its training context better than a trivial predictor that ignores physiological input. It is also hypothesized that a model utilizing a natural subset of physiological input features will have significantly lower cross-application error than a model utilizing all available physiological input features. These hypothesis are tested by comparing the distributions of RMSE generated over 40-fold cross-validation.

IV. Analysis and Results

This chapter illustrates the results of the hypothesis tests based on a statistical comparison of the RMSEs generated over 40-fold cross-validation. Cross-application error is compared to self-application error and trivial predictor error for tasks, conditions, and subjects. A secondary evaluation of cross-application error investigates the effect of reducing the set of physiological input features to one of six defined subsets.

Task Cross-Application

Figure 4 shows the distributions of RMSE for the cross-application of Surveillance and Tracking data. On the y-axis, the letter to the left of the arrow indicates the type of data a model was trained on, and the letter to the right of the arrow indicates the type of data tested on. The horizontal dashed lines visually separate the results of the three models and the trivial predictors. The solid vertical lines indicate the respective RMSE means of the trivial predictors.

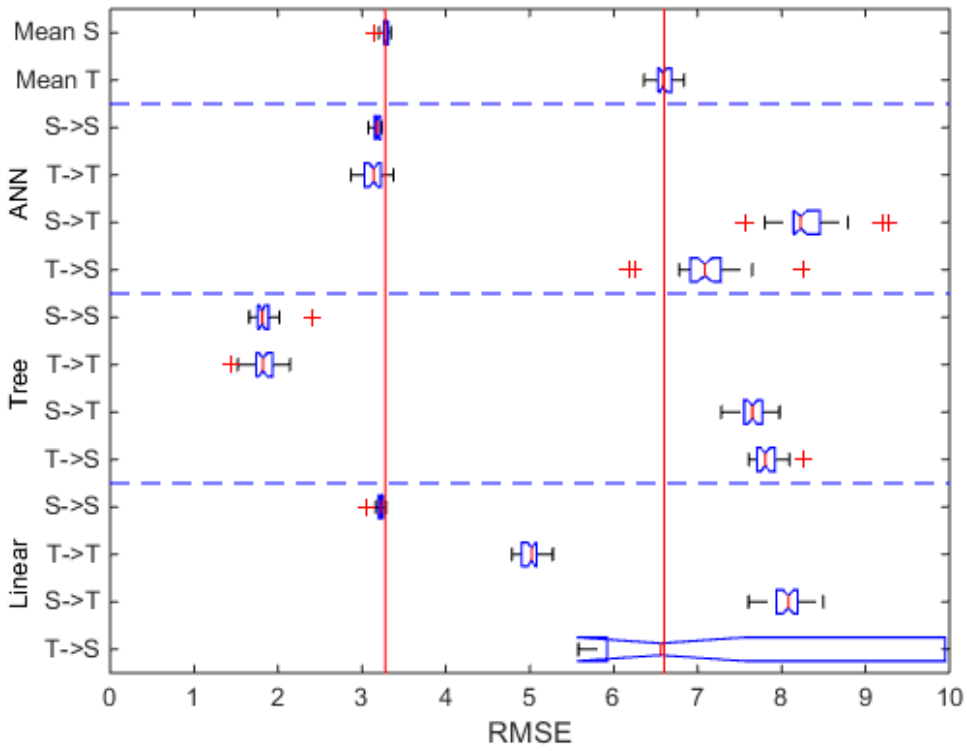

Figure 4 Task Cross-Application RMSE

As hypothesized, self-application (T→T, S→S) RMSE was significantly less than cross-application (T→S, S→T) RMSE for all three algorithms (p<0.01). No cross-application RMSE was significantly less than its respective trivial predictor RMSE (p>0.99). A summary of the results can be seen in Table 1. A bolded value indicates the null hypothesis could be rejected at a 95 percent confidence level.

Table 1 Task Cross-Application Mean RMSE Comparison

	$E_{A\to A} - E_{A\to B}$	$E_{A\to B} - E_{\mu(B)\to B}$		$E_{B\to B} - E_{B\to A}$	$E_{B\to A} - E_{\mu(A)\to A}$
Linear Regression	**-4.84**	1.46		**-8.56**	10.29
Regression Tree	**-5.82**	1.06		**-5.99**	4.54
ANN	**-5.13**	1.71		**-3.98**	3.82

A: Surveillance B: Tracking **Bold**: p-value<0.05

Figure 5 shows the distributions of task cross-application RMSE for the three algorithms relative to the 6 identified feature subsets. The vertical lines indicate the mean cross-application RMSE of the model utilizing all available features.

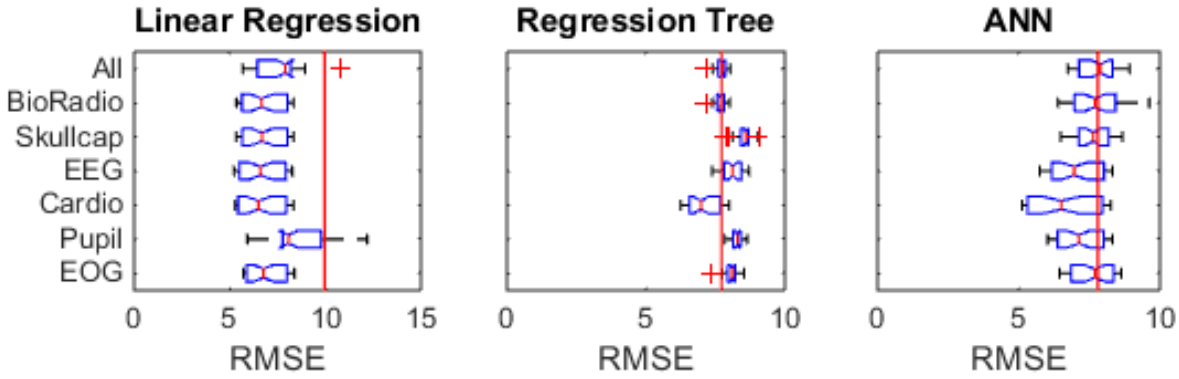

Figure 5 Task Cross-Application RMSE Relative to Input

Utilizing the EEG subset of features significantly reduced cross-application RMSE for all three algorithms, as did the EOG subset. A summary of the results can be seen in Table 2.

Table 2 Task Cross-Application Mean RMSE Relative to Input Comparison

	$E_{All} - E_{BioRadio}$	$E_{All} - E_{Skullcap}$	$E_{All} - E_{EEG}$	$E_{All} - E_{Cardio}$	$E_{All} - E_{Pupil}$	$E_{All} - E_{EOG}$
Linear Regression	4.22	**5.33**	**5.35**	1.54	-35.60	**5.28**
Regression Tree	**0.22**	**0.19**	**0.21**	**0.42**	-0.47	**0.47**
ANN	0.39	0.41	**0.64**	**0.68**	**0.50**	**0.53**

Bold: p-value<0.05

Condition Cross-Application

Figure 6 shows the distributions of RMSE for the cross-application of fuzz (F) and no fuzz (NF) condition data from the Surveillance task.

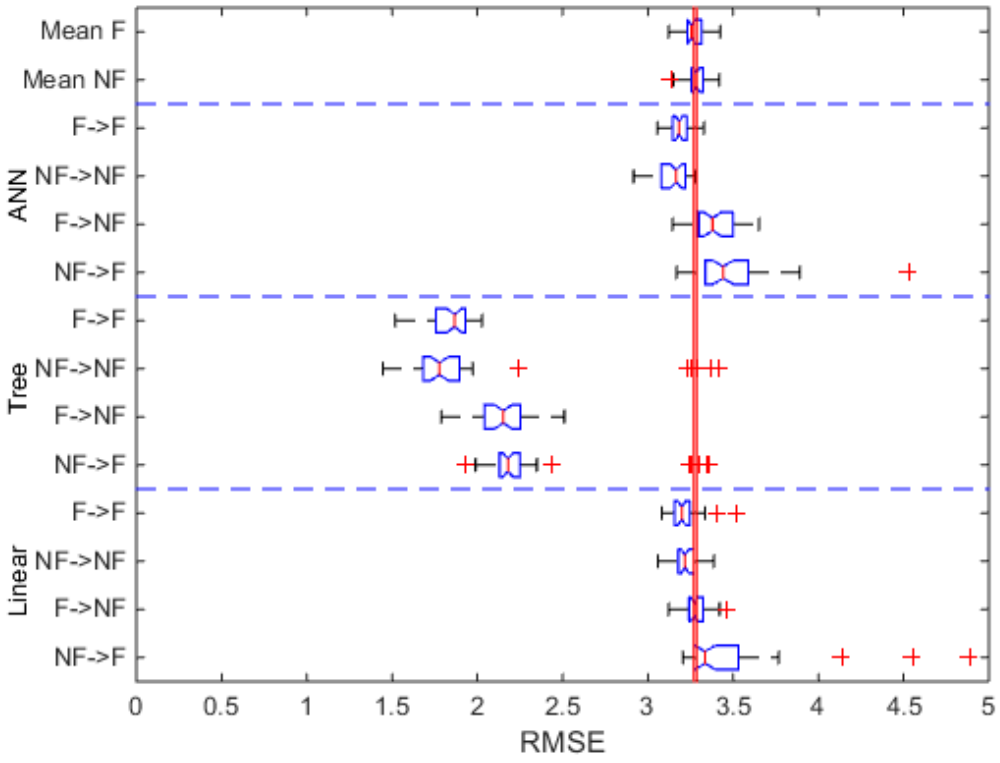

Figure 6 Fuzz Cross-Application RMSE

The results indicate that self-application RMSE was significantly less than cross-application RMSE for all algorithms (p<0.01). Linear regression cross-application RMSE was not significantly less than trivial predictor RMSE (p=0.43, 1.00), nor was ANN cross-application RMSE significantly less than trivial predictor RMSE (p=1.00, 0.99). A summary of the results can be seen in Table 3.

Table 3 Fuzz Cross-Application Mean RMSE Comparison

	$E_{A\rightarrow A} - E_{A\rightarrow B}$	$E_{A\rightarrow B} - E_{\mu(B)\rightarrow B}$		$E_{B\rightarrow B} - E_{B\rightarrow A}$	$E_{B\rightarrow A} - E_{\mu(A)\rightarrow A}$
Linear Regression	**-0.07**	-0.01		**-1.12**	1.07
Regression Tree	**-0.32**	**-1.13**		**-0.33**	**-0.97**
ANN	**-0.20**	0.10		**-0.41**	0.28

A: No Fuzz B: Fuzz **Bold**: p-value<0.05

Figure 7 shows the distributions of fuzz versus no fuzz cross-application RMSE for the three algorithms relative to the 6 identified feature subsets.

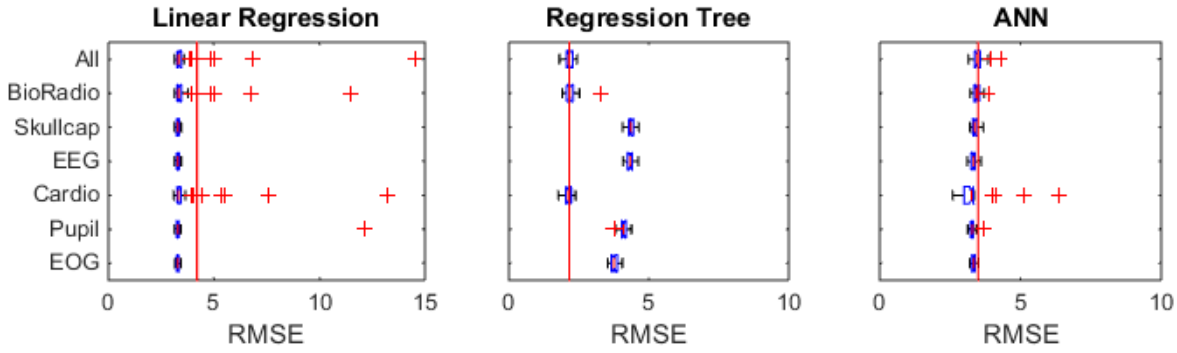

Figure 7 Fuzz Cross-Application RMSE Relative to Input

The linear EOG features model displayed the largest improvement in RMSE compared to the all-features model (p=0.03). A summary of the results can be seen in Table 4.

Table 4 Fuzz Cross-Application Mean RMSE Relative to Input Comparison

	$E_{All} - E_{BioRadio}$	$E_{All} - E_{Skullcap}$	$E_{All} - E_{EEG}$	$E_{All} - E_{Cardio}$	$E_{All} - E_{Pupil}$	$E_{All} - E_{EOG}$
Linear Regression	0.08	**0.89**	**0.90**	**-0.09**	0.69	**0.91**
Regression Tree	-0.04	-2.20	-2.18	0.04	-1.93	-1.60
ANN	0.05	**0.11**	**0.16**	**0.17**	**0.20**	**0.17**

Bold: p-value<0.05

Figure 8 shows the distributions of RMSE for the cross-application of high distractor (H) and low distractor (L) condition data from the Surveillance task.

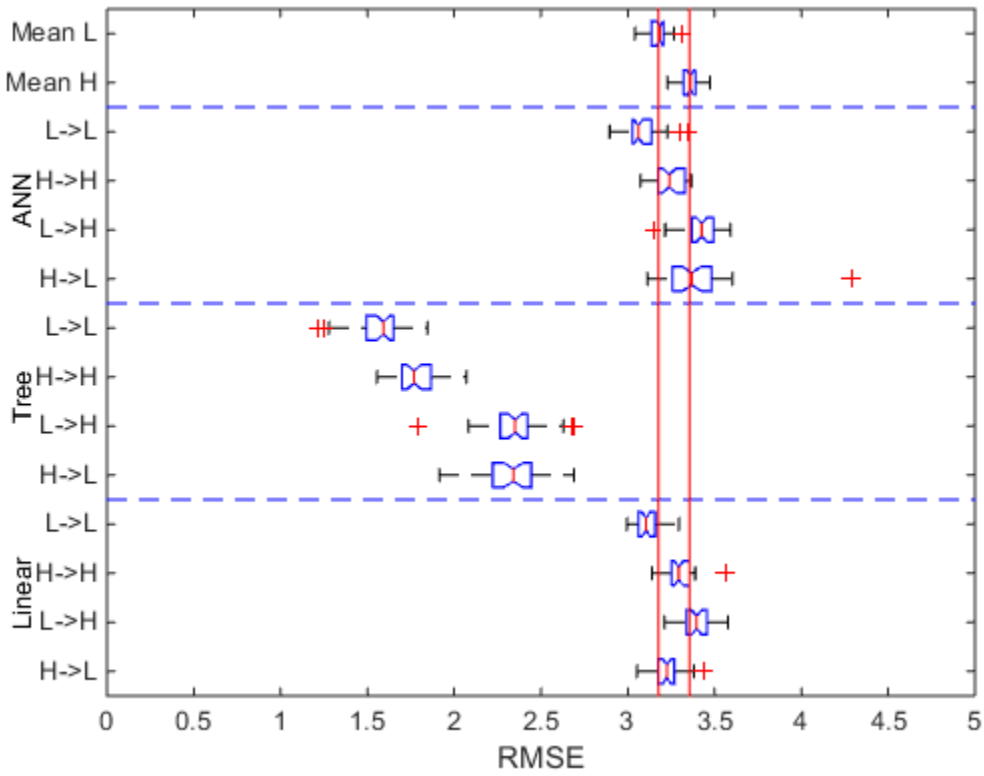

Figure 8 Distractors Cross-Application RMSE

Self-application RMSE was not significantly less than cross-application RMSE, with one exception for the linear model (p=0.999). Only the regression tree model was able to cross-estimate with significantly less error than the trivial predictors (p<0.01). A summary of the results can be seen in Table 5.

Table 5 Distractors Cross-Application Mean RMSE Comparison

	$E_{A\to A} - E_{A\to B}$	$E_{A\to B} - E_{\mu(B)\to B}$		$E_{B\to B} - E_{B\to A}$	$E_{B\to A} - E_{\mu(A)\to A}$
Linear Regression	**-0.32**	0.08		0.07	0.05
Regression Tree	**-0.79**	**-1.01**		**-0.55**	**-0.83**
ANN	**-0.34**	0.06		**-0.14**	0.20

A: Low Distractors B: High Distractors **Bold**: p-value<0.05

Figure 9 shows the distributions of high distractors versus low distractors cross-application RMSE for the three algorithms relative to the 6 identified feature subsets.

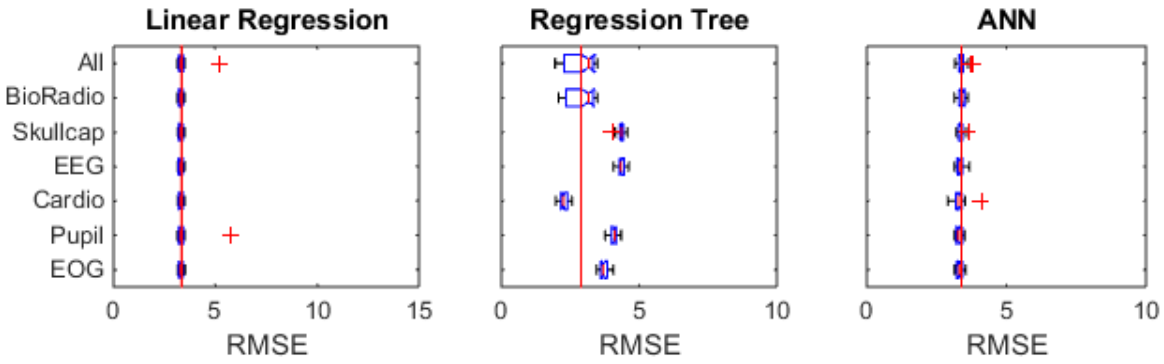

Figure 9 Distractors Cross-Application RMSE Relative to Input

Utilizing the Cardiopulmonary subset of features significantly reduced cross-application RMSE for both the regression tree (p=0.00) and ANN algorithms (p=0.01) compared to their respective all-features models. A summary of the results can be seen in Table 6.

Table 6 Distractors Cross-Application Mean RMSE Comparison

	E_{All} - $E_{BioRadio}$	E_{All} - $E_{Skullcap}$	E_{All} - E_{EEG}	E_{All} - E_{Cardio}	E_{All} - E_{Pupil}	E_{All} - E_{EOG}
Linear Regression	0.05	0.04	0.05	0.04	-0.02	0.04
Regression Tree	-0.01	-1.46	-1.46	**0.63**	-1.19	-0.82
ANN	-0.01	0.03	**0.05**	**0.09**	**0.09**	**0.07**

Bold: p-value<0.05

Figure 10 shows the distributions of RMSE for the cross-application of 1 HVT (1) and 2 HVT (2) condition data from the Tracking task.

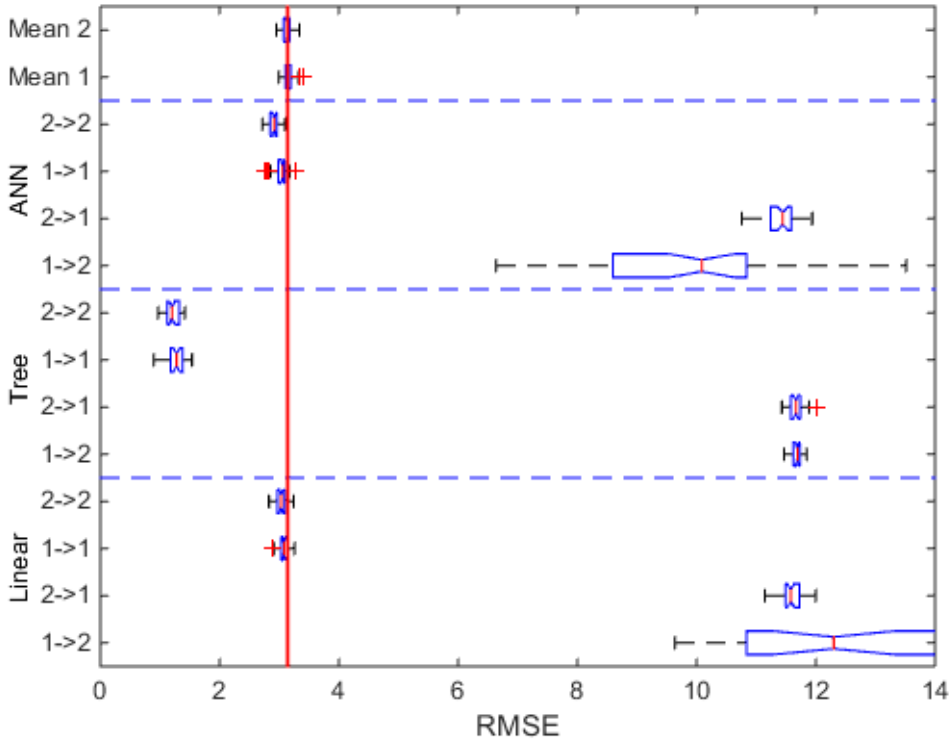

Figure 10 HVT Cross-Application RMSE

The results indicate that self-application RMSE was significantly less than cross-application RMSE for all algorithms (p<0.01). It is also indicated that cross-application RMSE was not significantly less than trivial predictor RMSE for any algorithm (p=1.00). A summary of the results can be seen in Table 7.

Table 7 HVT Cross-Application Mean RMSE Comparison

	$E_{A\rightarrow A} - E_{A\rightarrow B}$	$E_{A\rightarrow B} - E_{\mu(B)\rightarrow B}$		$E_{B\rightarrow B} - E_{B\rightarrow A}$	$E_{B\rightarrow A} - E_{\mu(A)\rightarrow A}$
Linear Regression	**-8.76**	8.64		**-9.92**	9.87
Regression Tree	**-10.46**	8.52		**-10.42**	8.55
ANN	**-8.51**	8.26		**-6.90**	6.80

A: 1 HVT B: 2 HVT **Bold**: p-value<0.05

Figure 11 shows the distributions of 1 HVT versus 2 HVT cross-application RMSE for the three algorithms relative to the 6 identified feature subsets.

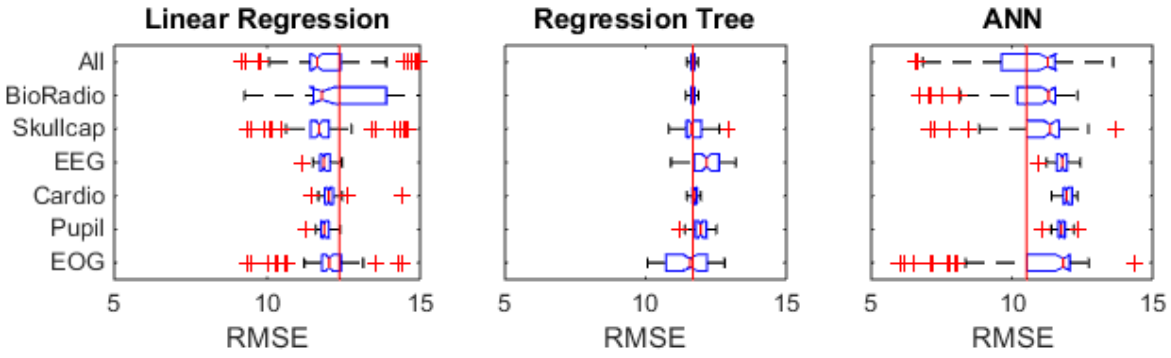

Figure 11 HVT Cross-Application RMSE Relative to Input

The largest significant decrease in cross-application RMSE is attributed to the linear regression model utilizing the EEG subset of input features (p=0.03). A summary of the results can be seen in Table 8. The linear regression model using only Pupilometry features also yielded a significantly lower cross-application RMSE than the all-features model (p=0.03).

Table 8 HVT Cross-Application Mean RMSE Comparison

	E_{All} - $E_{BioRadio}$	E_{All} - $E_{Skullcap}$	E_{All} - E_{EEG}	E_{All} - E_{Cardio}	E_{All} - E_{Pupil}	E_{All} - E_{EOG}
Linear Regression	-0.91	0.16	**0.48**	0.26	**0.46**	-0.37
Regression Tree	0.01	-0.01	-0.50	-0.05	-0.28	**0.21**
ANN	-0.16	-0.49	-1.26	-1.45	-1.25	-0.69

Bold: p-value<0.05

Figure 12 shows the distributions of RMSE for the cross-application of urban (U) and rural (R) condition data from the Tracking task.

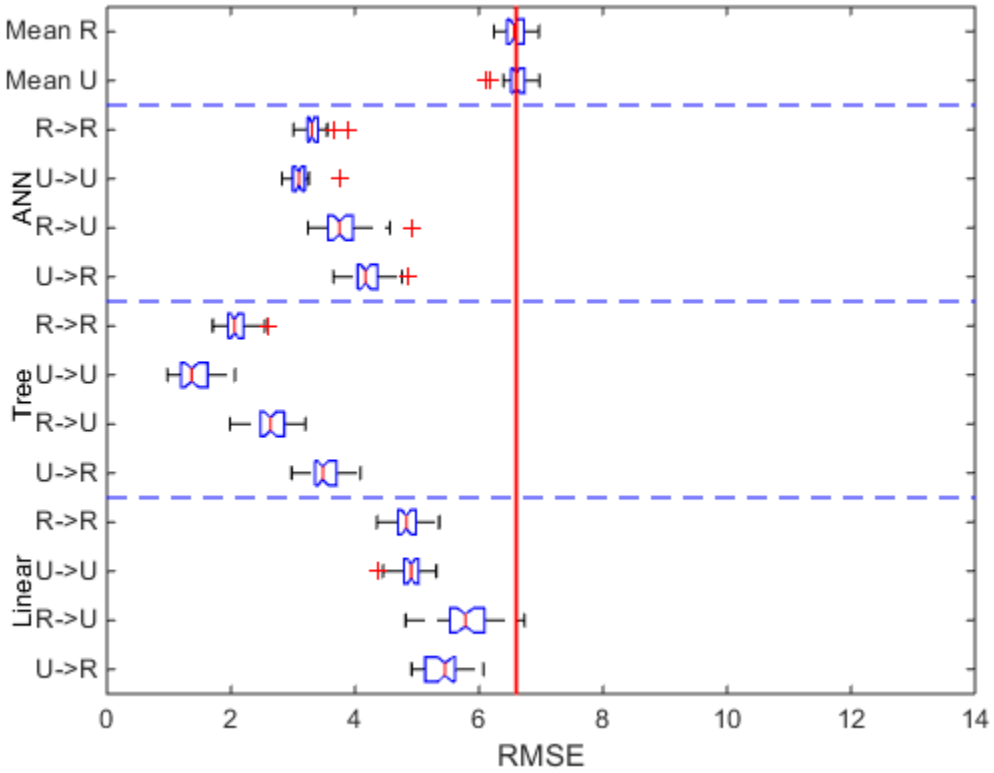

Figure 12 Route Cross-Application RMSE

The results indicate that self-application RMSE was significantly less than cross-application RMSE for all algorithms (p<0.01). It is also indicated that cross-application RMSE was significantly less than trivial predictor RMSE for all algorithms algorithm (p<0.01). A summary of the results can be seen in Table 9.

Table 9 Route Cross-Application Mean RMSE Comparison

	$E_{A\rightarrow A} - E_{A\rightarrow B}$	$E_{A\rightarrow B} - E_{\mu(B)\rightarrow B}$		$E_{B\rightarrow B} - E_{B\rightarrow A}$	$E_{B\rightarrow A} - E_{\mu(A)\rightarrow A}$
Linear Regression	**-0.98**	**-0.80**		**-0.54**	**-1.16**
Regression Tree	**-0.55**	**-3.97**		**-2.11**	**-3.08**
ANN	**-0.47**	**-2.82**		**-1.10**	**-2.38**

A: Rural B: Urban **Bold**: p-value<0.05

Figure 13 shows the distributions of urban versus rural cross-application RMSE for the three algorithms relative to the 6 identified feature subsets.

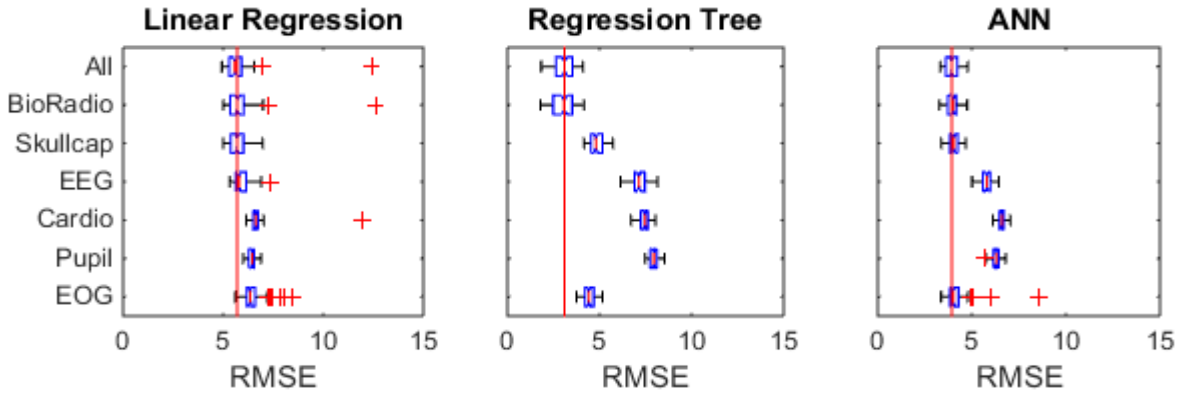

Figure 13 Route Cross-Application RMSE Relative to Input

The results indicate that none of the models utilizing an input feature subset significantly reduced cross-application RMSE compared to the all-features models. A summary of the results can be seen in Table 10.

Table 10 Route Cross-Application Mean RMSE Comparison

	E_{All} - $E_{BioRadio}$	E_{All} - $E_{Skullcap}$	E_{All} - E_{EEG}	E_{All} - E_{Cardio}	E_{All} - E_{Pupil}	E_{All} - E_{EOG}
Linear Regression	-0.13	-0.03	-0.20	-0.98	-0.72	-0.76
Regression Tree	0.06	-1.78	-4.07	-4.35	-4.84	-1.34
ANN	-0.02	-0.09	-1.86	-2.64	-2.33	-0.216

Bold: p-value<0.05

Subject Cross-Application

Figure 14 shows the distributions of RMSEs for the cross-application of data from 6 subjects and data from 1 subject. On the y-axis, "6" indicates data from 6 subjects and "1" indicates data from 1 subject.

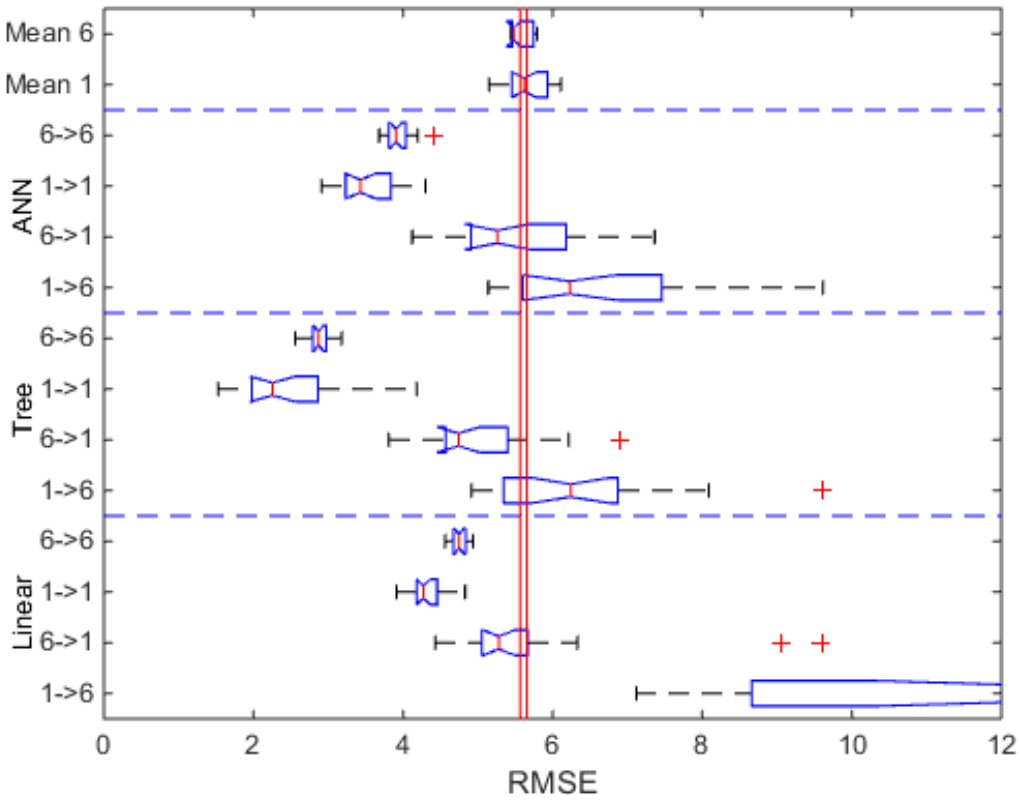

Figure 14 Subject Cross-Application RMSE

The results indicate that self-application RMSE was significantly less than cross-application RMSE for all algorithms (p<0.01). No cross-application RMSE was significantly less than the trivial predictor RMSE, with the exception of the regression tree 6→1 models (p<0.01). A summary of the results can be seen in Table 11.

Table 11 Subject Cross-Application Mean RMSE Comparison

	$E_{A\rightarrow A} - E_{A\rightarrow B}$	$E_{A\rightarrow B} - E_{\mu(B)\rightarrow B}$		$E_{B\rightarrow B} - E_{B\rightarrow A}$	$E_{B\rightarrow A} - E_{\mu(A)\rightarrow A}$
Linear Regression	**-0.95**	0.06		**-9.31**	8.04
Regression Tree	**-2.12**	**-0.65**		**-3.79**	0.76
ANN	**-1.61**	-0.11		**-3.44**	1.40

A: 6 Subjects B: 1 Subject **Bold**: p-value<0.05

Figure 15 shows the distributions of 6 subjects versus 1 subject cross-application RMSE for the three algorithms relative to the 6 identified feature subsets.

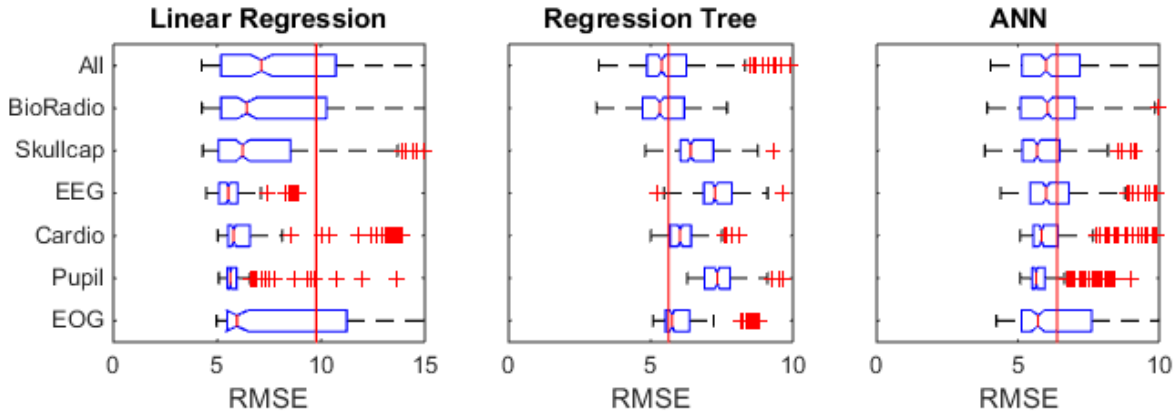

Figure 15 Subject Cross-Application RMSE Relative to Input

The results indicate that utilizing the Skullcap and Pupilometry feature subsets significantly reduced cross-application RMSE for both the linear regression and ANN algorithms (p<0.01). The largest decrease in cross-application RMSE was attributed to the linear model utilizing the EEG features subset (p=0.002). A summary of the results can be seen in Table 12.

Table 12 Subject Cross-Application Mean RMSE Comparison

	E_{All} - $E_{BioRadio}$	E_{All} - $E_{Skullcap}$	E_{All} - E_{EEG}	E_{All} - E_{Cardio}	E_{All} - E_{Pupil}	E_{All} - E_{EOG}
Linear Regression	0.91	**2.04**	**4.03**	**2.27**	**2.12**	0.54
Regression Tree	0.16	-0.98	-1.73	-0.50	-1.78	-0.39
ANN	0.04	**0.418**	0.08	-0.02	**0.51**	-0.88

Bold: p-value<0.05

Table 13 shows a comparison of the self-application RMSEs to cross-application RMSEs for all algorithms across all contexts.

Table 13 Mean RMSE Comparison

	Linear Regression			Regression Tree			ANN			
	Self Error	Cross Error	Δ	Self Error	Cross Error	Δ	Self Error	Cross Error	Δ	Avg. Δ
Surveillance	3.22	8.06	-4.84	1.83	7.66	-5.82	3.18	8.31	-5.13	-5.26
Tracking	5.00	13.57	-8.56	1.83	7.82	-5.99	3.13	7.11	-3.98	-6.18
Fuzz	3.21	3.28	-0.07	1.82	2.15	-0.32	3.19	3.39	-0.20	-0.20
No Fuzz	3.22	4.34	-1.13	1.96	2.30	-0.33	3.14	3.55	-0.41	-0.62
High Distractors	3.11	3.43	-0.32	1.56	2.35	-0.79	3.08	3.4	-0.34	-0.48
Low Distractors	3.30	3.23	0.07	1.80	2.35	-0.55	3.24	3.38	-0.14	-0.21
1 HVT	3.03	11.80	-8.76	1.21	11.67	-10.46	2.91	11.42	-8.51	-9.24
2 HVT	3.08	13.00	-9.92	1.27	11.69	-10.41	3.03	9.93	-6.90	-9.08
Rural	4.88	5.42	-0.54	1.40	3.51	-2.11	3.10	4.20	-1.10	-1.25
Urban	4.84	5.82	-0.98	2.10	2.65	-0.55	3.32	3.79	-0.47	-0.67
Subjects	4.76	5.71	-0.95	2.88	5.00	-2.12	3.93	5.54	-1.61	-1.56
Avg.	3.79	7.06	-3.27	1.79	5.38	-3.59	3.20	5.82	-2.62	

Investigative Questions Answered

Can a machine learning-based workload prediction model achieve a reasonable standard of cross-application error when applied to a diverse range of experimental contexts? Unfortunately, none of the algorithms were able to produce models that could estimate workload for samples outside their training context as well as they could estimate workload for samples inside their training context; the null hypothesis in Equation 3 was rejected for nearly all cases. However, the regression tree models were able to cross-estimate with significantly less error than the trivial predictors for both Surveillance conditions (see Tables 3 and 5), as well as the 6→1 subjects configuration (Table 11). All three algorithms produced models that were able to cross-estimate with significantly less error than the trivial predictors for the egress route (urban versus rural) condition (see Table 9).

Which algorithm is the best at cross-application workload estimation? The regression tree models had the lowest average cross-application error, followed by the ANN models and the linear regression models, respectively (see Table 13). On average, the ANN models exhibited the least difference between cross-application error and self-application error.

Which algorithm is the best at estimating workload within a specific context? The regression tree models exhibited the lowest self-application error for all contexts (see Table 13).

Are some contexts more generalizable than others? Yes, the high versus low distractors condition exhibited the least cross-application error, and the 1 versus 2 HVT condition exhibited the greatest cross-application error.

Can reducing the number of input features significantly decrease cross-application error? Yes, using a subset of the available input features significantly reduced cross-application error in all but one context (urban versus rural egress route). Models utilizing EEG features had significantly lower cross-prediction RMSE than then models using all features the most often, in 8 out of 18 possible instances. Although none of the six subsets strictly dominated the full feature set in all contexts, the results indicate that the complexity of the data collection apparatus can be reduced while also improving cross-application accuracy.

Summary

The algorithms were not able to construct models for this dataset that could estimate workload for samples inside of their training context and samples outside their training context with approximately equal accuracy. However, the regression tree models exhibited the most robustness in terms of minimal cross-application error, and were able to estimate workload across both Surveillance task conditions and human subjects within a reasonable limit. The regression tree models also exhibited the least error when training and testing on data from within the same context, making them the strongest candidate for developing accurate workload estimators for use in remotely piloted aircraft adaptive aiding systems. Reducing the set of physiological features simply based on how they were collected can significantly reduce cross-application error in some specific instances.

V. Conclusions and Recommendations

This chapter summarizes the significant findings derived from the conducted experiments. It contextualizes the role of this research effort within the field of human performance research, and makes recommendations towards how initial findings could be further expounded upon.

Conclusions of Research

The regression tree algorithm was the most successful at accurately estimating workload across experimental contexts, followed by the artificial neural network (ANN), with linear regression consistently performing the worst in terms of cross-application RMSE. The experimental results indicate that dynamic context-switch scenarios that change the nature of what the operator is doing (task type, number of HVTs) are more difficult to cross-predict than static contexts that merely alter the task environment (screen fuzz, non-target distractors). Cross-application error can be significantly reduced in some instances using a select subset of input features as opposed to the set of all features, with the added benefit of reducing the cost and cumbersomeness of the required physiological data collection device.

Significance of Research

Identifying workload estimation algorithms that can accurately cross-predict workload across multiple experimental variables, as well as classes of variables that hinder cross-application, is absolutely essential for the future development of robust workload estimation models for use in on-line adaptive aiding systems. The findings of this work contribute to the larger body of human performance research concerned with

accurate machine learning-based workload estimation. This research effort has contributed valuable information towards the future development of adaptive aiding systems that will allow a single operator to adeptly control multiple remotely piloted aircraft.

Recommendations for Future Research

Future research into increasing the robustness of machine learning-based workload estimation models may include identifying specific means of reducing workload cross-application error across dynamic contexts to the same level as static contexts. This may be achieved by utilizing physiological inputs or algorithms not utilized in this research effort. It may be beneficial to design and conduct an experiment that could titrate the minimum amount difference in a context-switch condition that significantly affects workload cross-prediction error, since every context in this experiment had that effect.

This research effort considered each sample independently and not as part of a time series. Future work might investigate if the trends identified here are maintained in a simulated on-line environment. Once truly robust workload estimation models have been developed, more nuanced on-line adaptive aiding systems that can take advantage of gradual allocation of automation can be implemented.

Summary

This research effort compared the ability of linear regression, regression trees, and artificial neural networks to create robust workload estimation models that could accurately cross-predict workload for a variety of experimental contexts. The results

indicate that the regression tree is the most adept at estimating workload within and across contexts and that dynamic contexts are harder to accurately cross-predict than static contexts. Reducing the set of input features based on means of measurement can significantly reduce cross-application error for certain contexts. The knowledge gained from this research effort will contribute to the ongoing development of accurate, robust workload estimation models that can effectively implement adaptive aiding strategies that may profoundly increase the capabilities of human operators in human-machine systems.

Appendix A: The Expanded VACP Scale

Channel	Value	Descriptors
VISUAL	0.0	No Visual Activity
	1.0	Visually Register/Detect (detect occurrence of image)
	3.0	Visually Inspect/Check (discrete inspection/static condition)
	4.0	Visually Locate/Align (selective orientation)
	4.4	Visually Track/Follow (maintain orientation)
	5.0	Visually Discriminate (detect visual difference)
	5.1	Visually Read (symbol)
	6.0	Visually Scan/Search/Monitor (continuous/serial inspection, multiple conditions)
AUDITORY	0.0	No Auditory Activity
	1.0	Detect/Register Sound (detect occurrence of sound).
	2.0	Orient to Sound (general orientation/attention)
	3.0	Interpret Semantic Content (speech, simple, 1-2 words)
	4.2	Orient to Sound (selective orientation/attention)
	4.3	Verify Auditory Feedback (detect occurrence of anticipated sound)
	6.0	Interpret Semantic Content (speech, complex, sentence)
	6.6	Discriminate Sound Characteristics (detect auditory differences)
	7.0	Interpret Sound Patterns (pulse rates, etc.)
COGNITIVE	0.0	No Cognitive Activity
	1.0	Automatic (simple association)
	1.2	Alternative Selection
	4.6	Evaluation/Judgment (consider single aspect)
	5.0	Sign/Signal Recognition
	5.3	Encoding/Decoding, Recall
	6.8	Evaluation/Judgment (consider several aspects)
	7.0	Estimation, Calculation, Conversion
FINE MOTOR	0.0	No Fine Motor Activity
	2.2	Discrete Actuation (button, toggle, trigger)
	2.6	Continuous Adjustive (flight controls, sensor control)
	4.6	Manipulative (tracking)
	5.5	Discrete Adjustment (rotary, vertical thumbwheel, lever position)
	6.5	Symbolic Production (writing)
	7.0	Serial Discrete Manipulation (keyboard entries)
GROSS MOTOR	0.0	No Gross Motor Activity
	1.0	Walking on level terrain
	2.0	Walking on uneven terrain
	3.0	Jogging on level terrain
	3.5	Heavy lifting
	5.0	Jogging on uneven terrain
	6.0	Complex climbing
SPEECH	0.0	No speech activity
	2.0	Simple (1-2 words)
	4.0	Complex (Sentence)
TACTILE	0.0	No tactile activity
	1.0	Alerting
	2.0	Simple discrimination
	4.0	Complex symbolic information

(Archer & Adkins, 1999)

Appendix B: Expanded VACP Scale Adapted for Study

Channel	Value	Descriptors
VISUAL	0.0	No Visual Activity
	4.4	Visually Track/Follow (maintain orientation)
	6.0	Visually Scan/Search/Monitor (continuous/serial inspection, multiple conditions)
	8.8	Visually Track/Follow (maintain orientation) x 2
	10.4	Visually Track/Follow + Visually Scan/Search/Monitor
	12.0	Visually Scan/Search/Monitor x 2
AUDITORY	0.0	No Auditory Activity
	6.0	Interpret Semantic Content (speech, complex, sentence)
COGNITIVE	0.0	No Cognitive Activity
	4.6	Evaluation/Judgment (consider single aspect)
	7.0	Estimation, Calculation, Conversion
	11.6	Evaluation/Judgment + Estimation, Calculation, Conversion
	16.2	Evaluation/Judgment + Estimation, Calculation, Conversion x 2
FINE MOTOR	0.0	No Fine Motor Activity
	2.2	Discrete Actuation (button, toggle, trigger)
	2.6	Continuous Adjustive (flight controls, sensor control)
	4.8	Manipulative (tracking)
	5.2	Discrete Adjustment (rotary, vertical thumbwheel, lever position)
	7.4	Serial Discrete Manipulation (keyboard entries)
SPEECH	0.0	No speech activity
	2.0	Simple (1-2 words)

Adapted from (Archer & Adkins, 1999)

Appendix C: Physiological Features List

No.	Name	Category	No.	Name	Category
1	F7 Alpha	EEG	40	T3 Gamma	EEG
2	F8 Alpha	EEG	41	Pz Gamma	EEG
3	Fz Alpha	EEG	42	O2 Gamma	EEG
4	T3 Alpha	EEG	43	F7 Delta	EEG
5	Pz Alpha	EEG	44	F8 Delta	EEG
6	O2 Alpha	EEG	45	Fz Delta	EEG
7	T4 Alpha	EEG	46	T3 Delta	EEG
8	F7 Beta	EEG	47	Pz Delta	EEG
9	F8 Beta	EEG	48	O2 Delta	EEG
10	Fz Beta	EEG	49	T4 Delta	EEG
11	T3 Beta	EEG	50	F7 Theta	EEG
12	Pz Beta	EEG	51	F8 Theta	EEG
13	O2 Beta	EEG	52	Fz Theta	EEG
14	T4 Beta	EEG	53	T3 Theta	EEG
15	F7 Gamma 1	EEG	54	Pz Theta	EEG
16	F8 Gamma 1	EEG	55	O2 Theta	EEG
17	Fz Gamma 1	EEG	56	T4 Theta	EEG
18	T3 Gamma 1	EEG	57	Heart Rate	Cardiopulmonary
19	Pz Gamma 1	EEG	58	Heart Rate Variability	Cardiopulmonary
20	O2 Gamma 1	EEG	59	Raw Pupil Diameter	Pupilometry
21	T4 Gamma 1	EEG	60	Raw Pupil Quality	Pupilometry
22	F7 Gamma 2	EEG	61	Filtered Pupil Diameter	Pupilometry
23	F8 Gamma 2	EEG	62	Filtered Pupil Quality	Pupilometry
24	Fz Gamma 2	EEG	63	Respiration Frequency	Cardiopulmonary
25	T3 Gamma 2	EEG	64	Respiration Amplitude	Cardiopulmonary
26	Pz Gamma 2	EEG	65	Blink Rate	EOG
27	O2 Gamma 2	EEG	66	Fixation	EOG
28	T4 Gamma 2	EEG			
29	F7 Gamma 3	EEG			
30	F8 Gamma 3	EEG			
31	Fz Gamma 3	EEG			
32	T3 Gamma 3	EEG			
33	Pz Gamma 3	EEG			
34	O2 Gamma 3	EEG			
35	T4 Gamma 3	EEG			
36	T4 Gamma	EEG			
37	F7 Gamma	EEG			
38	F8 Gamma	EEG			
39	Fz Gamma	EEG			

(Courtice et al., 2012)

Appendix D: Levenberg-Marquardt Algorithm for Neural Network Learning

Gradient descent is a simple method for artificial neural network learning (Hagan & Menhaj, 1994). Weights are updated according to

$$\Delta \boldsymbol{w_{ji}} = \boldsymbol{\eta \delta_j y_i} \tag{9}$$

Where

w_{ji} = the edge weight between nodes i and j

η = the learn rate parameter

δ_j = the local gradient of node j

y_i = the output of node i

The local gradient δ_j depends on whether neuron j is an output node or a hidden node. In the first case

$$\boldsymbol{\delta_j} = \boldsymbol{\varphi_j}'(\boldsymbol{v_j(n)})\,(\boldsymbol{y_j(n)} - \boldsymbol{d_j(n)}) \tag{10}$$

Where

$\varphi_j'(v_j(n))$ = the derivative of the activation function of node j

$y_j(n)$ = the output of node j given sample n

$d_j(n)$ = the desired output of node j given sample n

and in the case where node j is a hidden node, the local gradient is

$$\boldsymbol{\delta_j} = \boldsymbol{\varphi_j}'(\boldsymbol{v_j(n)}) \sum_{\boldsymbol{k}} \boldsymbol{\delta_k}\,(\boldsymbol{n}) \boldsymbol{w_{kj}(n)} \tag{11}$$

Where

$\varphi_j'(v_j(n))$ = the derivative of the activation function of node j

$\delta_k\,(n)$ = the local gradient of node k

$w_{kj}(n)$ = the edge weight between node *k* and node *j*

The Newtonian method dispenses with the tunable learn rate parameter by assuming that all local gradients are functions of linearly independent weights, such that

$$\Delta \boldsymbol{w} = -\boldsymbol{H}^{-1}\boldsymbol{\delta} \tag{12}$$

Where

w = the matrix of edge weights

H = the Hessian matrix i.e. the second-order derivatives of the error function with respect to the weights

δ = the local gradient vector

The Gauss-Newton algorithm avoids the difficulty of calculating the second-order derivative of the error function by approximating the Hessian matrix using the Jacobian matrix such that

$$\begin{aligned} \boldsymbol{H} &\approx \boldsymbol{J}^{\boldsymbol{T}}\boldsymbol{J} \\ \Delta \boldsymbol{w} &= -(\boldsymbol{J}^{\boldsymbol{T}}\boldsymbol{J})^{-1}\boldsymbol{J}\boldsymbol{e} \end{aligned} \tag{13}$$

Where

w = the matrix of edge weights

H = the Hessian matrix

J = the Jacobian matrix of all first-order partial derivatives of the total error function

e = the output error

The Levenberg–Marquardt algorithm can be considered a hybrid of the Gauss-Newton algorithm and gradient descent, altering its function based on the local solution space. It ensures that the approximation of the Hessian matrix (J^TJ) is always invertible by introducing another approximation.

$$\begin{aligned} \boldsymbol{H} &\approx \boldsymbol{J^T J} + \boldsymbol{\mu I} \\ \Delta \boldsymbol{w} &= -(\boldsymbol{J^T J} + \boldsymbol{\mu I})^{-1} \boldsymbol{J e_i} \end{aligned} \tag{14}$$

Where

H = the Hessian matrix

J = the Jacobian matrix

μ = the damping parameter

I = the identity matrix

When the combination coefficient is small, the Levenberg–Marquardt algorithm closely resembles the Gauss-Newton algorithm, which is faster and more accurate near a local minimum. The damping parameter is decremented with each training step that decreases performance error and is incremented when a training step increases performance error. This research effort utilized an initial damping parameter of 0.001, a decrement factor of 0.1, and an increment factor of 10.

Bibliography

Allender, L. E., Kelley, T. D., Archer, S., & Adkins, R. (1997). IMPRINT: The transition and further development of a soldier-system analysis tool. *MANPRINT Quarterly, 5*(1), 1-7.

Alpaydin, E. (2010). 4.6 regression. *Introduction to machine learning* (2nd ed., pp. 73). Cambridge, Mass.: MIT Press.

Archer, S., & Adkins, R. (1999). IMPRINT user's guide prepared for US army research laboratory. *Human Research and Engineering Directorate,*

Besson, P., Bourdin, C., Bringoux, L., Dousset, E., Maiano, C., Marqueste, T., . . . Vercher, J. (2013). Effectiveness of physiological and psychological features to estimate helicopter pilots' workload: A bayesian network approach. *IEEE Transactions on Intelligent Transportation Systems, 14*(4), 1872-1881.

Christensen, J. C., & Estepp, J. R. (2013). Coadaptive aiding and automation enhance operator performance. *Human Factors, 55*(5), 965-975.

Courtice, A., Geyer, A., Durkee, K., Caggiano, D., Pappada, S., Thoreson, J., . . . Hoepf, M. (2012). *Neurophysiological and behavioral data collection for defining operator states and performance assessment algorithms that drive adaptive aiding strategies.* (No. FWR20120132H). Dayton, OH: AFRL 711 HPW/RHCPA.

Dixon, S. R., Wickens, C. D., & Chang, D. (2004). Unmanned aerial vehicle flight control: False alarms versus misses. *Human Factors, 48*(1), 152-156.

Fong, A., Sibley, C., Cole, A., Baldwin, C., & Coyne, J. (2010). A comparison of artificial neural networks, logistic regressions, and classification trees for modeling mental workload in real-time. *Human Factors, 54*(19) 1709-1712.

Galster, S. M., & Johnson, E. M. (2013). *Sense-assess-augment: A Taxonomy for human effectiveness.* (No. AFRL-RH-WP-TM-2013-0002). Air Force Research Laboratory, Applied Neuroscience Branch, Wright-Patterson Air Force Base, OH 45433: AFRL.

Hagan, M. T., & Menhaj, M. B. (1994). Training feedforward networks with the marquardt algorithm. *IEEE Transactions on Neural Networks, 5*(6), 989-993.

Halverson, T., Estepp, J., Christensen, J., & Monnin, J. (2012). Classifying workload with eye movements in a complex task. *Human Factors, 56*(1), 168-172.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (task load index): Results of empirical and theoretical research. In P. Hancock, & N. Meshkati (Eds.), *Advances in psychology* (pp. 139-183). Amsterdam: North-Holland.

Haykin, S. (2009). Multilayer perceptrons. *Neural networks and learning machines* (3rd ed., pp. 122-229). Upper Saddle River, NJ: Pearson Prentice Hall.

Heger, D., Putze, F., & Schultz, T. (2010). Online workload recognition from EEG data during cognitive tests and human-machine interaction. *Lecture Notes in Computer Science, 6359*, 410-417.

Kaber, D. B., Wright, M. C., Prinzel, L. J., & Clamann, M. P. (2005). Adaptive automation of human-machine system information-processing functions. *Human Factors, 47*(4), 730-741.

Kahneman, D., Fredrickson, B. L., Schreiber, C. A., & Redelmeier, D. A. (1993). When more pain is preferred to less: Adding a better end. *Psychological Science, 4*(6), 401-405.

Kamzanova, A., Kustubayeva, A., & Matthews, G. (2012). Diagnostic monitoring of vigilance decrement using EEG workload indices. *Human Factors, 56*(1), 203-207.

Kramer, A. (1990). Physiological metrics of mental workload: A review of recent progress. In D. Damos (Ed.), *Multiple-task performance* (pp. 279-328). London: Taylor & Francis.

Lawrence, R. L., & Wright, A. (2001). Rule-based classification systems using classification and regression tree (CART) analysis. *Photogrammetric Engineering and Remote Sensing, 67*(10), 1137-1142.

Little, R., Dahl, S., Plott, B., Powers, J., Tillman, B., Davilla, D., . . . Wickens, C. (1993). *Crew reduction in armored vehicles ergonomic (CRAVES) study.* (No. A006). Hopkins, MN: Alliant Techsystems.

Marshall, S. P. (2007). Identifying cognitive state from eye metrics. *Aviation Space and Environmental Medicine, 78*(5), B165-B175.

McCracken, J. H., & Aldrich, T. B. (1984). *Analyses of selected LHX mission functions: Implications for operator workload and system automation goals.* (No. ASI479-024-84). Fort Rucker, AL: U.S. Army Research Institute Aviation Research and Development.

Nikolova, R. (2002). Functional determination of the operator state in the interaction of humans with automated systems. *NATO Research and Technology Organization Human Factors and Medicine Symposium,* Warsaw, Poland. *88*(1) 440-452.

Penaranda, B. N., & Baldwin, C. L. (2012). Temporal factors of EEG and artificial neural network classifiers of mental workload. *Human Factors, 56*(1), 188-192.

Smith, M. E., Gevins, A., Brown, H., Karnik, A., & Du, R. (2001). Monitoring task loading with multivariate EEG measures during complex forms of human-computer interaction. *Human Factors, 43*(3), 366-380.

Van Orden, K. F., Limbert, W., Makeig, S., & Jung, T. P. (2001). Eye activity correlates of workload during a visuospatial memory task. *Human Factors, 43*(1), 111-121.

Wickens, C. D. (1984). Processing resources in attention. In R. Parasuraman, & D. Davies (Eds.), *Varieties of attention* (pp. 63-102). San Diego: Academic Press.

Wilson, G. F., & Russell, C. A. (2004). Psychophysiologically determined adaptive aiding in a simulated UCAV task. In A. Vincenzi, M. Mustapha & A. Hancock (Eds.), *Human performance, situation awareness, and automation: Current research and trends* (pp. 200-204). Mahwah, NJ: Erlbaum Associates.

Wilson, G. F., & Russell, C. A. (2003). Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. *Human Factors, 45*(4), 635-643.

Wilson, G. F., & Russell, C. A. (2007). Performance enhancement in an uninhabited air vehicle task using psychophysiologically determined adaptive aiding. *Human Factors, 49*(6), 1005-1018.

Yin, Z., & Zhang, J. (2014). Operator functional state classification using least-square support vector machine based recursive feature elimination technique. *Computer Methods and Programs in Biomedicine, 113*(1), 101-115.

Yoo, H. (2012). Framework for designing adaptive automation. *Human Factors, 56*(1) 2133-2136.

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
26-03-2015	Master's Thesis	September 2013 – March 2015

4. TITLE AND SUBTITLE

Robust Models for Operator Workload Estimation

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Smith, Andrew M., 2nd Lieutenant, 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-15-M-064

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

14. ABSTRACT

When human-machine system operators are overwhelmed, judicious employment of automation can be beneficial. Ideally, a system which can accurately estimate current operator workload can make better choices when to employ automation. Supervised machine learning models can be trained to estimate workload in real time from operator physiological data. Unfortunately, estimating operator workload using trained models is limited: using a model trained in one context can yield poor estimation of workload in another. This research examines the utility of three algorithms (linear regression, regression trees, and Artificial Neural Networks) in terms of cross-application workload prediction. The study is conducted for a remotely piloted aircraft simulation under several context-switch scenarios – across two tasks, four task conditions, and seven human operators. Regression tree models were able to cross predict both task conditions of one task type within a reasonable level of error, and could accurately predict workload for one operator when trained on data from the other six. Six physiological input subsets were identified based on method of measurement, and were shown to produce superior cross-application models compared to models utilizing all input features in certain instances. Models utilizing only EEG features show the most potential for decreasing cross application error.

15. SUBJECT TERMS

Human Performance, Machine Learning, Regression Analysis, 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**	64	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 785-6565, ext 4612 brett.borghetti@afit.edu