DETECTING HUMAN MACHINE INTERACTION FINGERPRINTS IN CONTINUOUS EVENT DATA

A Dissertation

Submitted to the Faculty

of

Purdue University

by

Audrey E Reinert

In Partial Fulfillment of the

Requirements for the Degree

of

Doctor of Philosophy

August 2019

Purdue University

West Lafayette, Indiana

THE PURDUE UNIVERSITY GRADUATE SCHOOL
STATEMENT OF DISSERTATION APPROVAL

Dr. Steven Landry, Chair

School of Industrial Engineering

Dr. Barrett Caldwell

School of Industrial Engineering

Dr. Brandon Pitts

School of Industrial Engineering

Dr. Paul Parsons

Department of Computer Graphics Technology

Approved by:

Dr. Steven Landry

School of Industrial Engineering

To those who helped me get where I am today

ACKNOWLEDGMENTS

I am grateful to all of those with whom I have had the pleasure to work during this and other projects during my time at Purdue. Each member of my Dissertation Committee has provided me extensive personal and professional guidance and taught me a great deal about scientific research.

TABLE OF CONTENTS

Page

Page

Page

Page

LIST OF TABLES

Table Page

LIST OF FIGURES

Figure Page

ABSTRACT

Reinert, Audrey E. Ph.D., Purdue University, August 2019. Detecting Human Machine Interaction Fingerprints In Continuous Event Data. Major Professor: Steven J. Landry.

There is a problem facing human factors and human computing interaction researchers. While laboratory studies can provide direct measures of human performance, these methods are insufficient when trying to determine if similar changes in human performance are observable in high volumes of continuous event data. However, continuous event data does not contain direct measures of human performance, but it could contain indirect measures. It is not known if indirect measures of human performance present in continuous event data can be used to predict delay in responding to an unexpected event or assessing the operator‘s workload. By developing an interface with distinct difficulty levels that correlated with different measures of experienced workload we show that a set of variables exist that enable difficulty and response delay to be classified with 95% and 72% accuracy, respectively. Finally, there is evidence to suggest that the predictive accuracy is influenced by the sampling rate of the data and the size of the training set.

1. INTRODUCTION

1.1 Problem, Motivation and Approach

It is not presently known if a measure of human performance such as operator workload or reaction time can be inferred from data typically collected from an operating system. This is because the data collected by a system does not contain a clear indicator of human performance. For example, a car may automatically record information about lane position, acceleration and GPS location and use this data to model driving behavior. However, the car does not know if a change in driving behavior is due to the operator's emotional state, the operator's physiological state or a change in the environment.

There are reasons why using data recorded by a system would provide a more nuanced picture of operator performance. System data can be recorded at a near continuous rate. The data collected will be reflective of how the system is used in real operational conditions. Terabytes of data can be obtained at a marginal cost to a researcher. Finally, the data can be gathered unobtrusively without altering the driver. However, the data collected by a system does not contain a clear indicator of human performance.

Conversely, laboratory studies can directly measure changes in human performance. It is possible to measure a driver‘s blood alcohol content by having them use a breathalyzer when performing a task in the laboratory. There are downsides to relying exclusively on laboratory data. It is expensive and time consuming to collect a terabyte of performance data. The total number of participants in a laboratory study will be limited by financial and schedule constraints. Further, a researcher would need to demonstrate that performing the task in the laboratory is indistinguishable from performing the task in real operational conditions. This last point can be difficult

to clearly demonstrate. The distinction between these two types of data is further explained in the example below.

The purpose of this research is to show that human performance - in this case workload and reaction time - can be measured using data collected from a system. The following general method was used in this experiment. A web-application was constructed in which the amount of required to accomplish a task was manipulated. Participants performed the task under one of three difficulty conditions. Each difficulty condition corresponded with a different average subjective workload score as measured by a standard workload metric.

State data was recorded as the participant performed the task. Three types of categorization models were used to classify task difficulty, and three types of regression models were used to predict an individual‘s response time to an unexpected event. Various combinations of sampling rate, training set size and set of measures of participant experience were manipulated to test the sensitivity of the difficulty classification and delay prediction models.

1.2 Illustrative Example

Imagine two researchers have been approached by a car company and are tasked with identifying the set of behaviors that indicate when a driver is drunk. The first researcher designs an experiment where participants drive a car around a track a fixed number of times in a simulated environment. The participants are given a fixed amount of alcohol every few laps. The second researcher decides to buy one hundred terabytes of driving data from the car manufacturer.

The data produced in the laboratory by the first researcher contains a direct measure of performance in the form of blood alcohol content. This data can used to generate a model that correlates blood alcohol content with an observable measure of system performance such as braking distance. The first researcher could state with reasonable confidence that each fixed increase in blood alcohol content leads to

both a reduction in reaction time and an increases break distance. However, the first researcher may not be able to collect data from enough people to show that their findings will generalize to a larger population.

The second researcher will have access to data from hundreds of drivers that is naturally reflective of 'real'driving behavior. Any model the second researcher will produce will generalize to a large population. However, there is no objective record in the data of when the drivers were drinking. All the second researcher knows is that some of the data corresponds with a sober condition and some with a drinking condition.

This example demonstrates the relative strengths and weaknesses of both data sources. It is impossible to state with certainty that a subset of the data is indicative of a particular operator state without an objective measure of human performance. Laboratory data contains the objective measure of human performance but there may not be enough data to prove the findings will generalize.

1.3 Document Overview

The following thesis is divided eleven main chapters and three appendices. The background chapter contains an overview of human factors, data science and user experience research relevant to this work. The method chapter details the participant recruitment procedures, data collection techniques and experimental design.

The general data overview chapter details the results of the pilot study. This chapter contains summary statistics of the response data along with the results of statistical tests to determine if response delay and task difficulty are correlated. The next two chapters detail experimental results. The first of the two chapters details the results of models that predicted task difficulty. The second chapter the results of models that predicted response delay.

The sixth chapter discusses the results of difficulty prediction and response delay prediction models when the following features in the data were altered: a) the rate at

which data was sampled, b) the amount of data in the training set was altered, and c) measures of participant experience were included as predictor variables. The thesis closes with a summary of results, a discussion of the implications that arise from the work, the limitations of the current research and a discussion of future work.

There are three appendices included in the back of the thesis. The first appendix contains additional data visualizations of the results of response delay and task difficulty prediction models. The second appendix is a collection model performance tables. The first and second appendices show the same model performance data in different ways (See Appendixes A and B). The third appendix (Appendix C) provides additional academic context.

2. BACKGROUND

The following chapter is conceptually divided into four sections: (MATB-II), Workload, Gamification and Data Science. The (MATB-II) section details the development of the Multi-Attribute Task Battery and discusses how the tool is used in the human factors community. The Workload section discuss the different tools and techniques used to measure and evaluate subjective workload. The Gamification section discusses how games are used in human factors research. The data science sections provide an introduction to key data science concepts such as supervised learning. Additional context can be found in Appendix C

2.1 MATB-II

2.1.1 Definitions

The Multi-Attribute Task Battery (MATB-II) is a computer-based task designed to evaluate operator performance during system monitoring, tracking, communications monitoring, and resource management tasks. The MATB-II is analogous to activities airline crews perform in flight but can be used by non-subject matter experts. The interface consists of six parts: tracking, resource management, scheduling , system monitoring, communications, and pump status (Santiago-espada, Myer, Latorella, & Comstock, 2011) .

In the monitoring tasks, the participant responds to the presence of a red light, the absence of a green light and monitors for pointers for deviation from midpoint. Tracking tasks have the participant keeping the solid circle within the confines of the dotted box when the system is in manual. Otherwise, when the system is in automatic, they do not need to do anything. The scheduling window shows the beginnings and

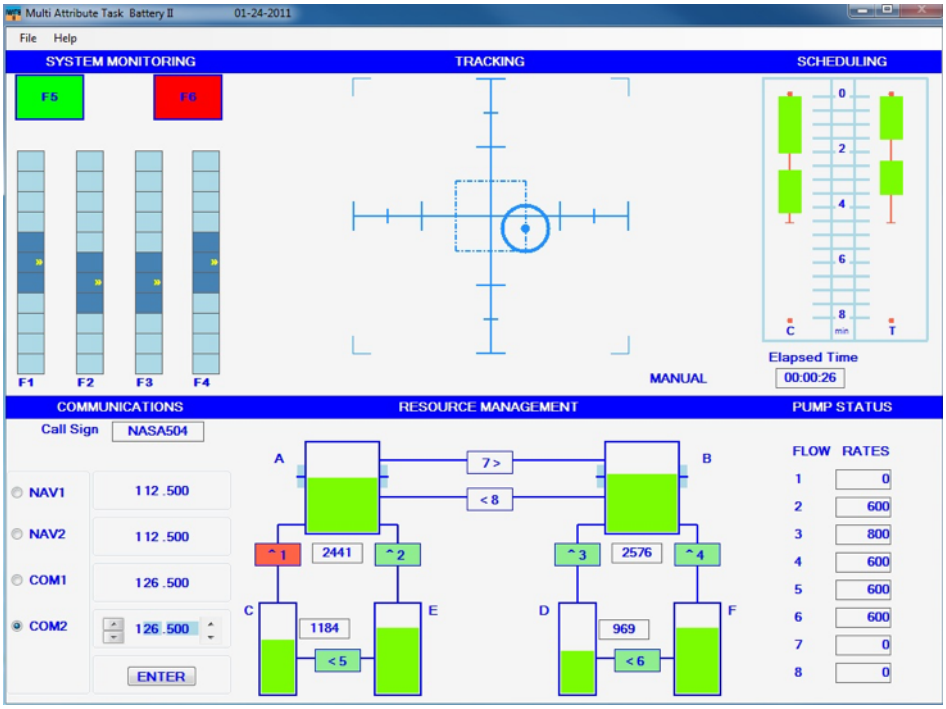


Fig. 2.1. MATB-II Interface

termination of tracking or communication tasks from the present time to 8 minutes into the future (Comstock Jr & Arnegard, 1992).

The communication task is designed to simulate receiving messages from air traffic control. The subject must discriminate their call sign from other three letter three number combinations. Finally, the goal of the resource management task is to maintain the fuel level of tanks A and B at a specific level. This is done by turning on and off specific pumps (Comstock Jr & Arnegard, 1992).

The MATB-II has a built-in Workload Assessment Tool based on the NASA Task Load Index This feature appears in a full window at specific pre-programmed intervals. The participant would use a slider feature to indicate how mentally, physically and temporally demanding they find the task in addition to a subjective measure of their current performance and effort. The last slider would indicate how frustrated the participant felt when performing the task.

MATB-II has been widely used in the human factors community as a research platform in part to the ease of adaption and extendibility. This widespread use has lead it to be used in studies on the effects of sleep deprivation on performance (Caldwell & Ramspott, 1998), task switching (Gutzwiller, Wickens, & Clegg, 2014) , attentional networks (Gutzwiller, Wickens, & Clegg, 2015) and time pressure (Gutzwiller, Wickens, & Clegg, 2016) . MATB-II can also be used as an assessment measure in multi-tasking studies (Chiappe, Conger, Liao, Caldwell, & Vu, 2013) and cognitive fatigue (Brooks, 2015).

2.1.2 Applications

Caldwell and Rampsott decomposed a thirty-minute MATB-II session into three ten-minute intervals to assess the impact of sleep deprivation on individual performance. Their study included eighteen male Army aviators and flight students. The authors noted that task durations of ten, twenty and thirty minutes were differentially sensitive to the effects of sleep loss. In particular, individuals reaction time to both communication calls and warning lights as well as tracking were more pronounced as the task duration increased (Caldwell & Ramspott, 1998).

Chiappe et al. examined where experience with action videogames could improve a participants ability to multi-task in high workload environments. The authors screened potential participants for game experience, selected fifty-three and divided them into groups-one control group and one manipulation group. Participants in the manipulation group were given a PS3 and encouraged to play five hours of games for ten weeks while the control group was given no such training. The authors found that the videogame treatment resulted in improved performance during the secondary tasks without interfering with the primary tasks of tracking and fuel management (Chiappe et al., 2013).

The authors noted that participants in the videogame condition could play more than five hours per week. They completed a correlational analysis of the number of

games played on performance. While the authors found negative correlations between reaction time and performance given gaming experience, in only one condition was this relationship significant (Chiappe et al., 2013).

A series of studies by Gutzwiller, Wickens and Clegg used the MATB-II to assess workload overload, individual differences in attentional networks and the effect of time on multi-task management. In their first study on workload, the authors told participants to perform four MATB-II tasks equally well or to prioritize the tracking task. The difficulty of the task was also modulated. The authors counted the number of task switches and found that participants were more likely to switch in easy conditions than in difficult conditions (Gutzwiller et al., 2014). The authors point to this finding as evidence a general switch cost avoidance behavior postulated by Kool (Kool, McGuire, Rosen, & Botvinick, 2010).

In Gutzwillers subsequent study, the authors posited that components of executive attention related to task switching behavior. In this study the authors included a measure of optimal switch times as a measure of executive attention. The authors found that when comparing low reaction time differences along the dimension of optimal verses non-optimal, that when optimal strategies were used participants made fewer switches. This result was not repeated for high reaction time differences (Gutzwiller et al., 2015).

The third and final Gutzwiller paper examined the role cognitive load plays on multi-task management. The authors studied task stability and cognitive load by having participants manage the sequential performance of four concurrent tasks. The authors found that the likelihood of task switching declines as memory load increases but increases as a function of task stability. In this context, task stability is derived from earlier work by Wickens (Wickens, 1986) which posits that dynamic systems have stable periods where an operator can neglect a specific task to focus on an alternate task (Gutzwiller et al., 2016) .

2.2 Workload

The simplest definition of workload according to Stephen Jex is the *amount of work an individual must exert to complete some task*. Workload can be classified in quantitative or qualitative terms (Jex, 1998) The research on workload distinguishes between perceived and actual workload. Further, workload can be used in terms of physical work ((how much does a participant need to lift?) or mental workload.

There are various tools that a researcher can employ to evaluate a participants mental workload. Per Meshkati et al. the majority of these measures fall into one of three techniques: physiological measures, performance-based measures and subjective measures. Performance based measures assume that an increase in the difficulty of a task will increase the demand on the participant. This increase in demand will result in lower performance. Physiological measures rest on the assumption that mental workload can be measured using levels of physiological activation. Finally, subjective measures posit that increased power expense is linked to perceived effort (Meshkati, Hancock, Rahimi, & M. Dawes, 1995).

No two workload measurement tools are equally suitable for the same task. Eggemeier et al. propose seven metrics a research can use to assess the suitability of a measure. These measures are: sensitivity, diagnosticity, selectivity, intrusiveness, reliability, implementation requirements and subject acceptability (Eggemeier, Wilson, Kramer, & Damos, 1991) . There are multiple outstanding measures of mental workload including: the Cooper-Harper Scale (Cooper & Harper Jr, 1969), the Bedford scale (Roscoe & Ellis, 1990) , the NASA Task Load Index (TLX) (Hart & Staveland, 1988), the Subjective Assessment Technique (Reid & Nygren, 1988) and the Workload Profile (Tsang & Velazquez, 1996).

The NASA-TLX is one of the more widely deployed measures of subjective workload and has been used in a variety of domains. The TLX measures workload using six sub-scales: mental demand, physical demand, temporal demand, effort, performance and frustration. Participants will provide a subjective rating -ranging from

0 to 100- for each of the sub-scales. The participant will then perform a pair-wise comparison. Each time a participant rates one scale as more salient than another, the greater the weight given to that sub-scale.

Cognitive load refers to the amount of mental effort being exerted. Cognitive load theory, developed by John Sweller, differentiates between three types of cognitive load: intrinsic, extraneous and germane Sweller1988Cognitive. Intrinsic cognitive load is the inherent difficulty of a specific topic. Extraneous load arises from the way information is presented to a user. Finally, germane cognitive load is the amount of effort devoted to developing and processing of schemas (Sweller, 1988).

Cognitive workload is recognized as a source of performance errors, but one of the key issues facing researchers who study cognitive workload is how to quantify it as a phenomena (Card, Moran, & Newell, 1983; Kirsh, 2000; Olson & Olson, 2000). A central problem when attempting to measuring workload is how to design a minimally invasive measure. A well-documented and validated method of measuring workload is the use of electroencephalography. Researchers can search for event related potentials(ERPs) in EEG data by looking for a change in amplitude (Gevins & Smith, 2003).

As a data source, EEGs change in predictable ways. This predictability allows for a degree of automatic detection using signal processing techniques. One signal component which is indicative of changes in workload is the P300 component of event-related brain signals. The P300 is recorded using electroencephalography and surfaces as a positive deflection of voltage with some latency (Polich, 2007). This component is elicited in the process of decision making. Fowler used the P300 as a measure of mental workload when studying a simulated aircraft landing task. Fowler found that the P300 amplitude was not strongly correlated with performance. However, the latency of the signal covaried with performance. According to Fowler, this latency was indicative of the slowing of perceptual and cognitive processes in response to increased workload (Fowler, 1994).

Tattersall and Foord found that workload measured using Instantaneous Self-Assessment (ISA) is consistent with other subjective measures of workload, it use

of this measure can interfere with the primary task (Tattersall & Foord, 1996). ISA measure was originally developed by the Civil Aviation authority of the United Kingdom to measure air traffic controllers workload (Jordan, 1992). ISA works by asking users to respond to a visual cue during as task using one of 5 keys. At the end of each interval, participants indicate their level of perceived workload. These levels are excessive, high, comfortable, relaxed, under-utilized.

Another psychophysiological measure of cognitive workload is the Index of Cognitive Activity. This measure works by measuring the individuals pupil dilation. This reliance on eye-tracking limits the ICA to visual displays (Marshall, 2002). Patten and colleagues studied measures of driver workload in two studies. The first study focused on driver distraction when using a cellphone as a function of task complexity. Participants drove on a series of different simulated roads (motorways, cities and rural areas) while having a conversation on a hands-free or hand-held device. Participants all drove the same route on a motorway. The researchers measured driver workload using the NASA TLX. Patten et. al. found that the content or complexity of the conversation had a greater influence over driver distraction than the complexity of the road (Patten, Kircher, stlund, & Nilsson, 2004).

Similar research addressed if a driver's level of experience changed their cognitive workload. The researchers found significant differences in experienced workload between experienced and inexperienced drivers. Further, inexperienced drivers were slower in reacting to a peripheral detection task. However, experienced drivers found high traffic and highly complex situations to be more taxing than their less experienced counterparts (Patten, Kircher, stlund, Nilsson, & Svenson, 2006).

We have all experienced both high and low physical and cognitive workload. This universal experience lends itself to an innate understanding of the concept. However, an activity that is seen as a high workload activity by one individual can be seen as a low workload activity by another because of subjective differences. This section outlined the techniques used by human factors engineers to transform the subjective experience of workload into a quantifiable metric.

2.3 Gamification

Gamification is the application of game playing elements - such as point scoring, competition, and rules of play- to other domains. The purpose of using gamification is to make a task feel more engaging by making an activity feel more challenging (Deterding, 2010a) or adding an overarching narrative structure (McGonigal, 2011). Gamification is typically applied to digital or computer based interactive experiences but the concept has historical roots.

Mark Nelson argues that the historical roots of gamification lie in marketing endeavors and educational systems in the form of scholastic achievement levels. Nelson noted that gamification efforts in the Soviet Union emphasized games as a tool to encourage productivity. Soviet games ranged from being purely competitive to morale-building exercises. The American model placed greater emphasis on capturing the sense of childhood play in an attempt to erase the work/play divide (Nelson, 2012).

Seaborn argues that interest in gamification has been renewed for three reasons. Over the past 20 years the computer games industry has grown and importance, leading to significant investments in research to understand what makes a game engaging. Web-based technologies, social media and mobile computing has changed how users and organizations modify, share and discuss experiences. Finally, firms are looking for new methods of connecting with influencing and learning about the experiences of customers and employees (Seaborn & Fels, 2015).

This leads to the question, what is a game? Some define games as intensely engaging but non-serious experiences structured by social boundaries and rules (Huizinga, 2007). Avedon and Sutton Smith defined games as a voluntary activity bounded by a set of rules but requiring conflict between equal parties and an unequal result (Avendon & Sutton-Smith, 1971). Others have defined games by six core features: rules, variables, quantifiable value laden outcomes, player effort and investment, and negotiable consequences (Juul, 2002).

Gamification works by directly manipulating the motivational controls of human behavior. Positive and negative reinforcing stimuli can be used to modify behavior (Skinner, 1938). Players will repeat behaviors which lead to satisfying outcomes while stopping those which leads to negative or undesired outcomes. Further, successful gamification is dependent on the repetition of outcomes (Robson, Plangger, Kietzmann, McCarthy, & Pitt, 2015).

There are questions about the theoretical ramifications of gamification. A meta-analysis by Hamari et al. found that 87% of applied gamification research failed to address theoretical foundations (Hamari, Koivisto, & Sarsa, 2014). When questions of theoretical foundations were raised, a few respondents referenced Deci et als. Work on intrinsic and external motivation (Deci, Koestner, & Ryan, 1999) . Other researchers cited Self-determination theory (Fallon et al., 2014; R. M. Ryan & Deci, 2000), operant conditioning (Skinner, 1938), ludic heuristics (Malone, 1982) and flow theory (Mirvis, Csikszentmihalyi, & Csikzentmihaly, 1991). Despite the mixed theoretical underpinnings, a consensus is emerging in two areas: intrinsic and extrinsic motivation and user centered design (Hamari & Tuunanen, 2014).

Modern video-games and gaming platforms collect gigabytes of data about a player's interactions with players, records of player movements and actions. Game designers use this data to drive player engagement. There is an explicit financial motive that compels game designers to collect this information. A poorly designed game leads to fewer players which leads to less money.

However, the same data that designers use to improve their designs can be employed by human factors engineers to study cooperation, communication and team behaviors. Games can be specifically modified to test how these behaviors emerge and changes in response to challenges. These behaviors can be recorded with the participant's explicit knowledge.

2.4 Data Science

This section discusses relevant research in data science. The first section focuses on the five Vs of Big data. The second section focuses on the link between Data Science and Human factors Engineering. The final section details research into bug data techniques

2.4.1 The Vs of Big Data

Big Data applications are defined by the volume, velocity, veracity, variety or variability, and value of the data being process. Volume in the context of big data applications refers to the amount of data being stored and operated upon. Velocity refers to the sampling or update frequency of the data itself. Data veracity refers to the degree which the data is trusted, precise and accurate. Variety, sometimes referred to as variability, is the number of different sources being sampled. Value is the ability to transform the data into a usable product (McAfee, Brynjolfsson, Davenport, et al., 2012).

2.4.2 Big data and human factors

The benefit of thinking of human factors problems through the lens of data science and big data has been well discussed in the literature (C. G. Drury, 2015). The benefits are often discussed in the context of digital sensors and how real time data can be used to reduce ergonomic risk (Walker & Strathie, 2016; Emanuele, 2016; Sharples & Houghton, 2017). None of these publications discuss how to use the acquired data in a predictive fashion.

There has been some research focusing on using continuous state data as to study human machine interaction problems . Dao et al. used pupil dilation as a way to measure pilot workload. The authors used information about error extent, gain, lag and delay from lateral path errors. They found that delay in responding to an error

was inversely proportional to workload (Dao, Parkinson, & Landry, 2016). However, this research did not use this data adaptively.

2.4.3 Supervised Learning

Supervised learning algorithms are a class of algorithms which use labeled training data to infer a function. The labeled training data consists of input-output pairs. The input objects usually take the form of multi-dimensional vectors (Mohri, Rostamizadeh, & Talwalkar, 2013). A supervised algorithm uses the training data to infer the function between the input vector and the output value. The aim is to build algorithms which can learn how to predict target output. However, algorithms are designed with one or more inductive biases. An Inductive bias is an assumption which an individual uses to predict specific outputs given inputs they have not trained on (Mitchell, 1980). Inductive biases-like heuristics- can be incorrect but without them learning algorithms would be able to predict or classify responses they had no prior experience with.

This section covers five types of supervised learning algorithms in detail: Neural Networks, Random Forests, Naive Bayesian Classifiers, Ensemble Learning and multinominal logistic regression. Each subsection details the specific uses cases of these algorithms along with their performance relative to each other.

Neural Networks

Neural Networks are a biologically inspired information processing paradigm. These networks are composed of many highly interconnected processing nodes which work in unison to solve a problem. These networks can be divided into three layers: an input layer, one of more hidden layers and an output layer.

How these layers communicate with each other determines the structure of the network. If the connections between the layers only feed from the input layer towards the output layer and do not form a cycle, then the network takes on a feed forward

structure (Zell, 1994). If the connections form a directed cycle, then a recurrent neural network is formed. Both neural network structures can be used for binary classification. As the name suggests, binary classification is the process of categorizing elements into two distinct groups. A specific type of linear binary classifier, which converts input into a numerical vector, is call are called the Perceptron (Freund & Schapire, 1999).

Neural networks 'learn 'how to classify an object or detect speech using one of two techniques: associative mapping and regularity detection. Associative mapping occurs when nodes associate certain firing patterns with specific patterns of input (Rojas, 1996) . Networks can also learning using regularity detection, where the nodes respond to properties of the input (Hasan, Choi, Neumann, Roy-Chowdhury, & Davis, 2016).

A programmer could use an associative learning paradigm to tune the weights between the connections in the following ways. Say the input is the number five. This activates output for eight, five and six. What the programmer can do is penalize connections between the nodes which lead to incorrect categorizations. Using regularity detection however, the programmer can encourage the network to look for sets of features such as the curve on one side but not the other.

This leads to the question, what are the trade offs which come from using a neural network. These networks can be trained to solve a variety of problems which would be difficult to solve using other methods. However, neural networks need numerical input which cannot have missing or incomplete values.

Random Forests

Random forests are an ensemble learning technique used for regression and classification tasks. A random forest operates by developing a multitude of decision trees during training. When presented with testing data, a random forest model will return the most common class in classification problems or the mean prediction in the

case of regression (Ho, 1995; Barandiaran, 1998). Random forest models have been employed in the following disciplines and problems : remote sensing (Belgiu & Drgu, 2016; Pal, 2005), Sleep stage identification (Fraiwan, Lweesy, Khasawneh, Wenz, & Dickhaus, 2012), ecology (Cutler et al., 2007), land usage (Rodriguez-Galiano, Chica-Olmo, Abarca-Hernandez, Atkinson, & Jeganathan, 2012),modeling the effect of climate change on coral (Fabina, Baskett, & Gross, 2015), and forest management (Atkins, Epstein, & Welsch, 2018).

Naive Bayesian Classifiers

Naive Bayesian Classifiers are similar to ensemble learning techniques in that they are a family of classifiers which rely on applications of Bayesian Reasoning. These classifiers assume that each of the features under consideration are strongly independent of each other. This independence is best explained through the following example. A fruit classifier would label an object a banana if it is 15 centimeters long, yellow and curved. However, each of these features contribute independently to the likelihood of the item being a banana despite any correlation between features.

Nave Bayesian Methods have been seen as the punching bag of classifiers despite having an extensive body of literature devoted to it from the field of information retrieval devoted to it (Lewis, 1998). This is not to say that criticisms of the algorithm are baseless as the algorithm tends to appear at the bottom of performance rankings when it comes to accuracy . However, the algorithm is still used as a baseline because other, more accurate algorithms, tend to be more complex (Rennie, Shih, Teevan, & Karger, 2003).

Here is how one would use a Naive Bayesian Classifier to categorize if an object was an apple or a banana. The measured features are color -measured in wavelength-length/height, and diameter. The researcher would first take a set of training data and build a classifier assuming a Gaussian distribution. This classifier would have the average wavelength of bananas and apples along with the variance of those samples.

Then, the testing sample would be presented. The research would then need to calculate the initial evidence that an item is either a banana or apple. Once the evidence value is known, the researcher needs only calculate the posterior probability of both apple and banana and pick the greater value.

Ensemble Learning

Machine learning can be characterized in terms of searching a hypothesis space for the most accurate hypothesis. There are cases, as explained by (Dietterich, 1997) , where a single best learner or hypothesis may not be found. First, the provided training data may not contain enough information to identify a best learner. There may even be cases where there are multiple bet learners. Second, the process used to obtain the best learner could be inaccurate or inefficient. Finally, the space being search may not contain the target function.

Ensemble learning is a means of overcoming these limitations. Ensemble learning is defined as:

> machine learning paradigm where multiple learners are trained to solve the same problem. In contrast to ordinary machine learning approaches which try to learn one hypothesis from training data, ensemble methods try to construct a set of hypotheses and combine them to use. (Chora et al., 2009)

Ensemble learning by-passes the lack of a single best hypothesis by generating multiple hypothesis which can approximate the best solution. Even if the method of finding the best hypothesis is inefficient, the process used to find 'good enough 'solutions may not be. Finally, ensemble solutions can form an approximation of the target function.

Each ensemble consists of a finite set of base learners. These base learners can be weak learners- meaning they perform only slightly better than a random guess - or strong learners- which can make highly accurate predictions. The process of turning

weak learners into strong leaners is called boosting and is discussed in more detail in a later section (Schapire, 1990). However, one advantage of using ensembles is that their collective predictive performance is better than the single best classifier among them (Hansen & Salamon, 1990).

Ensembles are constructed using parallel or serial approaches . Parallel ensembles combine accurate and diverse base learners which were constructed independently from each other. As such, each learner makes different errors when presented with new data but each learner must have an error rate better than random chance (Tuv, 2006). As a collective, these learners perform better than any individual since their errors cancel out. Serial ensembles rely on the previous generation of experts.

There are several types of ensembles discussed in the literature. However, most methods can be traced back to Bootstrap aggregation and Boosting. Bootstrap aggregation (Bagging) works by training each base learner on a different bootstrapped sample drawn from the original dataset (Breiman, 1996).

l

Multinomial Logistic Regression

Multinomial logistic regression (MLR) is a generalized form of a logistic regression that can be applied to multi-class problems. This technique is frequently used to analyze categorical response data using categorical or continuous variables (Bull & Donner, 1987). MLR is similar to ordinal logistic regression (OLR). Campbell et al. provide a clear explanation of the difference between these methods. An example of an ordinal ordinal would be whether an individual will be healthy, suffer an non-fatal heart attack or stroke and if they will suffer a fatal heart attack . Non-ordinal outcomes would be if a person was healthy, died of a heart attack or died of cancer (Campbell & Donner, 1989).

There are several advantages to using an MLR compared to other methods. First, MLR and by extension logistic regression- use an odds ration as an estimator for

the predictor variable. Unlike, discriminant analysis- a separate type of classification model- MLR models can use categorical and continuous variables as predictors (Stevens, 1992). Finally, unlike discrimination analyses, MLR does not require the independent variables to follow a normal distribution (Hossain, Wright, & Petersen, 2002).

There are also limitations to using MLR in classification problems. The researcher needs a sufficiently large sample size across all levels of both the independent and dependent variable (Hossain et al., 2002). There is a possibility that MLR models can become difficult to interpret if there are more than 4 dependent variables (Hosmer & Lemeshow, 2000).

MLR models have been extensively adopted by social work and human factors practitioners. One such analysis was conducted in 2012 by Ghulam et al. who used multi-nominal logistic regression to investigate single and multi-vehicle accidents (H., S., & R., 2012). The social work discipline has used MLR to identify juvenile delinquency risk factors, the likelihood of psychiatric disorders among adults, and demographic characteristics of children who had experienced phsycial and emotional abuse (Afifi, Brownridge, Cox, & Sareen, 2006; Mandell, Walrath, Manteuffel, Sgro, & Pinto-Martin, 2005; J. P. Ryan, Hernandez, & Herz, 2007).

2.5 Summary

The link between games, data science, workload and the MAT-II is not immediately obvious. However, the MATB-II does have some characteristics that make it a game. There are specific rules a user has to follow, the task requires effort and emotional investment and there is a set of value laden outcomes associated with success and failure. While MATB-II does not have a game over screen informing the player of failure, there are still consequences for failure. Thinking of MATB-II as a game prompts a reexamination of how human performance data can be collected for use in human factors studies.

Modern video-games collect data about player actions without the player's awareness. Game designers use the collected data to asses player interactions or identify where players are exploiting the game's mechanics. Simple games have been used as a research tool to study interpersonal communication, individual motivation and player cooperation. The next reasonable step is to ask if workload or another measure of performance can be studied without the player'knowledge.

Games produce data at a rate and volume that the human factors discipline does not have the tools or training to process or represent. Further, many of the tools human factors practitioners are intrusive. Machine learning and data science techniques can be applied to a large volume of data to identify trends and patterns. The learning techniques can be supervised- where the outcome variable is explicitly labeled - or unsupervised - where the outcome variable is not explicitly labeled. These methods can be adopted by human factors as a way to process and derive meaning from a large volume of data.

3. METHODS

The following chapter details: a) how participants were recruited for the study , b) a description of the survey and main task participants were expected to complete and c) model testing and evaluation procedures.

3.1 Research Plan

The research was divided into seven stages. The first stage entailed the development of a Web-based version of MATB-II which was used to collect the data used for subsequent analysis. The second stage was a pilot study intended to identify the delineations in the data between high, medium, and low workload conditions. Once clear delineations in the data were defined, the MATB-II task was distributed via Prolific Academic to collect a high volume of continuous system state data.

Once the data was cleaned, a series of classification algorithms were constructed which use derived measures in the data to categorize participants into the categories of high, medium and low workload. These classification algorithms were trained on a subset of the data and tested on the remaining subset. Stage six used a similar process to determine if indirect (derived measures), or direct measures, can correctly classify the delay in responding to a malfunction. A sensitivity analysis was performed in the seventh and final stage to determine if the data's sampling rate, the size of the training set and or measures of participant's experience affected the ability to classify workload or delay.

3.1.1 Prolific Academic

Prolific Academic is an online subject recruitment platform that matches research requests with participants from around the world Prolific. It is similar to services such as Amazon's Mechanic Turk. Prolific Academic was selected for this study for two reasons. Researchers have the ability to target their study towards participants with specific demographic traits. Further, Prolific contains tools that help a researcher manage a large subject pool including subject payment tools.

3.2 Participant Recruitment

Participants were recruited through Prolific Academic. Demographic information about the participant's recent experience flying an aircraft or playing videogames in the past week and past month was collected using a survey. Participants would select from one of six response options when responding to these questions. These options are reprinted in Table 3.1 and Table 3.2. The selection of these levels was derived from Prolific Academic' s pre-screening questionnaire.

The participant would sign the online consent form using their Prolific Academic ID number. A participant's Prolific Academic ID number was a unique 24-character string that contained no personal identifying information. This ID number was used to link a participant's survey results to the data generated by the task website. Once the participant had completed the questionnaire they would move to the main task.

Participants were paid one dollar and fifty cents for their time.

Number of Participants

A total of 1,000 participants were recruited for participation in the study. Three hundred and sixty five individuals either left the task half way through or submitted incomplete responses. These participants were removed from the final study as their results were incomplete.

Table 3.1.

Levels for Gamers

Per Week	**Per Month**
1-2	1-5
3-4	5-10
5-6	11-15
7-8	16-20
9-10	21-25
10+	25+

Table 3.2.

Levels for Pilots

Per Week	**Per Month**
0-5	1-10
6-10	11-20
11-15	21-30
16-20	31-40
21-25	41-50
25+	50+

3.3 Description of Task

The NASA MATB-II was selected as the inspiration for this task for the following reasons. MATB-II was designed as a modular system. The modular design enables an experimenter to select specific task combinations for participants to perform by turning off task units. This design choice served as a precedent to justify presenting

participants with two tasks instead of five. Further, the MATB-II has been used to study cognitive workload, time pressures and the effects of fatigue on performance. Finally, the task described below is reproducible in the desktop version of MATB-II. Developing a custom task was considered but ultimately disregarded as a custom task could not be readily reproduced.

Participants were asked to interact with a custom-built website modeled after the NASA MATB-II. The interface only contained the tracking and resource management tasks. These were the two tasks that were the easiest to explain and had the fewest confounds. The audio task was not included as it could not be determined that poor performance could not be attributed to audio quality. A screen shot of this interface is shown in Figure 3.1.

Participants were presented with the following cover story:

> You are flying an older aircraft. It is prone to mechanical malfunctions and does not have autopilot. You will need to fly by hand and watch for and correct pump failures.

The task consisted of two components: a tracking task and a resource management task. The goal of the tracking task was to bring a floating reticle back to the center of the cross hairs. Participants would use their mouse to 'pull 'the reticle back to center. For example, if the reticle moved to the upper right of the screen, the participant would need to move the mouse towards the lower left.

The goal of the resource management task was to keep the fuel level of tanks A and B within 500 units of their respective starting value. Participants would use the number keys 1 to 8 to turn each pump on and off. One of these pumps would fail during the last 30 seconds of the experiment. This failure was indicated by the pump turning orange. A participant would need to 'restart'the pump by pressing the corresponding key. This website recorded the variables shown in Table 3.3 at a sampling rate of 60 Hz. The data was recorded as a JSON file labeled using the participant's Prolific Academic ID number.

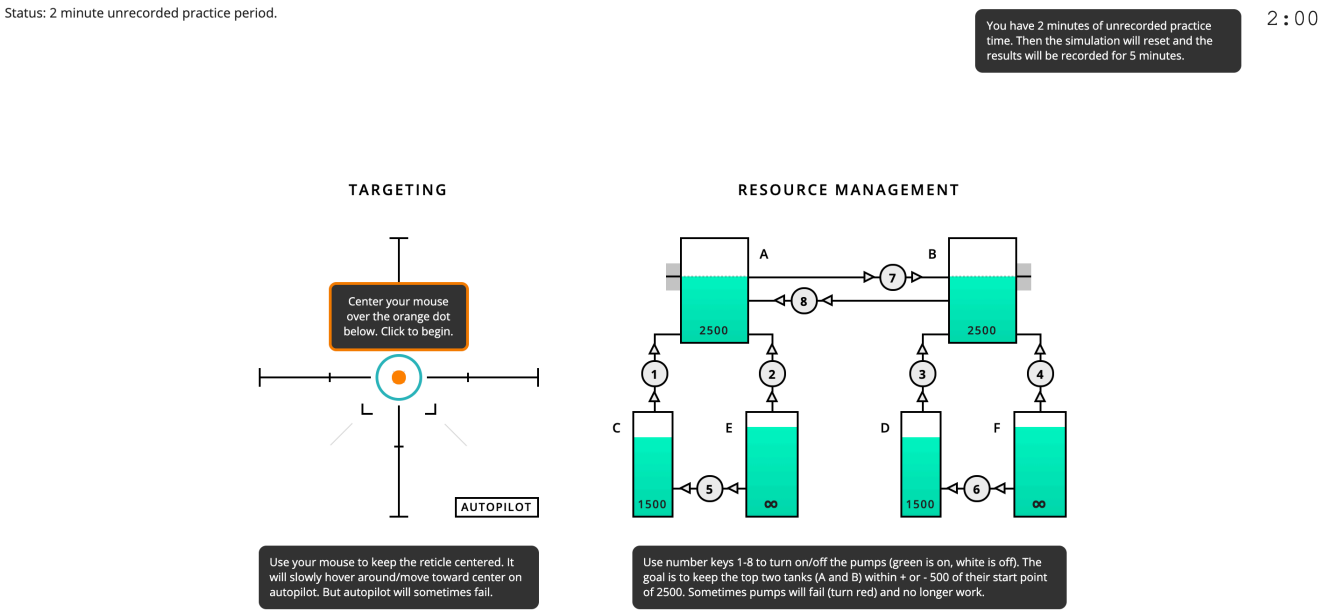


Fig. 3.1. Image of testing website

Table 3.3.
List recorded values

Variable Name	**Value**
Time	System Time
Mouse X / Mouse Y	Position of Mouse (pixels)
Reticle X / Reticle Y	Position of Reticle (pixels)
Tank Values (6 total)	Liters
Pump State (8 total)	On/Off/Failed
Difficulty	Easy/Medium/Hard

3.3.1 Interaction Modes

The two interaction modalities used in this experiment —mouse movement and key board— were selected for the following reasons. These are interaction modes

that the largest number of participants would have ready access to and the greatest experience with. While a joystick would fit the theme of the experiment, it was assumed that most participants would not have access to a joystick.

3.3.2 Difficulty Settings

The difficulty of the interaction task was determined using the following variables listed in 3.4. The mover force variables refer to the max force that will be present throwing the moving reticle off during auto-pilot failure the amount of time auto-pilot fails for before coming back (in radians) the minor turbulence always corrects back toward center upon change, this sets the max deviation allowed from pointing directly back to center. The pump flow rate is used to determine how quickly fuel flows through the pump using the following equation.

$$tickFlowRate = flowRate/(FRAMERATE * 60); \tag{3.1}$$

3.3.3 Pilot Study

A pilot study was conducted to assess how task difficulty was manipulated and how long the experiment should run. The eccentricity of the reticle and the rate of flow between the tanks were altered in each difficulty condition. In the easy condition the reticle could move up to fifty pixels away from the center of the cross hairs. However, the participant could move their mouse more than fifty pixels away to exert control on the reticle. While reticle position was directly dependent on difficulty, the mouse position would not be dependent.

The flow rate of the tanks was manipulated in all conditions. A direct measure of the rate of change in tank volume would directly correlate with difficulty. The participants were instructed to keep the main tanks within 500 units of their starting value for each tank. They were not told to keep both tanks balanced with each other.

Table 3.4.
Variables used to manipulate task difficulty

Variable	Easy	Medium	Hard
MOVER$_{A}UTOPILOT_{F}AILURE_{L}ENGTH$	3000	4000	5000
MOVER$_{M}AJOR_{M}AX_{F}ORCE$	0,4	0,9	1,4
MOVER$_{M}AJOR_{M}IN_{F}ORCE$	0,3	0,9	1
PUMP$_{1F}LOW$	800	800	900
PUMP$_{2F}LOW$	600	450	400
PUMP$_{3F}LOW$	800	700	650
PUMP$_{4F}LOW$	600	600	550
PUMP$_{5F}LOW$	600	350	220
PUMP$_{6F}LOW$	600	400	350
PUMP$_{7F}LOW$	400	300	200
PUMP$_{8F}LOW$	400	300	300
TANK$_{AC}ONSUMPTION$	800	800	800
TANK$_{BC}ONSUMPTION$	800	800	800

There is a reason to suspect that a participant would allow one main tank to become unbalanced at the expense of the other.

Finally, the pilot study determined that two minutes of training was sufficient to learning the controls without needing to consider learning effects. Further, the pilot study also showed that a pump failure in the last thirty seconds of the experiment was sufficiently unexpected.

The results of the pilot study indicated that participants in the Easy condition experience a lower subjective workload as measured by the NASA-TLX than participants in the Medium and Hard conditions. These results were confirmed by the

results of the talk aloud debriefing. The results of the pilot study are discussed in detail in the Data Exploration Chapter.

3.4 Data Preparation

A custom set of R scripts would read each JSON file separately into a data table. This script raw converted the recorded time into a fixed millisecond sequence to standardized time values across sampling rates. Each data table had 30,000 rows.

Once the time values had been reformatted, the sampling rate of the original 60 Hz data was artificially manipulated reformatted by resampling the original data at fixed intervals. This sampling reduced the total number of rows by a fixed amount. The processed data was then split into two subsets. Once subset contained all records before the pump failure and the other contained all rows after the pump failure.

The pre-pump failure subset was analyzed for the number of button presses recorded by each pump, the mean time between those presses and the standard deviation of the presses. The position the mouse and reticle were smoothed to reduce local noise in the pre-pump failure subset. A separate script would search the post failure subset for the failed pump and then calculate the time between the first failure and next state change. The processed data was stored in multiple SQL databases.

3.4.1 Sampling Rates

The 60 Hz data was resampled at the following intervals: 30 Hz, 12 Hz, 6 Hz, 3 Hz, 2 Hz, 1 Hz, Hz, 1/4 Hz and 1/10 Hz. The majority of commercial monitors refresh at or 60 Hz and 30 Hz. With the exception of the 12 Hz and 1/10 Hz sampling rates, each sampling rate is the Nyquist frequency of the proceeding value.

The sampling rates corresponded to the number of rows the resampling algorithm would skip in between samples. For example, at 30 Hz the resampling algorithm would take every second row while at 1 Hz the resampling algorithm took every 60th

row. This meant that at lower sampling rates, there were fewer rows of data that could be used to generate a derived measure.

12 Hz was used in lieu of 15 Hz because the sampling sequence would be: 7.5 Hz ,3.25 Hz, 1.125 Hz, etc. The reampling algorithm would need to take every 8th, 18.46th, and 53.3rd row respectively. No meaningful representation of these fractional rows could be found.

3.5 Model Selection

The modeling techniques used in this experiment were selected based on two criteria: a) a modeling method needed an extensive publication record and b) the modeling method needed extensive coding documentation. These criteria were implemented to ensure that the performance of the model could be assessed without needing to consider the validity of the method.

Four models were selected that satisfied these criteria: general linear models, neural networks, random forests and multinomial logistic regression. General linear models were only used regression problems and multinomial logistic regression was only used in categorization problems. Neural networks and random forests were used in both classification and regression problems.

The regression models were a general linear model, a neural net regression and a random forest regression. The categorization methods were a multinomial logistic regression, a neural network and a random forest. The general linear model and the multinomial logistic regression model were selected because they represent the purest translation of input variables to outcome. Their relative simplicity served as a means of validating the results of other regression and categorization.

3.6 Model Generation in R

The following general method was used in the regression and categorization experiments. The processed data was split into a training and test set using an 80/20 split.

The training data was divided into two groups of variables: the predictor variables and the outcome variable. A model would be given the training predictors and would map an association between the predictor and independent variables. A trained model would then use predictor variables drawn from the test set to predict the outcome t variable. This predicted outcome would be compared to the 'true'outcome.

The caret package in R automatically provides the following summary statistics for categorization models using the confusionmatrix() function: Accuracy, the 95% confidence interval, the No information rate and a p-value. The no information rate is the accuracy of a model that was guessing at category membership. The p-value was the result of a binomial exact test of the likelihood that the accuracy value was greater than the no information rate model (Kuhn, 2008).

Table 3.5.

Example Confusion Matrix

	Real			
Predicted		A	B	C
	A	66	15	18
	B	0	3	0
	C	1	0	10

The output of the confusion matrix function contains a table like the one shown in Table 3.5. The binomial test function used by *caret* uses the sum of the diagonal of this matrix as the number of successful trials. The sum of the entire total is the maximum number of trials. The default probability can be set by the analyst or will be set equal to the sum of the largest column over the total number of cases. This percentage is termed the no information rate model and represents the outcome of guessing at categorization.

In the case of Table 3.5, the equation would look like the equation below. The p-value of this equation is 0.0052.

$$p = binom.test(79, 113, (79/113)) \tag{3.2}$$

The Mean Average Error, r^2 and Root Mean Squared Error of regression models was obtained using the postresample function. This function takes the predicted and observed values as input would return the MAE, r^2 and RMSE of the model.

3.7 Model Inputs

Each model was constructed using five classes of derived variables: number of key presses, mean time between key presses, standard deviation of the key presses, mouse gain and tank differential (see Table 3.6). All variables were measured from the start of the session to four minutes and thirty seconds. This time marked the predetermined point where the pump was supposed to fail. A model could also be supplied with the participants weekly and monthly experience.

Table 3.6.

List of derived measures

Derived Measures	**Number**
Number of Keypresses	8
Mean Time Between Press	8
SD Time between Presses	8
Absolute Distance from origin	16
Tank Differential	21

Five distinct tests were conducted using the following variables to predict delay. These combinations were selected because they contained a direct or indirect measure of time and are listed below:

1. Weekly and Monthly Experience Only (2 variables)
2. Number of Key Presses Only (8 Variables)
3. Key Presses and Experience (10 variables)
4. Key presses, mean time between key presses and standard deviation (24 variables)
5. Test 4 with the addition of Experience (26 variables)

Seven distinct tests were conducted using the following variables to predict difficulty These tests are listed below:

1. Weekly and Monthly Experience Only (2 variables)
2. Mouse Position Only (16 Variables)
3. Mouse Position and Experience (18 variables)
4. Tank differential Only (21 variables)
5. Tank differential and experience (23 variables)
6. Tank differential and Mouse position (39 variables)
7. Test 6 with the addition of Experience (41 variables)

3.8 Model Evaluation

A categorization model was considered to accurately predict delay group or difficultly if the following criteria were met.

1. The average accuracy of the model was above chance

2. The lower bound of the 95% confidence interval was above chance
3. The performance of the model had to be significantly different than random guessing

The second and third criteria were implemented to ensure that the analyst was 95% certainty that the true prediction mean was above chance. The caret package in R automatically calculates the confidence interval and average accuracy using a binomial exact test. A model was deemed statistically significant if the p-value of the binomial exact test comparing the model to a random chance model was less than 0,05.

3.8.1 Random Guess Models

As previously mentioned, a model was considered to be statistically significant if it performed better than chance. The CARET package in R automatically generates a no information model. This no information model is a model that guesses at category membership given no additional information.

Chance was defined relative to two factors: a) the total number of categories and b) the distribution of category membership. There were three possible categories in the difficulty classification condition and these categories had an approximately equal number of class members. There were three possible categories in the delay classification condition but these categories had an unequal number of class members.

Chance for the difficulty classification condition was equal to one in three. This makes intuitive sense as if no additional information was provided and you blindly guessed at the outcome, one would be correct about one in three times. However, the distribution of the three classes in the delay categorization condition was uneven. One group contained approximately sixty percentage of all responses. If you did not know the distribution, one could say that a correct random guess would approach 1/3. However, when the distribution is accounted for the correct random guess rate would be 60%

3.9 Hypothesis & Research Questions

The research questions this dissertation addresses were motivated by a widening gulf between the data science and human factors disciplines in terms of understanding human performance. There has been an increase in the diversity of data sources that can be used by data scientists and human factors practitioners over the past twenty years. This increase in diversity is attributable to multiple factors such as increased adoption of mobile devices and other technological improvements.

The techniques and tools used by human factors professionals were not designed to process large volumes of continuous event data. Data science has the capabilities to process large volumes of data in near real time. However, data scientists do not necessarily possess a nuanced understanding of human performance. What a data scientist may label as poor human performance may not account for individual variations. Conversely, what human factors practitioners consider poor performance may be treated as normal human performance by a data scientist.

This led to the question *can the tools and techniques used by both disciplines be used to in conjunction to model human performance*? This broad question was operationalized in the following three questions.

RQ1 : Can a measure of human performance be inferred using continuous state data collected by a live system?

RQ 2: Can an individual's response to an unexpected event be predicting using continuous state data collected by a live system?

RQ 3: Can a measure of human performance be inferred using continuous state data **alone**?

These research questions informed the following hypothesis.

Null Hypothesis 1 : The first null hypothesis states that the number of correct classifications of task difficulty is equal to chance (1/3).

Hypothesis 1 : The first hypothesis is that the number of correct classifications of participant workload is greater than chance.

Null Hypothesis 2 : The second null hypothesis states that the number of correct classifications of participant delay groupings is equal to chance (1/N) where N is the number of clusters which best describes the data.

Hypothesis 2 : The second hypothesis states that there is a sampling rate and a training/test split below which a model would perform no better than chance.

Null Hypothesis 3 : The third null hypothesis states that the inclusion of demographic measures would not decrease the confidence interval of a prediction.

Hypothesis 3 : The third hypothesis states that the inclusion of demographic measures would decrease the confidence interval of a prediction.

4. GENERAL DATA EXPLORATION

The following sections detail the results of a pilot study conducted in late fall 2018 to determine of the three difficulty levels corresponded to distinct NASA—TLX workload scores. Further sections provide a general overview of the mean and median response times. Finally, the chapter closes with a discussion of the effect of experience on response time.

4.1 Results of Pilot Study

A short pilot study was conducted in October of 2018 to determine if the three task difficulty levels- Easy, Medium and Hard- corresponded to distinct NASA-TLX workload scores. 10 students from the School of Industrial Engineering responded to a post recruitment flier.

Table 4.1 shows the average TLX scores from the pilot study.

Table 4.1.

Results of Pilot Study

Condition	Participants	Average TLX Score	Average Score(no outliers)
Easy	4	51	43
Medium	3	58	58
Hard	3	64	64

The results of the pilot study indicate that the three difficulty levels are quantifiably different according to the NASA-TLX. On average, participants in the Easy condition experienced a lower workload during completion of the task than partici-

Table 4.2.
NASA-TLX Weighted Ratings

Easy	40	43	46	75
Medium	37	60	76	-
Hard	46	71	73	-

pants in the Medium and Hard Condition did (4.1). The greatest variability in TLX scores occurred in the Medium condition while the scores for the Easy and Hard condition appear to converge towards the low 40's and low 70's respectively.

The participants consistently assigned a higher ranking to the $\leq MentalDemand' and `Perforna$

4.2 General Analysis

The first phase of the analysis entailed a comprehensive exploration of the data. This analysis was necessary to address if response delay and task difficulty correlated and is there a statistically significant difference between the average delay in responding to an on-screen malfunction in the Easy/Medium/Hard task conditions. Finally, does participant experience have a statistically significant impact on response time.

4.2.1 Determining Correlation between Difficult and Delay

The first stage of the model building process entailed an examination of the mean and median response times of all participants by difficulty condition. The purpose of this stage was to determine if response delay was predictive of task difficulty and vice versa. The results of this analysis also provided insight into the skew of the data and potential statistical similarity between conditions.

Figure 4.1 is a graphical representation of the mean and median response times by condition. The figure does show a consistent skew in the direction of quicker

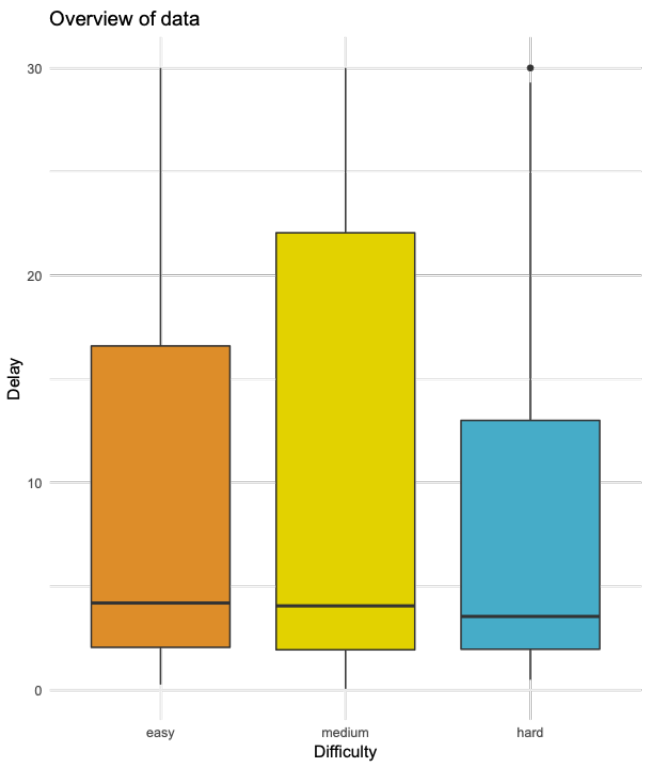


Fig. 4.1. Overview of data

Table 4.3.
Mean and Median Response time by condition

Condition	**Mean response time (secs)**	**Median response time**
Easy	10.33	4.7
Medium	11.07	4.28
Hard	9.64	3.56

responses across the three task difficulty conditions. Per the results shown in Table 4.3 and Table 4.4 there is no statistically significant difference between the mean response times for the Easy, Medium and Hard Conditions. This finding, along with

Table 4.4.
Results of T-test between conditions

Condition	P-value
Easy versus Medium	.507
Easy versus Medium	.5338
Medium versus Hard	.2076

the similarity in median response times, suggested that response time alone is not predictive of task difficulty nor is difficulty predictive of response time.

This result means that a model that predicted difficulty would not provide any information about response delay. Two models will need to be developed: one to predict difficulty and one to predict delay.

4.2.2 Differences in means by experience level

The next phase of the analysis focused on determining the effect of experience on response time once it had been established that the three conditions were not statistically different from each other. The question this phase sought to answer was do participants with similar levels of experience per week and per month have similar response times across conditions.

Based on Table 4.5 it is correct to say that on average higher levels of weekly and monthly experience correspond to a faster response to a pump failure. The fastest average response times occurred when the participant was at level 2 or 3 in weekly experience and 5 or 6 for monthly experience. This suggests that individuals with a higher level of monthly experience should have faster responses times.

However, when the response times are analyzed by difficulty levels the trend described above does not necessarily hold. Participants who were at the level one for both weekly and monthly experience experienced progressively slower reaction times

Table 4.5.
Mean response time by Week and Month

		Level by month						
		1	2	3	4	5	6	Average
Level by week	1	11,01	12,41	21,77	21,25	5,54	9,90	13,6
	2		11,43	10,34	7,73	10,61	7,12	9,4
	3		5,77	10,84	11,60	7,03	8,17	8,7
	4			17,94	14,37	10,74	9,57	13,2
	5		29,23	10,01	12,96	7,20	9,61	13,8
	6				22,99	6,95	10,56	13,5
	Average	11	15	14	15	8	9	

as difficulty increased. Beyond level two experience per month, additional experience per week does not correspond to a decrease in response time across all difficulty conditions. Higher levels of experience in the past month and week penalize individuals in the easy difficult condition but provide a slight benefit in higher difficulties.

These results indicate that measures of experience alone - at least as defined in this experiment- are insufficient predictors of difficulty and delay. Further, the benefit of experience does not scale in a predictable manner. This does not mean that a model built using delay as an input will necessarily be more or less accurate than a model built without these input features.

Table 4.6.
Mean response time by Week and Month for easy condition

		Level by month						
		1	2	3	4	5	6	
Level by week	1	5,75	23,59	20,25			30,00	19,9
	2		13,21	5,06	1,75	5,80	12,14	7,6
	3		9,68	13,93	10,21	12,23	10,85	11,4
	4			29,12	19,82	10,96	7,60	16,9
	5			1,87	10,78	10,36	10,57	8,4
	6					6,47	9,74	8,1
		6	15	14	11	9	13	

Table 4.7.
Mean response time by Week and Month for medium condition

		Level by month						Average
		1	2	3	4	5	6	
Level by week	1	10,77	11,22	16,62	18,51	4,84	8,67	11,8
	2		3,07	14,64	16,13	17,55	2,44	10,8
	3		5,44	6,75	16,49	4,16	7,47	8,1
	4			10,30	8,80	13,20	12,76	11,3
	5			15,67	10,01	2,46	7,18	8,8
	6				29,13	10,10	12,23	17,2
	Average	11	7	13	17	9	8	

Table 4.8.
Mean response time by Week and Month for hard condition

		Level by month						Average
		1	2	3	4	5	6	
Level by week	1	17,46	10,28	29,21	26,72	6,95	2,03	15,4
	2		12,27	10,45	9,50	12,34	1,93	9,3
	3		2,82	9,03	9,89	2,80	3,87	5,7
	4			3,22	14,67	7,93	8,34	8,5
	5		29,23	1,17	29,13	7,72	11,10	15,7
	6				10,73	5,81	9,52	8,7
	Average	17	14	11	17	7	6	

5. PREDICTING DIFFICULTY

The features used to predict task difficulty were divisible into three categories: participant experience, mouse gain and tank gain. Mouse gain was measured by counting the number of unique instances where the participant‘s mouse was a fixed distance away from the origin of the on-screen cross hairs. There were sixteen categorical mouse gain variables measured in ten-pixel increments. Tank gain was measured by examining the absolute between the values of the two main tanks on a scale of 0-500 broken down into fixed intervals.

Tank gain and mouse gain were selected as predictive features for the following reasons. Unlike the number of key presses-which has a positive correlation with difficulty- mouse gain and tank gain are not immediately correlated with difficulty. There is no *a priori* reason to suspect that individuals in the low difficulty would tolerate a greater difference in tank values. Further, while the eccentricity of the reticle was dependent on task difficulty, there was nothing limiting where the individual‘s mouse could be. Thus, even if the reticle could move up to fifty pixels, a participant could move their mouse up to three time as much.

A complete listing of all difficulty classification results can be found in graphical for in A and as tables in B.

5.1 Overview

Seven distinct tests were conducted using the following variables to predict difficulty. These tests varied the number of variables that could be used to develop a prediction. These tests are listed below:

1. Weekly and Monthly Experience Only (2 variables)

2. Mouse Position Only (16 Variables)

3. Mouse Position and Experience (18 variables)

4. Tank differential Only (21 variables)

5. Tank differential and experience (23 variables)

6. . Tank differential and Mouse position (39 variables)

7. Test 6 with the addition of Experience (41 variables)

Table 5.1 shows the average accuracy of the three classes of difficulty prediction models trained using 90% of the available data sampled at 60Hz. For additional results, see B. The cells in grey are statistically insignificant results.

Table 5.1.
Average accuracy of workload classifiers

Model	**1**	**2**	**3**	**4**	**5**	**6**	**7**
Multinomial	30%	93%	89%	51%	51%	88%	91%
Neural Network	33%	89%	88%	54%	47%	81%	91%
Random Forest	30%	95%	89%	51%	54%	91%	95%

The following conclusions can be drawn from this table. First, models trained using experience alone failed to be statistically significant. This outcome was expected as task difficulty was randomly assigned. Thus, any correlation between experience and difficulty is spurious. The second conclusion is that the inclusion of measures of experience has a mixed effect on the accuracy of a model. There are some instances where the model will gain a few points of predictive accuracy (columns 6 and 7 of Table 5.1) and instances where the inclusion of experience results in a loss in predictive accuracy. Finally, models that include information about mouse differential performed

Table 5.2.
Average Performance of statistically significant models

Model	Lower bound	Average	Upper Bound
Multinomial	68%	74%	79%
Neural Network	70%	75%	80%
Random Forest	71%	77%	81%

35% better than models that did not have tank differential Table 5.2shows the average performance of all statistically significant models. The results suggest that random forest models are the most accurate. However, no one modeling method significantly out performs any other.

5.1.1 Experience Only

Table5.3shows the results of the multinomial logistic regression, net network and random forest models built using participant experience as input and trained using 90% of the data as a training set. All three models failed to satisfy all three of the criteria listed previously. No model had an average accuracy above chance or a confidence interval that was above chance.

5.1.2 Mouse Data Only

Table5.4shows the results of the multinomial logistic regression, net network and random forest models built using participant experience as input and trained using 90% of the data as a training set. All three models failed to satisfy all three of the criteria listed previously. No model had an average accuracy above chance or a confidence interval that was above chance.

Table 5.3.
Mean Model Accuracy using Experience Only

Model	**Lower bound**	**Accuracy**	**Upper bound**	**Model>Chance**
Multinomial	18%	30%	43%	0.96
Neural Network	21%	33%	47%	0.88
Random Forest	18%	30%	43%	0.96

Table 5.4.
Model accuracy using mouse position data

Model	**Lower bound**	**Accuracy**	**Upper bound**	**Model>Chance**
Multinomial	83%	93%	98%	0
Neural Network	78%	89%	96%	0
Random Forest	85%	95%	99%	0

5.1.3 Mouse and Experience Data

Table 5.5 shows the results of the multinomial logistic regression, net network and random forest models built using mouse data as input and trained using 90% of the data as a train set. None of the models shown performed worse than chance and all three satisfied the test criteria. However, these models do perform slightly worse than models that do not include experience.

Table 5.5.
Model accuracy using mouse data and experience

Model	**Lower bound**	**Accuracy**	**Upper bound**	**Model>Chance**
Multinomial	78%	89%	96%	0
Neural Network	76%	88%	95%	0
Random Forest	78%	89%	96%	0

Table 5.6.
Model accuracy using tank differential data

Model	**Lower bound**	**Accuracy**	**Upper bound**	**Model>Chance**
Multinomial	37%	51%	64%	0.07
Neural Network	41%	54%	68%	0.02
Random Forest	37%	51%	64%	0.07

5.1.4 Tank Data Only

Table 5.6 shows the results of the multinomial logistic regression, net network and random forest models built using tank differential data as input and trained using 90% of the data as a train set. The multinomial logistic regression and random forest models were both statistically indistinguishable from a random guess model. This is because the lower bound did not exceed chance.

Table 5.7.
Model accuracy using tank differential and experience

Model	**Lower bound**	**Accuracy**	**Upper bound**	**Model>Chance**
Multinomial	45%	51%	57%	0
Neural Network	41%	47%	53%	0
Random Forest	48%	54%	60%	0

5.1.5 Tank and Experience

Table 5.7 shows the results of the multinomial logistic regression, net network and random forest models built using tank differential data as input and trained using 90% of the data as a train set. The multinomial logistic regression and random forest models were both statistically indistinguishable from a random guess model. This is because the lower bound did not exceed chance.

5.1.6 Tank, Mouse and Experience Combinations

Table 5.8.
Model accuracy using mouse position and tank differential

Model	**Lower bound**	**Accuracy**	**Upper bound**	**Model>Chance**
Multinomial	76%	88%	95%	0
Neural Network	68%	81%	90%	0
Random Forest	81%	91%	97%	0

Table 5.9.
Model Accuracy using mouse position, tank differential and experience

Model	**Lower bound**	**Accuracy**	**Upper bound**	**Model>Chance**
Multinomial	81%	91%	97%	0
Neural Network	81%	91%	97%	0
Random Forest	85%	95%	99%	0

Table 5.8 shows the results of models built using mouse position, tank differential as input. Table 5.9 shows the results of models built using mouse position, tank differential and experience as inputs. While both groups of models are statistically significant, the models that include experience are 5% more accurate on average.

5.2 Interpretations of Results

The results of these analyses are evidence that task difficulty can be predicted at a rate above chance. These results indicate that measures of participant experience are not necessary to create a statistically significant classification model. Excluding experience only models, the least accurate but statistically significant models are one built using tank differential and experience. Models built using tank differential alone were not guaranteed to be statistically significant. The most accurate models were those built using information about the position of the mouse.

Evidence shows that the inclusion of tank differential to models built using mouse position data are slightly less accurate on average than mouse data alone. However, the mean predictive accuracy of these two classes of models are not significantly statistically different from each other. Further, models that include experience as a

predictor variable are not significantly statistically different from models that do not have experience as a predictor variable.

6. DELAY PREDICTION MODELS

Five distinct tests were conducted using the following combinations of variables to predict delay in seconds. These tests varied the number of variables that could be used to develop a prediction. These tests are listed below:

1. Weekly and Monthly Experience Only (2 variables)
2. Number of Key Presses Only (8 Variables)
3. Key Presses and Experience (10 variables)
4. Key presses, mean time between key presses and standard deviation (24 variables)
5. Test 4 with the addition of Experience (26 variables)

Three regression methods were used to model the relationship between input and output variables. These methods were: general linear modeling, neural networks and random forests. In total, fifteen separate models were created for a given sampling rate. The results shown herein were generated by models using 90% of the available data sampled at 60Hz. A complete listing of all delay classification results can be found in graphical for in A and as tables in B.

Table 6.1 shows the RMSE values of fifteen models given the five possible combinations of input variables. These values take on greater significance when compared to the mean response delay. The mean response delay was 11.2 seconds. These models are off by over a third of the total range.

Table 6.1.
RMSE values of regression models at 60Hz

Model	Exp Only	Key Press	Key Press Exp	Key Press, Mean and SD	Key press, mean, SD, and Exp
Linear	11.8	12.4	12.5	14	12.6
Neural Network	12	10.5	11.6	10.5	11.4
Random Forest	12.9	10.3	11.5	10.7	11

6.1 Model Performance

To say that the regression models performed poorly would be an understatement. The mean average error of the models does reach a minimum value of 4 seconds. This occurred when the sampling frequency was equal to a 1/10th of a hertz. There are reasons to suspect that models built using data at this sampling rate are prediction a distorted view of real events. Excluding sampling rates below 1/2Hz , the models will incorrectly predict delay by at least 8 seconds.

Table 6.2.
Regression model performance using experience only

Model	**RMSE**	r^2	**MAE**
General Linear Model	11.863	0.000	10.662
Neural Network	12.023	0.025	10.819
Random Forest	12.293	0.001	11.065

Table 6.3.
Regression model performance using keypress data

Model	**RMSE**	r^2	**MAE**
General Linear Model	12.460	0.016	11.068
Neural Network	10.496	0.233	8.675
Random Forest	10.375	0.249	8.371

Table 6.4.
Regression model performance using keypress and experience

Model	**RMSE**	r^2	**MAE**
General Linear Model	1.,537	0.025	11.178
Neural Network	11.655	0.094	9.535
Random Forest	11.581	0.110	9.474

Table 6.5.
Regression model performance using keypress and interaction times

Model	**RMSE**	r^2	**MAE**
General Linear Model	14.544	0.046	10.717
Neural Network	10.537	0.202	8.325
Random Forest	10.772	0.170	8.566

When the r^2 value of the models is taken into consideration, it becomes apparent that the models are essentially predicting the mean response time of the data set. This is evidence by the low r^2 values. There are three explanations for these outcomes.

Table 6.6.
Regression model performance using keypress, interaction times and experience

Model	**RMSE**	r^2	**MAE**
General Linear Model	12.265	0.020	10.915
Neural Network	11.450	0.122	9.168
Random Forest	11.014	0.173	9.298

The first explanation is that the underlying distribution of the data is skewed to the left. This evidence for this assertion comes from

Table 6.6 where it is apparent that the mean response time is larger than the median response time. The second explanation is that the selected variables are not simply not predictive of delay. The third and final explanation is that the selected variables could be valid, but the relationship is not linear.

6.2 Predicting Delay Group

No regression model could be found that accurately predicted delay. The decision was made to convert the regression problem into a classification problem by clustering the delay responses based on similarity. The logic was that rather than predicting a participant delay in second, it would be possible to predict which delay cluster they would belong in based on a set of variables. A K-Nearest Neighbors clustering algorithm was applied to the delay data. The resulting clusters are shown below in Figure 6.1. The results show that delay can be divided into three clusters centered at: 2.54, 10.92 and 28.26 seconds. What is not shown in this image is the number of individuals in each cluster. 60% of the participants belong to cluster one while the reaming 40% were split between the remaining two clusters. Given this class imbalance, a model that was guessing cluster membership would be correct 60% of

Fig. 6.1. Results of Cluster Analysis.

the time. In the context of the success criteria outlined previously, a model would be considered significant if the mean accuracy was better than chance and the confidence interval was above chance.

6.2.1 Result

Table 6.7 shows the mean average accuracy of delay group prediction models training using data sampled at 60 Hz. Only neural networks build using key press variables were statistically significant at 72% with an upper and lower bound on the prediction at 58.6% and 83%. These results prove that a delay can be classified at a rate above chance. However, the preponderance of models that failed to be statistically significant are evidence that the selected combination of predictor variables is insufficient.

There are two possible solutions to the class imbalance problem under consideration. The first possible solution is to collect more data to assess if the imbalance

Table 6.7.
Mean Accuracy of Delay Prediction

Model	**Exp. Only**	**Key Press**	**Key Press Exp.**	**Key Press, Mean, and SD**	**Key press, Mean, SD, and Exp.**
Multinomial	60%	60%	56%	65%	65%
Neural Network	60%	72%	65%	67%	67%
Random Forest	54%	68%	64%	67%	70%

persists. The second option is to use cost sensitive learning techniques to penalize the more common class. This method has been previously shown to correct class imbalance problems (Elkan, 2001; Zhang & Oles, 2001) .

7. EFFECT OF SAMPLING RATE AND TRAINING SET SIZE

The following section is divided into three parts. Section 7.1 provides a short overview of summary of the mean, median and number of unique values generated at a given sampling rate. Section 7.2 details the effects of experience, sampling rate and training set size on the categorization accuracy of difficulty classification models. Section 7.3 details the effects of experience, sampling rate and training set size on the categorization accuracy of delay classification models.

A complete listing of the results can be found in graphical for in A and as tables in B.

7.1 Means, Medians and Unique values

Table 7.1 shows the mean and median values of the analyzed data at different sampling rates along with the number of unique delay times in each of the data sets. As the sampling rate decreases, the median value of the dataset will increase while the number of unique entries decreases. This is a consequence of how the algorithms calculated delay.

The algorithm sequentially searched for a failure flag in each of the eight columns of pump data. Once the failure flag had been found, the algorithm would count the number of rows between the failure state and the correction of that state. The sampling rate was altered by removing rows at a particular interval. If the change in the size of the sampling window is small enough, the difference in recorded reaction times is not immediately apparent.

Table 7.2 shows a graphical representation of how sampling rate affected the data. Each row in the table is a fixed time step. At 60Hz, all 12 rows will be included in the

Table 7.1.
Table of Means, Medians and Unique Values

Sampling Frequency (Hz)	**Mean**	**Median**	**Min**	**Max**	**Unique**
60	11.36	4.20	0.10	29.98	279
30	11.36	4.18	0.10	29.96	225
12	11.32	4.16	0.08	29.91	148
6	11.29	4.16	0.33	29.83	99
3	11.18	4.00	0.33	29.66	68
2	11.09	4.00	0.50	29.50	49
1	11.18	5.00	1.00	29.00	29
1/2	12.57	10.17	2.00	28.00	15
1/4	12.199	7.86	4.00	24.00	7
1/10	10.627	6.52	6,52	20.00	3

processed data. The 30Hz sample will contain half the number of rows as the 60Hz data. The 60Hz data accurately represents when the pump switched from failed to active (row 8). Data sampled at 30Hz shows the same failure being corrected at row 9 while data sampled at 12Hz captures the change in state at row 11.

The difference between the rows is 0.0166 seconds. This time is small enough that detecting the same event one row later does not significantly alter the outcome. When the sampling rate is slow enough, the time difference between sampled rows can exceed two seconds. The system is essentially blind to state changes that occur faster than the sampling window and will round a participant‘s response up to the nearest whole unit. This is why the minimum possible recorded value increases as the sampling rate decreases. Delay values that fall between sampled rows are coerced up to the nearest sampled value. For example, if the sampling rate captured every tenth

Table 7.2.
Graphical Explanation of the effect of sampling rate

Row Number	State	60Hz	30Hz	12 Hz	Time (secs)
1	Fail	Yes	Yes	Yes	250.0000
2	Fail	Yes	No	No	250.0167
3	Fail	Yes	Yes	No	250.0333
4	Fail	Yes	No	No	250.0500
5	Fail	Yes	Yes	No	250.0667
6	Fail	Yes	No	Yes	250.0833
7	Fail	Yes	Yes	No	250.1000
8	Active	Yes	No	No	250.1167
9	Active	Yes	Yes	No	250.1333
10	Active	Yes	No	No	250.1500
11	Active	Yes	Yes	Yes	250.1667
12	Active	Yes	No	No	250.1833

row but the participant responded at the 36th row, then their response time would be set equal to the time value in the 40th rows.

7.2 Difficulty Classification

The following sections contain two types of heat maps. One Heat-map shows the 95% confidence interval for a given modeling method built using tank variables and tank and experience variables respectively. These two families of models had the lowest predictive accuracy of all models excluding experience only models. The bluer values represent smaller ranges. The second Heat-maps show the classification accuracy for given modeling method built using tank variables and tank and experience variables respectively.

A complete listing of difficulty classification results can be found in Appendices A and B.

7.2.1 Multinomial Logistic Regression

Effect of Experience

Figure 7.1 and Figure 7.2 are Heat-maps showing the 95% confidence interval for multinomial logistic regression models built using tank variables and tank and experience variables respectively. These two families of models had the lowest predictive accuracy of all models excluding experience only models. The bluer values represent smaller ranges.

Three conclusions can be drawn from these figures. The first conclusion is that increasing the ratio of data in the training set corresponds with a larger range between the upper and lower bound of the confidence interval for models built with and without experience as an input. The second conclusion is that altering the sampling rate does not have an effect on the range of the confidence interval for models built with and without experience. However, the inclusion of experience reduces the range of the confidence interval at training and testing splits above 75%.

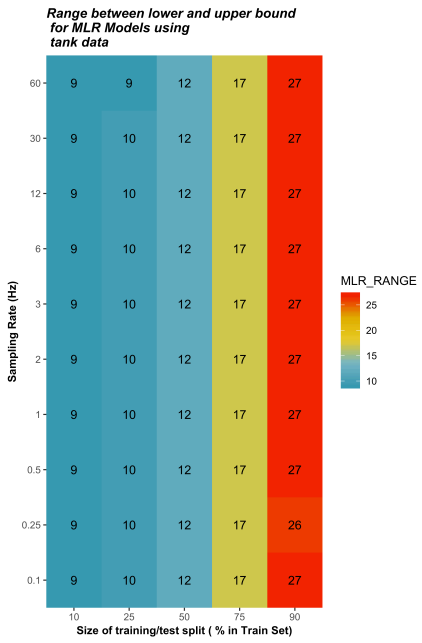


Fig. 7.1. Heat-map of Confidence Intervals for multinomial logistic regression models without experience

Effect of Sampling Rate and Training Set Size

Figure 7.3 and Figure 7.4 are Heat-maps showing the classification accuracy for multinomial logistic regression models built using tank variables and tank and experience variables respectively. These two families of models had the lowest predictive accuracy of all models excluding experience only models. The bluer values represent higher accuracies. The numbers are the p value for the model.

Three conclusions can be drawn from these figures. The first conclusion is that reducing the sampling rate of the data neither significantly increases nor decreases classification accuracy for a given training ratio. The second conclusion is that increasing the training ratio corresponds with a marginal gain in accuracy for a given sampling rate. Finally, there is an effect of training ratio on the statistical significance of a model when experience is included. Models built without experience are guaranteed to be statistically significant at training splits below 50% while models

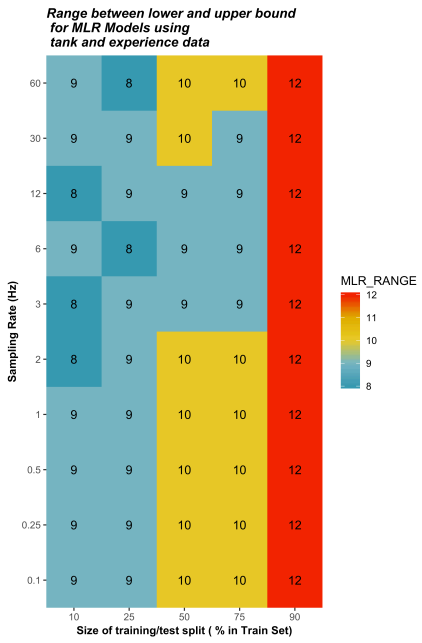


Fig. 7.2. Heat-map of Confidence Intervals for multinomial logistic regression Models with experience

with experience are guaranteed to be statistically significant at training splits above 50%.

Interpretation

The evidence shown above demonstrates that the inclusion of experience has the following effects on classification accuracy of multinomial logistic regression models. First, experience will reduce the range of the confidence interval in models trained using 50% or more of the available data. The inclusion of experience into a model implies that a larger training set is required to guarantee a statistically significant prediction. Finally, the inclusion experience does not increase the accuracy of the model.

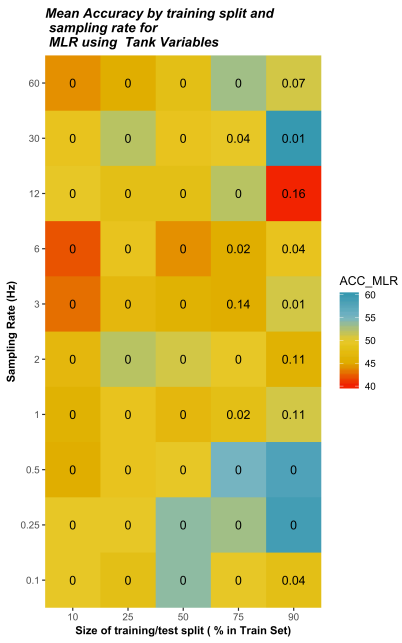


Fig. 7.3. Heat-map of Prediction Accuracy for multinomial logistic regression Models without experience

7.2.2 Neural Networks

Effect of Experience

Figure 7.5 and Figure7.6 are Heat-maps showing the 95% confidence interval for neural net models built using mouse and tank variables and mouse, tank and experience variables respectively. The bluer values represent smaller ranges. Three conclusions can be drawn from these figures. The first conclusion is that increasing the ratio of data in the training set corresponds with a larger range between the upper and lower bound of the confidence interval for models built with and without experience as an input. The second conclusion is that altering the sampling rate does not have an effect on the range of the confidence interval for models built with and without experience. However, the inclusion of experience increases the range of the confidence interval at training and testing splits above 75%.

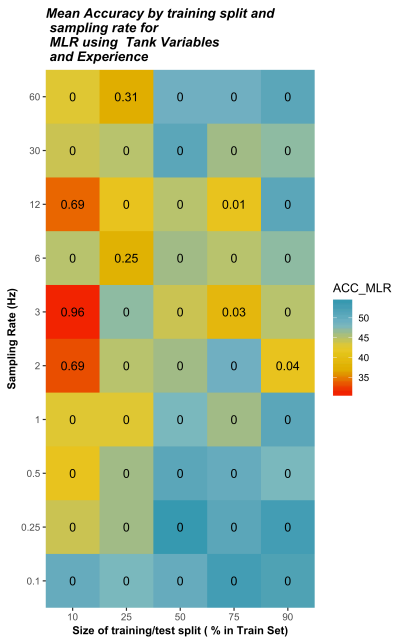


Fig. 7.4. Heat-map of Prediction Accuracy for multinomial logistic regression models with experience

Effect of Sampling Rate and Training Set Size

Figure 7.7 and Figure 7.8 are Heat-maps showing the classification accuracy for neural net models built using mouse and tank variables and mouse, tank and experience variables respectively. These two families of models had the lowest predictive accuracy of all models excluding experience only models. The bluer values represent higher accuracy predictions. The numbers are the p value for the model.

Three conclusions can be drawn from these figures. The first conclusion is that reducing the sampling rate of the data does not consistently decrease the accuracy of a model. The second conclusion is that increasing the training ratio corresponds with a marginal gain in accuracy for a given sampling rate. Finally, there is an effect of training ratio on the accuracy of a model when experience is included. The inclusion of experience in models built using less than 10% of the data perform marginally above chance. The inclusion of experience will decrease the accuracy of a model.

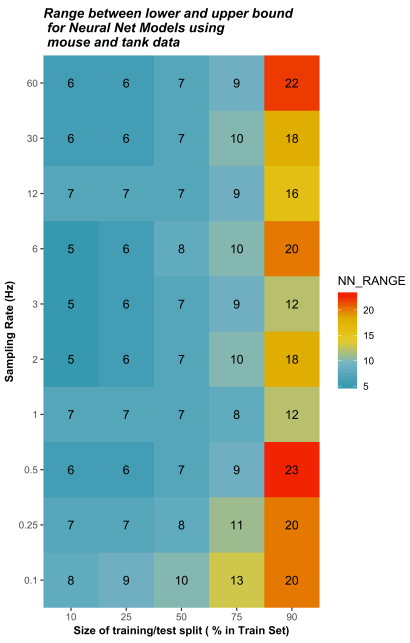


Fig. 7.5. Heat-map of Confidence Intervals for Neural Net Models without experience

Interpretation

The inclusion of experience does not change the confidence interval of models built using less than 50% of the data. However, the inclusion of experience can make models statistically insignificant when training split and sampling rate are controlled. For larger training splits experience will only marginally reduce the confidence interval of a prediction.

7.2.3 Random Forests

Effect of Experience

Figure 7.9 and Figure7.10 are heat-maps showing the 95% confidence interval for neural net models built using mouse and tank variables and mouse, tank and experience variables respectively. The bluer values represent smaller ranges. Three

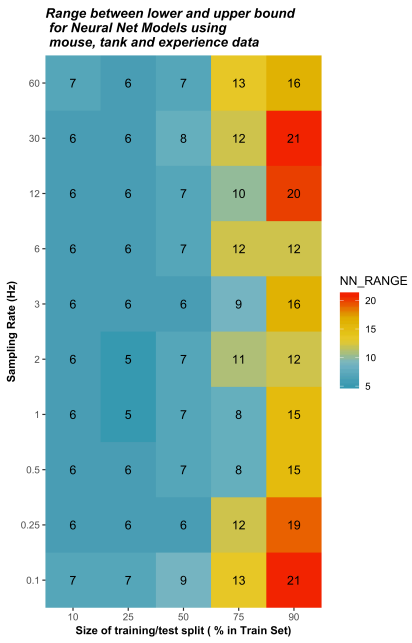


Fig. 7.6. Heat-map of Confidence Intervals for Neural Net Models with experience

conclusions can be drawn from these figures. The first conclusion is that increasing the ratio of data in the training set corresponds with a larger range between the upper and lower bound of the confidence interval for models built with and without experience as an input. The second conclusion is that altering the sampling rate does not have an effect on the range of the confidence interval for models built with and without experience. However, the inclusion of experience reduces the range of the confidence interval at training and testing splits above 75%.

Effect of Sampling Rate and Training Set Size

Figure 7.11 and Figure 7.12 are heat-maps showing the classification accuracy for random forest models built using mouse and tank variables and mouse, tank and experience variables respectively. These two families of models had the lowest predictive accuracy of all models excluding experience only models. The bluer values

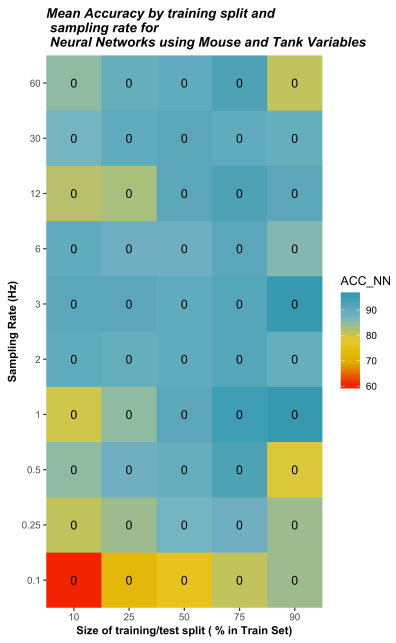


Fig. 7.7. Heat-map of Prediction Accuracy for Neural Net Models without experience

represent relatively more accurate predictions. The numbers are the p value for the model.

Three conclusions can be drawn from these figures. The first conclusion is that reducing the sampling rate of the data does not consistently decrease the accuracy of a model for training splits under 75%. The second conclusion is that increasing the training ratio does correspond with a marginal gain in accuracy for a given sampling rate for models built with experience. Finally, there is an effect of experience on the accuracy across all sampling conditions. The models are less accurate but will always be statistically significant.

Interpretation

The inclusion of experience will reduce the range of the confidence interval for models built using a training split greater than 75%. However, the inclusion of expe-

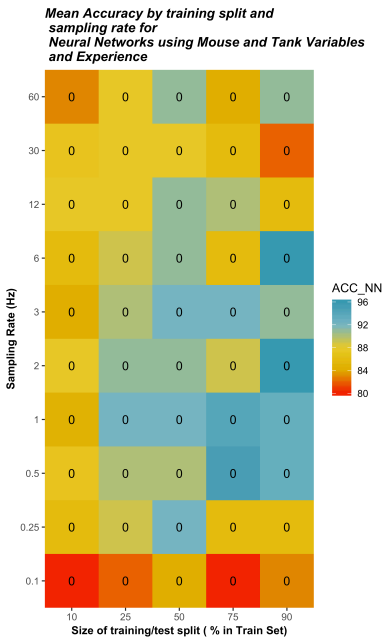


Fig. 7.8. Heat-map of Prediction Accuracy for multinomial logistic regression models with experience

rience reduces the accuracy of the models but will ensure the prediction is statistically significant. The most accurate models do occur at higher training splits but if the sampling rate and training split are reduced a model of equivalent accuracy will be generated.

7.2.4 Discussion

The models shown in the previous sections were difficulty classification models built using tank variables and tank and experience variables. The models produced using different combinations of predictor variables were consistently performing above 80% classification accuracy. However, the data yields three important findings. The first is that including experience will reduce the confidence interval of a prediction and decrease the accuracy across multiple training splits and sampling rates.

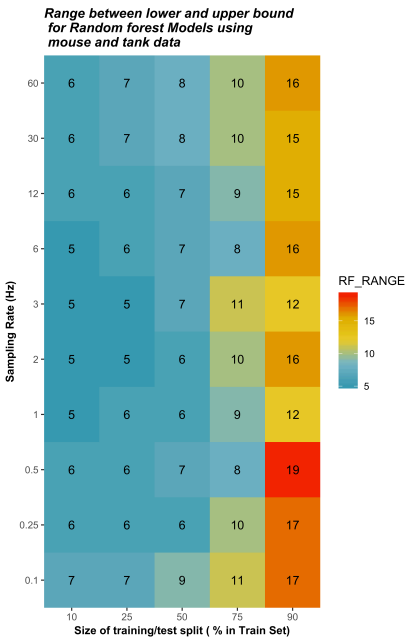


Fig. 7.9. Heat-map of Confidence Intervals for Random Forest Models without experience

The second finding is that altering the size of the training split does have an effect on the statistical significance of a model when controlling for experience. Models built using experience require a larger training split to guarantee statistical significance. Finally, altering sampling rate had a marginal effect on predictive accuracy. There is no sampling rate where all models fail to be significant.

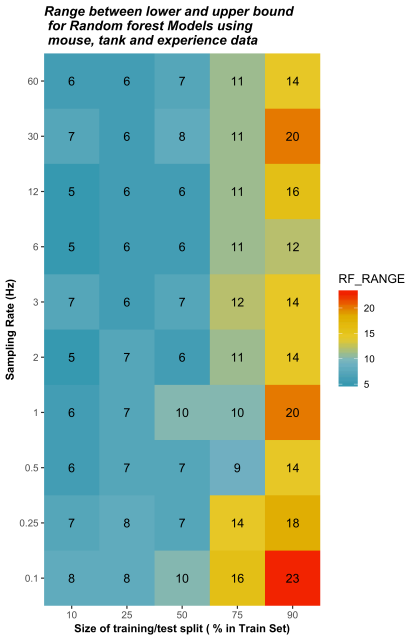


Fig. 7.10. Heat-map of Confidence Intervals for Random Forest Models with experience

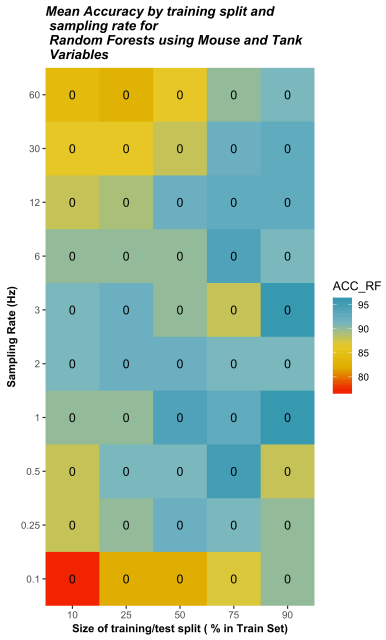


Fig. 7.11. heat-map of Prediction Accuracy for Random Forest Models without experience

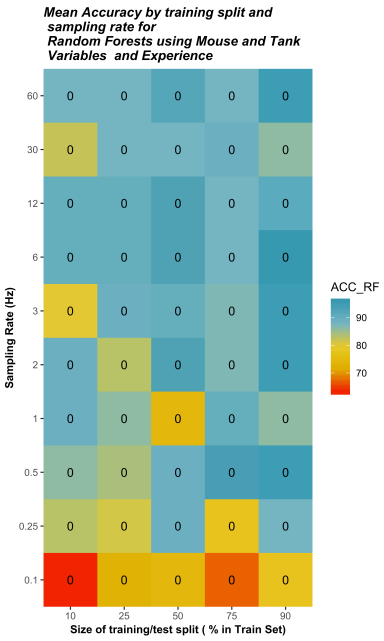


Fig. 7.12. heat-map of Prediction Accuracy for Random Forest Models with experience

7.3 Delay Classification

The following sections contain two types of heat maps. One heat-map shows the 95% confidence interval for a given modeling method built using the number of keypress and the number of keypress and experience variables respectively. These two families of models had the best predictive accuracy of all models excluding experience only models. These two families of models had the lowest predictive accuracy of all models excluding experience only models. The bluer values represent smaller ranges. The second heat-maps show the classification accuracy for a given modeling method built using keypress and the number of keypress and experience variables respectively.

A complete listing of delay classification results can be found in classification results Appendices A and B.

7.3.1 General Linear Regression

Effect of Experience

Figure 7.13 and Figure 7.14 heat-maps showing the 95% confidence interval for general linear models built using the number of keypress and the number of keypress and experience variables respectively. These two families of models had the best predictive accuracy of all models excluding experience only models. The bluer values represent smaller ranges.

Two conclusions can be drawn from these figures. The first conclusion is that the inclusion of experience does not alter the size of confidence intervals. The second conclusion is that models produced using a 90% training split and data sampled at 1/4 Hz have the greatest range.

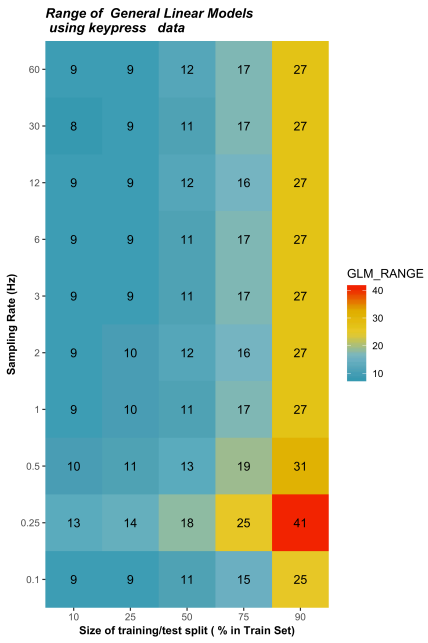


Fig. 7.13. heat-map of Confidence Intervals for multinomial logistic regression models without experience

Effect of Sampling Rate and Training Set Size

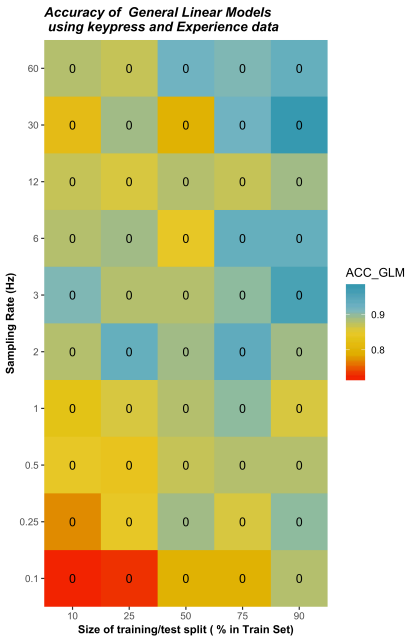


Fig. 7.16. heat-map of Prediction Accuracy for general linear models with experience

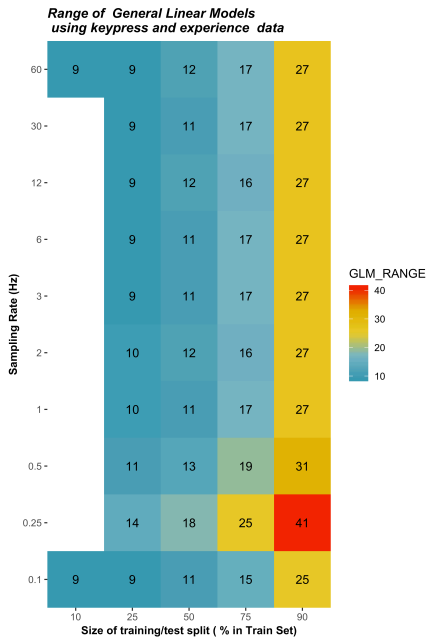


Fig. 7.14. heat-map of Confidence Intervals for multinomial logistic regression Models with experience

Figure 7.15 is a heat-map showing the classification accuracy for general linear models using the number of keypress. Figure 7.16 is a heat-map showing the classification accuracy for general linear models using the number of keypress and experience variables . These two families of models had the greatest predictive accuracy of all models excluding experience only models. The bluer values represent relatively more accurate predictions. The numbers are the p value for the model.

There were only eight models that accurately predicted delay group at a rate above chance in Figure7.15 and only one model that accurately predicted delay group at a rate above chance in Figure 7.16. This is strong evidence that experience penalizes model performance. The second conclusion is that for models that do not have experience and are built using data sampled faster than 1/Hz, the only statistically significant models occur at 30Hz, 6Hz, 3Hz, 2Hz and 1 Hz.

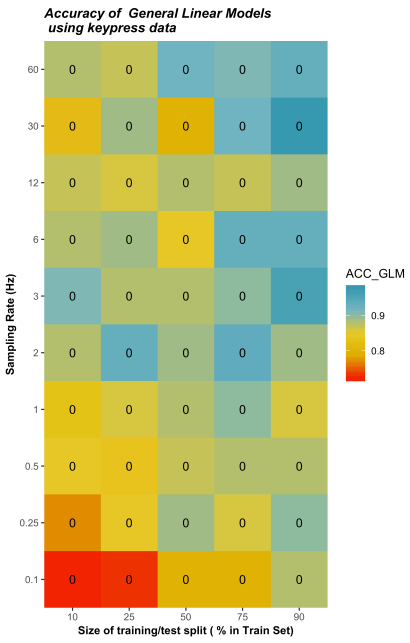


Fig. 7.15. heat-map of Prediction Accuracy for general linear models Models without experience

7.3.2 Neural Networks

Effect of Experience

Figure **??** and Figure**??** are heat-maps 95% confidence interval for neural net models built using the number of keypress and the number of keypress and experience variables respectively. These two families of models had the best predictive accuracy of all models excluding experience only models. The bluer values represent smaller ranges. Two conclusions can be drawn from these figures. The first conclusion is that the inclusion of experience does not alter the size of confidence intervals. The second conclusion is that models produced using a 90% training split and data sampled at 1/4 Hz have the greatest range.

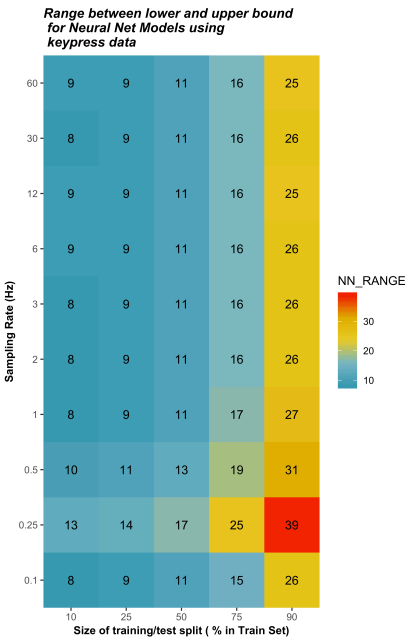


Fig. 7.17. heat-map of Confidence Intervals for Neural Net Models without experience

Effect of Sampling Rate and Training Set Size

Figure 7.19 and Figure 7.20 are heat-maps showing the classification accuracy for neural net models built using mouse and tank variables and mouse, tank and experience variables respectively. These two families of models had the lowest predictive accuracy of all models excluding experience only models. The bluer values represent relatively more accurate predictions. The numbers are the p value for the model.

Three conclusions can be drawn from these figures. The first conclusion is that a little more than half of all combinations of training split and sampling rate will result in a statistically significant model when experience is not included. There are 18 combinations of training split and sampling rate will result in a statistically significant model when experience is not included. This is evidence that experience does penalize a models significance. There is a range of sampling frequencies at the 25% training split that consistently produce statistically significant models when

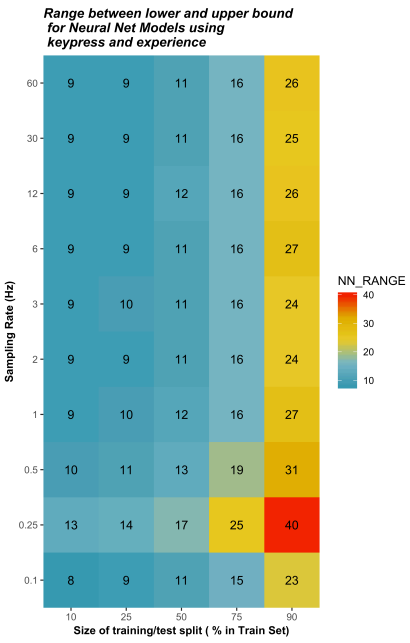


Fig. 7.18. heat-map of Confidence Intervals for Neural Net Models with experience

experience is and is not included. Experience does have an impact on the required sampling frequency at the 90% training split. Models that do not have experience will be statistically significant at 60Hz and 12Hz while models with experience will be statistically significant at 3Hz and 2Hz.

Interpretation

The evidence demonstrates that neural network models can predict delay group under specific training split conditions. Models built using 25 to 50% of the available data are more likely to be statistically significant at variety of sampling frequencies. There is an interesting trend that occurs models built using data sampled under of a hertz that is discussed later in the paper.

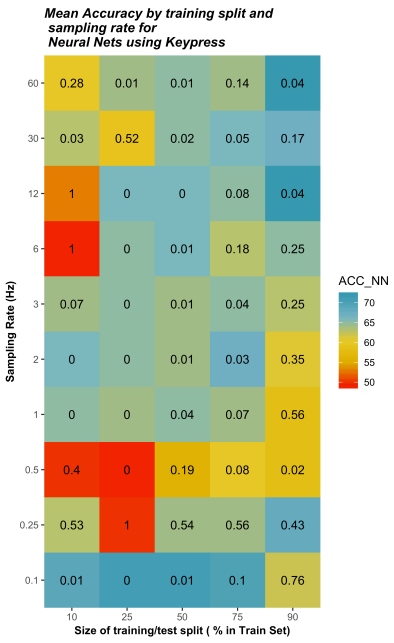


Fig. 7.19. heat-map of Prediction Accuracy for Neural Net Models without experience

7.3.3 Random Forests

Effect of Experience

Figure 7.21 and Figure7.22 are heat-maps showing the 95% confidence interval for neural net models built using the number of keypress and the number of keypress and experience variables respectively. These two families of models had the best predictive accuracy of all models excluding experience only models. The bluer values represent smaller ranges. Two conclusions can be drawn from these figures. The first conclusion is that the inclusion of experience does not alter the size of confidence intervals. The second conclusion is that models produced using a 90% training split and data sampled at 1/4 Hz have the greatest range.

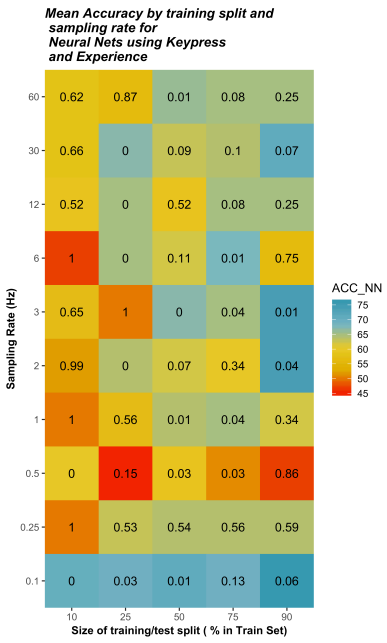


Fig. 7.20. heat-map of Prediction Accuracy for multinomial logistic regression models with experience

Effect of Sampling Rate and Training Set Size

Figure 7.23 and Figure 7.24 are heat-maps showing the classification accuracy for random forest models built using the number of keypress and the number of keypress and experience variables respectively. These two families of models had the best predictive accuracy of all models excluding experience only models. The bluer values represent higher accuracy. The numbers are the p value for the model.

The first conclusion is that a little under a quarter of all combinations of training split and sampling rate will result in a statistically significant model when experience is not included. There are 16 combinations of training split and sampling rate will result in a statistically significant model when experience is not included. This is further evidence that experience does provide a smaller benefit to model‘s significance. However, models built with experience are not more accurate than models built without experience.

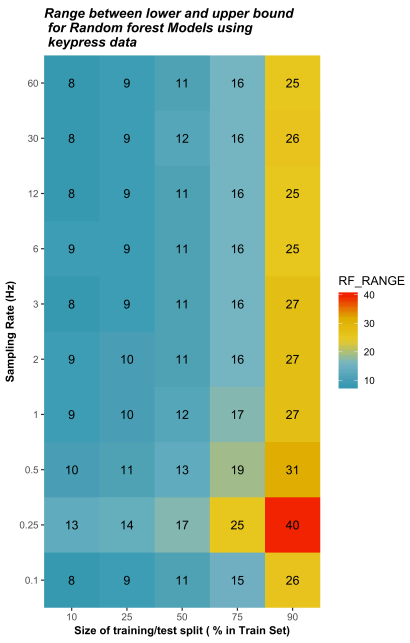


Fig. 7.21. heat-map of Confidence Intervals for Random Forest Models without experience

The inclusion of experience to will increase the size of the training split required to produce a statistically prediction. Further models built using data sampled at slower rates did perform worse than models built using data sampled at a higher sampling rate. The best models were built using 90% of the data which included experience and was greater than 6Hz.

Interpretation

The evidence demonstrates that neural network models can predict delay group under specific training split conditions. Models built using 25 to 50% of the available data are more likely to be statistically significant at variety of sampling frequencies when experience is not a factor. The inclusion of experience does increase the size of the training split required to produce a statistically significant model.

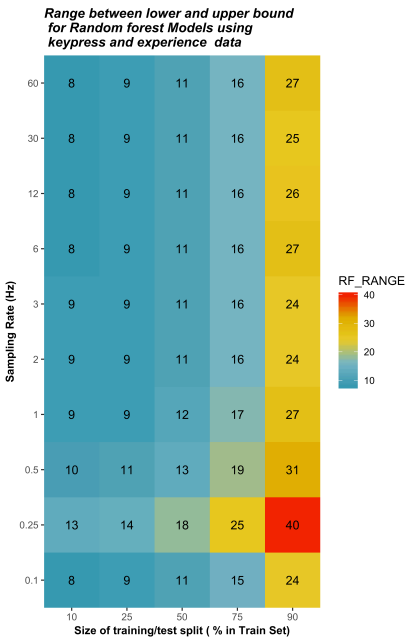


Fig. 7.22. heat-map of Confidence Intervals for Random Forest Models with experience

Hz and 1/10 Hz

These two sampling frequencies represent a special case for all delay prediction models regardless of model type, training set split and combination of predictor variables. No model built using data sampled at a 1/4Hz was statistically significant. Models built using data sampled faster than this frequency were not guaranteed to be statistically significant but for the majority of combinations of experience, sampling rate and training split there was at least one statistically significant model. The evidence shows that the minimum viable sampling frequency is a Hz.

Models built at 1/10 could be statistically significant, but they are excluded from consideration for the following reason. Per Table 7.1 the minimum recorded response time was a little over 6 seconds while the maximum response time was 20 seconds. Further, there were only three possible response outcomes. Considering that the 'real

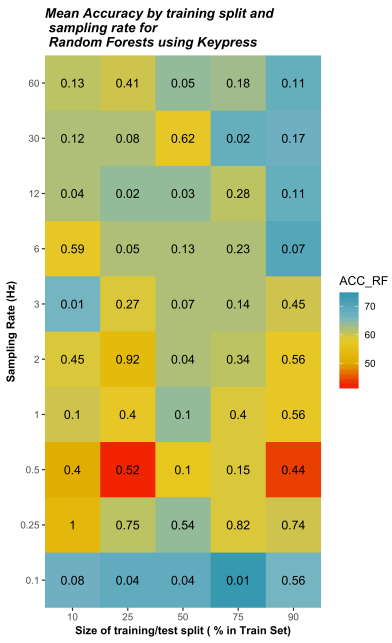


Fig. 7.23. heat-map of Prediction Accuracy for Random Forest Models without experience

'response time was between 0 and 30, sampling at a 1/10 of a hertz fundamentally misrepresents the world.

7.3.4 Discussion

The models shown in the previous sections were delay classification models built using the number of keypress and the number of keypress and experience variables respectively. These two families of models had the best predictive accuracy of all models excluding experience only models. The evidence confirms that it is possible to construct a model that will predict delay group at a rate above chance.

Three additional findings can be derived from the data. The first is that neural networks are the most consistent models when it comes to predicting delay group while multinomial logistic regression models are the least consistent. Further there is evidence that a statistically significant model can be constructed for most combi-

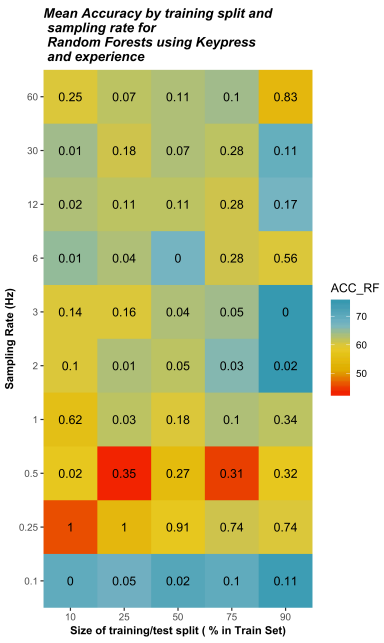


Fig. 7.24. heat-map of Prediction Accuracy for Random Forest Models with experience

nations of predictor variables, sampling frequency and training split. There is strong evidence that proves that the minimum sampling rate is 1/2 Hz.

8. SUMMARY AND DISCUSSION OF RESULTS

The following hypotheses were tested during this experiment:

Null Hypothesis 1 : The first null hypothesis states that the number of correct classifications of task difficulty is equal to chance (1/3).

Hypothesis 1 : The first hypothesis is that the number of correct classifications of participant workload is greater than chance.

Null Hypothesis 2 : The second null hypothesis states that the number of correct classifications of participant delay groupings is equal to chance (1/N) where N is the number of clusters which best describes the data.

Hypothesis 2 : The second hypothesis states that there is a sampling rate and a training/test split below which a model would perform no better than chance.

Null Hypothesis 3 : The third null hypothesis states that the inclusion of demographic measures would not decrease the confidence interval of a prediction.

Hypothesis 3 : The third hypothesis states that the inclusion of demographic measures would decrease the confidence interval of a prediction.

There is sufficient evidence to reject Null Hypothesis 1 in favor of Hypothesis 1. The results of this research project indicate that task difficulty, and by proxy subjective workload, can be determined using recorded and derived system state data at a rate above chance. The evidence further demonstrates that an individual‘s delay in responding to an on-screen event cannot be consistently predicted using regression models. Further, there is sufficient evidence to reject Null Hypothesis 2 in favor of Hypothesis 2. The evidence demonstrate that the delay cluster a participant will belong to can be predicted at a rate above chance.

There is sufficient evidence to reject Null Hypothesis 3 in favor of Hypothesis 3. The inclusion of experience will reduce the confidence interval of a categorization prediction but decrease the accuracy of said prediction across multiple training splits and sampling rates. In the context of delay prediction, the inclusion of experience will increase the size of the training split required to produce a statistically significant prediction for random forest models. The inclusion of experience will have a detrimental effect on the prediction accuracy of multinomial logistic regression and neural network models by reducing the total number of significant models. These results indicate that an analyst can produce a statistically significant difficulty and delay classification model both with and without the benefit of knowing the participant‘s experience.

The evidence does not support the rejection of Null Hypothesis 2 for difficulty prediction models. There was no combination of training set split and sampling rate that would result in a statistically insignificant difficulty prediction model. Hypothesis 2 could be falsified for delay prediction models. There were multiple combinations of sampling rates and raining set splits that results in a statistically insignificant prediction. Further work is needed to determine if this outcome is attributable to the existence of a class imbalance.

8.1 Discussion

One of the main inspirations for this work stemmed from a paper that demonstrated that mobile phones can be used to predict individual activity levels. The same paper showed that it was possible to predict if an individual user had Parkinson's disease based on how the phone was moving (Antos, Albert, & Kording, 2014). The ability to detect if a phone was carried on a belt, in a pocket, or in a purse is not broadly useful to the human factors community. The source and volume of data used to make those predictions and the method used to collect the data is useful to the human factors community.

The rapid proliferation of the types of data available in conjunction with an increase in the rate at which data is produced presents a problem for the human factors discipline. The tools and techniques used by human factors practitioners are not designed to model the performance of a human operator in real time or collect performance data an unintrusive manner. The tools presently available to practitioners can precisesly assess an operator's cognitive load when performing a task. In order to measure operator workload a researcher either needs to wait for the task to conclude or interrupt the operator while performing the task. The literature shows that interrupting an operator during a task has a negative impact on task performance (Bailey & Konstan, 2006; Bailey, Konstan, & Carlis, 2001; Adamczyk & Bailey, 2004).

The results of this research show that a measure of human performance can be derived from continuous system state data without interfering with task performance. These results should be read in the same way as the ability to detect where a phone was being carried. Both studies show that information about the operator can be interred using data collected by a system without the operator's explicit knowledge. The results of both studies point to the need to employ supervised learning methods when making a prediction about the operator.

The results further demonstrate that response delay can be predicted under certain conditions. The following caveats do apply to this result. First, the variables used to predict response delay were dependent on time. Changes to the sampling rate caused a change to the values of the predictor and outcome variables. While the change was on the order of milliseconds, the consequences of that change are profound.

A unique consequence of the current research is the ability to measure human performance without knowing demographic information such as gender or age. The implications of this consequence cuts two ways. A positive implication of this work is that an accurate model of operator performance can be derived using system data alone. This allows for a general model of operator performance to be constructed. However, a general performance model would not account for variations in individual performance. For example, a general model of driving behavior may misclassify a

driver as being drunk when in fact they may have neurological issues. Demographic information should be used to tailor a model to to better assess a specific operator.

The results of the paper shows that the inclusion participant experience at least in how it was defined in this study does not improve model performance. This does not mean that experience does not play a role in individual performance; Rather, the result suggests that the operational definition of experience used in this experiment was possibly over-broad.

The results of manipulating the data sampling rate demonstrate that the minimum sampling rate necessary to produce a statistically significant categorization of response delay is 1/2Hz. Data sampled below this frequency either misrepresents the operator‘s true performance or results in a statistically insignificant model. There was no sampling frequency below which task difficulty could not be classified. Finally, the evidence demonstrates that the size of the training/test split has an impact on model performance. Delay and task difficulty classification models trained using a smaller train/test split had smaller confidence intervals but were not necessarily more likely to be statistically significant.

In short, this research shows that a measure of human performance can be inferred from continuous event data. Thus, it is possible to say that *state y was observed therefore, the operator was in state x.*

8.1.1 Modeling Techniques and Data

None of the classification techniques, multi-nominal logistic regression, neural networks, and random forests, clearly performed worse than the others. All three models performed had an average classification accuracy of 75% , excluding statistically insignificant predictions. No clear recommendation can be made on the basis of accuracy.

Multi-nominal logistic regression models were the most likely to be statistically insignificant when predicting response delay. Neural network and random forest mod-

els were less likely to be statistically insignificant under the same conditions. Both neural networks and random forests had similar predictive accuracy when predicting task difficulty and response delay. Random forests should be used in future research as they are less computationally intensive.

Alterations to features in the data have a greater impact on the prediction of response time. The consequences of these changes should be interpreted through the lens of signal processing. Decreasing the rate at which the data is sampled leads to the system loosing information. This is most evident when calculating response time.

An individual may have responded to the pump failure in two and a half seconds. If the sampling window is larger than a half second, then the individual's response will be collected at the next whole interval. This would transform a 2 1/2 second response into a 3 second response. This loss of a 1/2 second may be imperceptible to an operator but it could lead a system to acting in an unexpected way. This does not occur when predicting task difficulty. The data used to predict difficulty was not dependent on time, but it was dependent on position. It is possible to meaningfully reconstruct the mouse trajectory even if rows are missing.

9. IMPLICATIONS

The tools and techniques used by human factors practitioners were not intended to collect and process data in near real-time. Imagine you are measuring operator workload under various road conditions using the NASA-TLX under laboratory conditions. Ask the driver to complete the TLX too frequently and their driving performance will suffer. You are left with the question: did drivers perform the task poorly because you kept distracting them or because the task was actually hard? Ask the driver to complete the TLX too infrequently and you cannot easily identify what aspects of the task required the driver to exert more effort.

This example points to multiple growth direction for the human factors discipline. The first direction entails the development of tools that measure human performance in near real-time. This development process would be slow and expensive. The other direction would combine elements of data science with human factors engineering techniques.

A researcher would need to clearly link a state of the system with a measure of operator performance in order to properly combine data science and human factors engineering techniques. This link would need to be verified under laboratory conditions. During this stage the research would also formulate operator archetypes. These archetypes would be used to describe collections of individuals with similar performance characteristics.

It possible to find multiple operator archetypes given a sufficiently large volume of data by using unstructured learning methods. Once these archetypes have been identified, you could designate one group as safe drivers and another as unsafe drivers. However, this blind classification would not account for individual variability or external environmental factors. Laboratory studies would be needed to ensure that the archetypes were reflective of real operators.

The data science and human factors disciplines complement each other. The results of human factors studies help guide data science research by providing an objective measure of a behavior. Data science provides human factors engineers with data from hundreds of participants. This allows for human factors experiments to be conducted at scale.

The results of this research have implications for research related to human automation pairings. One of the key questions in automation design is when to exclude or include the human in a control loop. The ability to predict how long an operator will take to respond to an unexpected event is beneficial for determining when the automation could take preemptive action. A similar argument could be constructed for when to delegate control to the human or to the automation based on operator workload.

However, these solutions may create more problems than they solve. A blanket rule stating that any nonoptimal response is 'incorrect'ignores the contexts which operators will find themselves in. Many solutions that appear suboptimal are perfectly valid alternate solutions. For example, rerouting fuel through the working parts of a system to bypass the failed parts is an equally valid solution to restarting a broken pump.

This research has further implications on the design of systems that employ rule based automation. Imagine a system that operated using the following rules:

- When workload is high, take control from the operator until stable workload is reached.
- When workload is low, give control to the operator until stable workload is reached.

The goal of the automation is to keep operator workload at some predetermined equilibrium. However, there are circumstances where high workload is desirable and removing control may hinder the operator. If these rules are applied without consid-

eration to operational context, then the automation may contribute to performance issues.

9.1 Experimental Designs

The data collection method used in this experiment has implications for future work involving crowd-sourced human factors experiments. The task used in this experiment was designed to quickly capture data from a large subject pool with minimal investment. The ability to quickly collect responses from a large and diverse subject pool allows a researcher to capture a wider variety of responses at a marginal cost. The behavior of the task can be easily altered in response to new discoveries in the data. For example, the task can be altered so that the pump fails two minutes into the experiment by changing one line of code.

There are risks with this low-cost, high-yield design method. The lack of direct oversight makes it difficult to identify low effort responses in real-time. The task instructions have to be as unambiguous as possible to prevent confusion. For example, one of the participants in the study contacted the experimenter asking what 'pull the reticle' meant. They explained that English was not their first language and had never seen the verb pull used in this context before.

The goal of this experiment is not to replace or supplant laboratory experiments. There are certain tasks, such as driving, that are best studied in a laboratory setting. Laboratory experiments also permit a degree of control over when and how stimuli are presented to subjects. Critically, laboratory experiments allow for a clear link between between stimulus and response to be established.

Ideally, laboratory studies and low-cost, high-yield studies would be used to complement. Laboratory experiments are needed to clearly establish if the task captures the phenomena of interest. The low-cost, high-yield studies are needed to capture the phenomena at scale.

9.2 Data Driven Discrimination

It is irresponsible to minimize the negative implications of this research while overselling the positive implications.

> You know that this is every pilot or machine operator‘s worst nightmare because it would open up the possibility for an airline or employer to establish an expected response time to a particular event? This would create all sorts of instances of people being penalized for not meeting some form of expected response time or manner of their response. Commercial pilots on the more modern airlines have just about every aspect of their performance analyzed by their employer as it is which adds pressure to an already complicated and stressful work environment.

The quote above was sent in response to a participant recruitment advert and perfectly captures the negative consequences of this research. The ability to predict reaction times and assess the state of the operator is beneficial. Doing so without an understanding of individual variation or consideration of context may result in harmful outcomes.

One could easily imagine a situation where an insurance company altered your rates in response to your driving behavior. This could potentially be used to discourage dangerous behaviors such as erratic driving by imposing a financial penalty on the behavior. If the system did not know if a driver has a neurological impairment, it is possible that individual's behavior could be seen as erratic driving. This introduces the real possibility of algorithmic discrimination.

The question is not if data will be used in such a manner; the question is when.

9.3 Beyond the Data

There are several claims that are suggested but cannot be supported by the data. First, demographic data may be sufficient but not necessary when predicting individ-

ual workload. Demographic data was not included as a predictor in this experiment, yet statistically significant prediction models were constructed.

The data suggests that the perceived difficulty of a task can be dynamically manipulated in real time. Dynamic difficulty adjustment (DDA) is employed by game designers to tailor the game play experience to an individual. Dynamic difficulty adjustment could be used in automated driving or situations where humans are supervising automated agents.

Finally, the data suggests that the tools and methods employed could be used to detect other aspects of human performance. The web-based interface is based on an existing interface that has been used to study human performance. It is reasonable to assume that the current interface could be modified to measure operator distraction.

9.4 Human Factors Education

The human factors discipline has reached an inflection point in regards to data. It is an undeniable fact that data is going to be produced at faster rates and higher volumes. The tools and techniques used by human factors professionals were not designed to process large volumes of data in near real-time.

Human factors education will need to adapt to better reflect the changing data landscape that researchers and practitioners will face in the future. This could happen in several ways. Educational institutions could encourage human factors students to take courses in data science to teach them how to navigate large, complex, and multi-faceted data sets. Machine learning and programming courses could become required.

The goal is not to make human factors engineers expert data scientists. Data science is a distinct discipline that draws from a different skill set. The aim of such education would be to: a) provide human factors practitioners with enough understanding of data science to properly apply these techniques to human factors problems, and b) preserve the discipline by applying a skeptical lens to data science methods. There is the temptation among data scientists to label a sub-set of performance data as the

result of sub-optimal human performance. Human factors practitioners will need to use their expertise to rein in these potentially erroneous classifications.

10. LIMITATIONS

The results of this research are subject to the several limitations. The MATB-II has previously been used in studies involving both pilots and gamers. This is why gamers and pilots were chosen for this study.

One can directly compare the experience level of two pilots by asking what certifications they have obtained and their total logged flight hours. Similarly, it is possible to compare the experiences of two games by asking how many hours of gaming they play in a week. However, the experiences between the two groups may not be directly comparable.

The measure of experience used in this experiment only counted the number of hours per week and per month a participant had participated in either activity. This measure excludes additional information (e.g. the type of games played or type certification) that can be used to describe participant experience. The resulting measure of experience is possibly uninformative. This limits the ability to definitively state that experience has a negligible impact when predicting either response delay or task difficulty.

The results are further limited by the number of gamers and pilots who participated in the study. Only 30 pilots completed the study. This could be attributed to one of two factors. There is a financial barrier of entry to being a pilot. This barrier would naturally reduce the number of pilots that can be studied. Gaming does not have this financial barrier. This barrier to access limits the ability to clearly determine if there was a significant difference between pilot performance and gamer performance.

10.1 Technical Limitations

The results of this research are subject to the following technical limitations. The data was resampled at the following intervals: 30Hz, 12Hz, 6Hz, 3Hz, 2Hz, 1Hz, Hz, 1/4Hz and 1/10Hz as described in 3.4.1. With the exception of the 12Hz and 1/10Hz sampling rates, each sampling rate is the Nyquist frequency of the proceeding value. This sequence guaranteed that the resampling algorithm would need to sample rows at a whole interval .The algorithm would have needed to interpolate the value of a fractional row which would have injected additional error into the data.

The second technical limitation stems from the non-exhaustive search all possible modeling techniques and resampling parameters. The CARET package in R allows for the researcher to specify a wide range of model tuning and resampling parameters. Further, the CARET package comes with hundreds of different modeling techniques that can be used. Given the computing resources available to the researcher, the performance of each class of predictive model was not examined. The models selected for this experiment were models that had a long publication record and easily accessible coding documentation.

Finally, the present research did not investigate if non-linear models would be able to accurately predict task difficulty or delay. The relationship between the response delay and predictor variables may not be linear. As such, applying a linear method would result in meaningless predictions.

11. FUTURE RESEARCH

The following chapter has been divided into two sections dedicated to future research: the first section details the logical next steps that can be taken in the short term to expand upon the findings presented here; the second section discusses the long term research projects that will stem from the current research. The third section discusses how this research could be scaled up.

11.1 Next Steps

There are four suggested directions in the immediate future research could move. The first research direction wise to examine if delay prediction can be solved by examining the delay in the context of a sliding window problem. It may be possible that the current prediction variables reliably predict delay when examining data 30 to 60 seconds before an unexpected failure. This would mean that a system would only need to consider recent user interactions when making a prediction.

The second direction would examine if the inclusion of explicit demographic measures-such as age, gender, etc- improve the classification accuracy of a model. These demographic measures were not included in the current research for the following reasons. There are many factors-including age, gender, handidness and and fatigue- that effect reaction time (Karia, Ghuntla, Mehta, Gokhale, & Shah, 2012). The demographic data did contain records of age and gender; however, there was no a priori reason to suspect that age or gender would be predictive of task difficulty. Furthermore, the majority of participants were in their twenties.

This is not to say that age or gender would not have an effect on response time to a pump failure. However, such a study would need to control for which pump had failed and how many times a participant had interacted with each pump. If a pump

that participants frequently interacted with failed, it would be expected that they would have a faster reaction time.

Future research could also investigate if non-linear methods can be used to predict response delay. The current results show that delay group can be predicted under certain combinations of sampling rate and sampling set size. Additional work is needed to determine if there are combinations of that consistently predict response delay.

Finally, subsequent research will focus on applying the tools and techniques developed by this work to different domains. This would happen in one of two ways. The first would be to apply these methods to data collected from a live system such as a fleet of cars to determine a measure of operator performance. This would represent a next logical step in demonstrating that measures of human performance can be detected in a live system.

Alternatively, the research would employ the same data collection methodology but look for a different measure of human performance. The MATB-II has been used to study phenomena such as cognitive control and distraction. Given that the web-interface is a recreation of MAT-II, it may be possible to use the tool to measure other elements of human performance.

11.2 Human Factors sans Humans?

The title of this section is designed to be deliberately inflammatory. Proposing to study human performance without measuring or observing human behavior is akin to an entomologist wanting to study insect behavior without collecting samples. This scenario is ridiculous when considered through the lens of laboratory based human factors experiments. However, when viewed through the lens of a low-cost, high-yield design methodology the question does become reasonable.

One of the long term questions raised by the current work is how should demographic information be included into a model to develop an individualized prediction

model? The question is not how demographic data be included but how should that data be collected. The demographic data necessary to produce an individualized model could be used to identify individuals in unexpected ways. There are clear protocols on how to handle individual identifiable data in a laboratory setting. The protocols for handling data generated by a system are less clear.

Researchers will need to answer two questions regarding informed consent. First, does voluntary use of a product that automatically collects data imply consent for that data to be used in research? Second, would a disclosure of system data pose a risk to individual participants? These questions did arise when the Institutional Review Broad reviewed this study. Individuals who use Prolific Academic could be argued to be giving implied consent for their data to be used as a function of their membership. However, this blanket implied consent many not apply to specific studies.

The system data collected in this study reflects individual performance at an arbitrary task. No one would be at risk of losing their job or would come to serious harm if their performance data leaked. This question becomes non-trivial when the discussion focuses on driving data.

This research does have implications for automation design. One of the questions that will need to be addressed is how and when to transfer control from the human to the automated system or vice-versa. The current research suggests that it is possible to detect when transfer is necessary. These results offer little insight into how to transfer control.

A final long term question raised by this research is: what should the role of human factors be when it comes to data science? The systems humans interact with are only getting more complex and are capable of producing data at high rates and volumes. Using unstructured learning it is possible to look at a large set of data, identify some clusters in the data, and declare those clusters as representative of a particular behavior. Without a proper understanding of the nuances of human behavior, this methodology would lead to unfortunate outcomes or support a discriminatory system.

11.3 Human Factors at Scale

One of the sources of inspiration for this thesis came from a harrowing experience while driving. The author was driving home from a sporting event late at night. There was little illumination and the weather had the hint of an autumn chill. This combination of factors lead the driver of the car the author was in to fall asleep for 10 seconds.

A reoccurring example in this paper has been drunk driving. This example was used because: it is an event individuals have first or second hand experience with through media exposure, it is a clearly deleterious and harmful behavior, and it is a notable economic cost. The ability for a car to detect when a driver was impaired, either because of drugs or sleep deprivation, would save lives.

There are several questions that must be answered before this technology could be implemented. What sort of data does the car need to record? Who will have access to that data? Should the system error on the side of caution and assume that a driver is impaired when they may have a neurological issue? These are the sort of questions that will need to be addressed when thinking about human factors at scale.

REFERENCES

REFERENCES

Abdelwahab, O., Bahgat, M., Lowrance, C. J., & Elmaghraby, A. (2015, 12 7). Effect of training set size on svm and naive bayes for twitter sentiment analysis. In (pp. 46–51). doi: 10.1109/ISSPIT.2015.7394379

Abernethy, B., Neal, R. J., & Koning, P. (1994). Visualperceptual and cognitive differences between expert, intermediate, and novice snooker players. *Applied Cognitive Psychology*, *8*(3), 185–211. doi: 10.1002/acp.2350080302

Adamczyk, P. D., & Bailey, B. P. (2004). If not now, when?: the effects of interruption at different moments within task execution. In *Proceedings of the sigchi conference on human factors in computing systems* (pp. 271–278).

Afifi, T. O., Brownridge, D. A., Cox, B. J., & Sareen, J. (2006, 10). Physical punishment, childhood abuse and psychiatric disorders. *Child Abuse Neglect*, *30*(10), 1093–1103. doi: 10.1016/j.chiabu.2006.04.006

Aggarwal, C. C., Watson, T. J., Ctr, R., Han, J., Wang, J., & Yu, P. S. (2003). A framework for clustering evolving data streams. *Proc. of the 29th int. conf. on Very large data bases*, 81–92. doi: 10.1.1.13.8650

Ahituv, N., Igbaria, M., & Sella, A. V. (1998). The effects of time pressure and completeness of information on decision making. *Journal of management information systems*, *15*(2), 153–172.

Andrews, F. M., & Farris, G. F. (1972). Time pressure and performance of scientists and engineers: A five-year panel study. *Organizational Behavior and Human Performance*, *8*(2), 185–200. doi: 10.1016/0030-5073(72)90045-1

Antos, S. A., Albert, M. V., & Kording, K. P. (2014). Hand, belt, pocket or bag: Practical activity tracking with mobile phones. *Journal of neuroscience methods*, *231*, 22–30.

Atkins, J. W., Epstein, H. E., & Welsch, D. L. (2018, 10). Using landsat imagery to map understory shrub expansion relative to landscape position in a mid-appalachian watershed. *Ecosphere*, *9*(10), e02404. doi: 10.1002/ecs2.2404

Avendon, E. M., & Sutton-Smith, B. (1971). The study of games. In *The study of games.*

Ayyagari, R., Grover, V., & Purvis, R. (2011). Technostress: technological antecedents and implications. *MIS Quarterly*, *35*(4), 831–858. doi: 10.1093/bja/aeq366

Bacher, I., Mac Namee, B., & Kelleher, J. D. (2018). Scoped: Evaluating a composite visualisation of the scope chain hierarchy within source code. IEEE.

Baer, M., & Oldham, G. R. (2006). The curvilinear relation between experienced creative time pressure and creativity: moderating effects of openness to experience and support for creativity. *Journal of Applied Psychology*, *91*(4), 963.

Bagley, K. S. (2012). *Conceptual mile markers to improve time-to-value for exploratory search sessions* (Unpublished doctoral dissertation). (AAI3536690)

Bailey, B. P., & Konstan, J. A. (2006). On the need for attention-aware systems: Measuring effects of interruption on task performance, error rate, and affective state. *Computers in human behavior*, *22*(4), 685–708.

Bailey, B. P., Konstan, J. A., & Carlis, J. V. (2001). The effects of interruptions on task performance, annoyance, and anxiety in the user interface. In *Interact* (Vol. 1, pp. 593–601).

Baldwin, D. C., & Daugherty, S. R. (2004). Sleep deprivation and fatigue in residency training: results of a national survey of first- and second-year residents. *Sleep*, *27*(2), 217–23. doi: http://dx.doi.org/10.1093/sleep/27.2.217

Banbury, S. P., Macken, W. J., Tremblay, S., & Jones, D. M. (2001). Auditory distraction and short-term memory: Phenomena and practical implications. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *43*(1), 12–29. doi: 10.1518/001872001775992462

Bansal, N., Blum, A., & Chawla, S. (2004). Correlation clustering. *Machine Learning*, *56*(1-3), 89–113.

Barandiaran, I. (1998). The random subspace method for constructing decision forests. *IEEE transactions on pattern analysis and machine intelligence*, *20*(8).

Bartle, R. M. L. (1996). Hearts, clubs, diamonds, spades: Players who suit muds. *Journal of MUD research*, *6*(1), 39. doi: 10.1007/s00256-004-0875-6

Beach, L. R., & Mitchell, T. R. (1978). A contingency model for the selection of decision strategies. *The Academy of Management Review*, *3*(3), 439–449.

Belgiu, M., & Drgu, L. (2016, 4). Random forest in remote sensing: A review of applications and future directions. *ISPRS Journal of Photogrammetry and Remote Sensing*, *114*, 24–31. doi: 10.1016/j.isprsjprs.2016.01.011

Bergersen, G. R., Sjberg, D. I., & Dyb\a a, T. (2014). Construction and validation of an instrument for measuring programming skill. *IEEE Transactions on Software Engineering*, *40*(12).

Bevan, G., & Hood, C. (2006). What's measured is what matters: Targets and gaming in the english public health care system. *Public Administration*, *84*(3), 517–538. doi: 10.1111/j.1467-9299.2006.00600.x

Bigus, J. P. (1996). *Data mining with neural networks: solving business problems from application development to decision support*. McGraw-Hill, Inc.

Bischl, B., Lang, M., Kotthoff, L., Schiffner, J., Richter, J., Studerus, E., ... Jones, Z. M. (2016). mlr: Machine learning in r. *The Journal of Machine Learning Research*, *17*(1), 59385942.

Blaauw, G. J. (1982, 8). Driving experience and task demands in simulator and instrumented car: A validation study. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *24*(4), 473–486. doi: 10.1177/001872088202400408

Blashfield, R. K. (1976). Mixture model tests of cluster analysis: Accuracy of four agglomerative hierarchical methods. *Psychological Bulletin*, *83*(3), 377.

Boot, W. R., Kramer, A. F., Simons, D. J., Fabiani, M., & Gratton, G. (2008). The effects of video game playing on attention, memory, and executive control. *Acta Psychologica*, *129*(3), 387–398. doi: 10.1016/j.actpsy.2008.09.005

Borst, J. P., Taatgen, N. A., & van Rijn, H. (2010). The problem state: A cognitive bottleneck in multitasking. *Journal of Experimental Psychology: Learning, memory, and cognition*, *36*(2), 363.

Bostan, B., & Ogut, S. (2009). Game challenges and difficulty levels: lessons learned from rpgs. *International Simulation and Gaming Association Conference*, 1–11.

Bottleneck theory. (0). Retrieved from `https://www.alleydog.com/glossary/definition-cit.p` ([Online; accessed 2017-12-24])

Bowey, J. T., Birk, V. M., & Mandryk, R. L. (2015). Manipulating leaderboards to induce player experience. In (pp. 115–120). doi: 10.1145/2793107.2793138

Breiman, L. (1996). Bagging predictors. *Machine Learning*, *24*(2), 123–140. doi: 10.1007/BF00058655

Broach, D. M., & Dollar, C. S. (2002). *Relationship of employee attitudes and supervisor-controller ratio to en route operational error rates* (Tech. Rep.).

Brooks, J. (2015). Cognitive fatigue: Exploring the relationship between the fatigue effect and action video-game experience.

Bryan, J. F., & Locke, E. A. (1967). Parkinson's law as a goal-setting phenomenon. *Organizational Behavior and Human Performance*, *2*, 258–275. doi: 10.1016/0030-5073(67)90021-9

Bull, S. B., & Donner, A. (1987, 12). The efficiency of multinomial logistic regression compared with multiple group discriminant analysis. *Journal of the American Statistical Association*, *82*(400), 1118–1122. doi: 10.1080/01621459.1987.10478548

Bunse, C. (2006). Using patterns for the refinement and translationof uml models: A controlled experiment. *Empirical Software Engineering*, *11*(2), 227267.

Burke, M. J., & Day, R. R. (1986). A cumulative study of the effectiveness of managerial training. *Journal of applied Psychology*, *71*(2), 232.

Caldwell, J. A., & Ramspott, S. (1998). Effects of task duration on sensitivity to sleep deprivation using the multi-attribute task battery. *Behavior Research Methods, Instruments, Computers*, *30*(4), 651–660.

Campbell, M. K., & Donner, A. (1989, 6). Classification efficiency of multinomial logistic regression relative to ordinal logistic regression. *Journal of the American Statistical Association*, *84*(406), 587–591. doi: 10.1080/01621459.1989.10478807

Card, S. K., English, W. K., & Burr, B. J. (1978). Evaluation of mouse, rate-controlled isometric joystick, step keys, and text keys for text selection on a crt. *Ergonomics*, *21*(8), 601–613.

Card, S. K., Moran, T. P., & Newell, A. (1983). *The psychology of human-computer interaction* (No. 1). doi: 10.1007/s13398-014-0173-7.2

Carthey, J., Walker, S., Deelchand, V., Vincent, C., & Griffiths, W. H. (2011). Breaking the rules: understanding non-compliance with policies and guidelines. *BMJ: British Medical Journal (Online)*, *343*.

Cechanowicz, J. E., Gutwin, C., Bateman, S., Mandryk, R., & Stavness, I. (2014). Improving player balancing in racing games. In (pp. 47–56).

Chernbumroong, S., Sureephong, P., & Muangmoon, O. O. (2017). The effect of leaderboard in different goal-setting levels. In (pp. 230–234). doi: 10.1109/ICDAMT.2017.7904967

Chiappe, D., Conger, M., Liao, J., Caldwell, J. L., & Vu, K.-P. L. (2013). Improving multi-tasking ability through action videogames. *Applied ergonomics*, *44*(2), 278–284.

Chora, M., Bhanu, B., Chen, H., Champod, C., Komatsu, N., Nakano, M., ... Liu, Z. (2009). Ensemble Learning. *Encyclopedia of Biometrics*, 270–273. Retrieved from `http://www.springerlink.com/index/10.1007/978-0-387-73003-5`$_2$93 doi: 10.1007/978-0-387-73003-5$_2$93

Comstock Jr, J. R., & Arnegard, R. J. (1992). The multi-attribute task battery for human operator workload and strategic behavior research.

Cooper, G. E., & Harper Jr, R. P. (1969). *The use of pilot rating in the evaluation of aircraft handling qualities* (Tech. Rep.).

Cutler, D. R., Edwards, T. C., Beard, K. H., Cutler, A., Hess, K. T., Gibson, J., & Lawler, J. J. (2007, 11). Random forests for classification in ecology. *Ecology*, *88*(11), 2783–2792. doi: 10.1890/07-0539.1

Dang, X. H., Lee, V. C., Ng, W. K., & Ong, K. L. (2009). Incremental and adaptive clustering stream data over sliding window. In (Vol. 5690 LNCS, pp. 660–674). doi: 10.1007/978-3-642-03573-9$_5$5

Dao, A.-Q. V., Parkinson, J. R., & Landry, S. J. (2016, 9). Identifying human-machine interaction problems in continuous state data. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *60*(1), 86–90. doi: 10.1177/1541931213601019

Deci, E. L., Koestner, R., & Ryan, R. M. (1999). A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. *Psychological Bulletin*, *125*(6), 627–668. doi: 10.1037/0033-2909.125.6.627

Deci, E. L., & Ryan, R. M. (1980). The empirical exploration of intrinsic motivational processes. *Advances in Experimental Social Psychology*, *13*(C), 39–80. doi: 10.1016/S0065-2601(08)60130-6

Deterding, S. (2010a). Just add points? what ux can (and cannot) learn from game design. *UXCamp Europe, Berlin, 30 May 2010*.

Deterding, S. (2010b). Pawned. gamification and its discontents. *Playful 2010*.

Dietterich, T. G. (1997). Machine learning research: Four current directions. *AI Magazine*, *18*, 97–136.

Dourish, P., & Bellotti, V. (1992). Awareness and coordination in shared workspaces. *Proc. Intl. Conf. on Computer-Supported Cooperative Work*(November), 107–114. doi: 10.1145/143457.143468

Drews, F. A., Pasupathi, M., & Strayer, D. L. (2008). Passenger and cell phone conversations in simulated driving. *Journal of Experimental Psychology: Applied*, *14*(4), 392.

Drury, C. G. (2015, 5 4). Human factors/ergonomics implications of big data analytics: Chartered institute of ergonomics and human factors annual lecture. *Ergonomics*, *58*(5), 659–673. doi: 10.1080/00140139.2015.1025106

Drury, J. L., & Scott, S. D. (2008). Awareness in unmanned aerial vehicle operations. *The International C2 Journal*, *2*(1), 1–10.

Eastridge, B. J., Hamilton, E. C., O'Keefe, G. E., Rege, V. R., Valentine, R. J., Jones, D. J., ... Thal, E. R. (2003). Effect of sleep deprivation on the performance of simulated laparoscopic surgical skill. In (Vol. 186, pp. 169–174). doi: 10.1016/S0002-9610(03)00183-1

Eckhardt, A., Maier, C., & Buettner, R. (2012). The influence of pressure to perform and experience on changing perceptions and user performance: a multi-method experimental analysis.

Eggemeier, F., Wilson, G., Kramer, A., & Damos, D. (1991). Workload assessment in multi-task environments, chapter 9. multiple-task performance.

ElBardissi, A. W., Wiegmann, D. A., Dearani, J. A., Daly, R. C., & Sundt, T. M. (2007). Application of the human factors analysis and classification system methodology to the cardiovascular surgery operating room. *The Annals of Thoracic Surgery*, *83*(4), 1412–1419.

Elkan, C. (2001). The foundations of cost-sensitive learning. In (Vol. 17, p. 973978). Lawrence Erlbaum Associates Ltd.

Emanuele, C. F. (2016). Big data analytics as a tool for reducing ergonomics risk. *Journal of Ergonomics*, *07*(01). Retrieved from `https://www.omicsgroup.org/journals/big-data-analytics-as-a-tool-for-reducing-erg` ([Online; accessed 2019-02-28]) doi: 10.4172/2165-7556.1000e164

Faber, V. (1994). Clustering and the continuous k-means algorithm. *Los Alamos Science*, *22*.

Fabina, N. S., Baskett, M. L., & Gross, K. (2015, 9). The differential effects of increasing frequency and magnitude of extreme events on coral populations. *Ecological Applications*, *25*(6), 1534–1545. doi: 10.1890/14-0273.1

Fallon, C. K., Panganiban, A. R., Wohleber, R., Matthews, G., Kustubayeva, A. M., Roberts, R., ... Villeneuve, M. (2014). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary educational psychology*, *25*(2), 54–67. doi: 10.1006/ceps.1999.1020

Falstein, N. (2005). Understanding fun–the theory of natural funativity. *Introduction to game development*, 71–98.

Feigenspan, J., Kastner, C., Liebig, J., Apel, S., & Hanenberg, S. (2012, 6). Measuring programming experience. In (pp. 73–82). Passau, Germany: IEEE. Retrieved from `http://ieeexplore.ieee.org/document/6240511/` ([Online; accessed 2019-04-04]) doi: 10.1109/ICPC.2012.6240511

Feigenspan, J., & Siegmund, N. (2012). Supporting comprehension experiments with human subjects. IEEE.

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

Foody, G., McCulloch, M. B., & Yates, W. B. (1995, 6 1). The effect of training set size and composition on artificial neural network classification. *International Journal of Remote Sensing*, *16*(9), 1707–1723. doi: 10.1080/01431169508954507

Fowler, B. (1994). P300 as a measure of workload during a simulated aircraft landing task. *Human factors*, *36*(4), 670–683.

Fraiwan, L., Lweesy, K., Khasawneh, N., Wenz, H., & Dickhaus, H. (2012, 10). Automated sleep stage identification system based on timefrequency analysis of a single eeg channel and random forest classifier. *Computer Methods and Programs in Biomedicine*, *108*(1), 10–19. doi: 10.1016/j.cmpb.2011.11.005

Frederick, C. M., & Ryan, R. M. (1993). Differences between sport and physical activity and motivation. *Journal of Sport Behavior*, *16*(3), 125–145.

Frederick, C. M., & Ryan, R. M. (1995). Self-determination in sport: A review using cognitive evaluation theory. *International Journal of Sport Psychology*, *26*(1), 5–23.

Freund, Y., & Schapire, R. E. (1999). Large margin classification using the perceptron algorithm. *Machine Learning*, *37*(3), 277–296. doi: 10.1023/A:1007662407062

Garrett, J. W., & Teizer, J. (2009). Human factors analysis classification system relating to human error awareness taxonomy in construction safety. *Journal of Construction Engineering and Management*, *135*(8), 754–763.

General, A., Da Silva, B., Esteves, D., Halleran, M., & Liut, M. (n.d.). A comparative analysis between the mouse, trackpad and the leap motion.

Gerling, K. M., Miller, M., Mandryk, R. L., Birk, M. V., & Smeddinck, J. D. (2014). Effects of balancing for physical abilities on player performance, experience and self-esteem in exergames. In (pp. 2201–2210).

Gevins, A., & Smith, M. E. (2003). Neurophysiological measures of cognitive workload during human-computer interaction. *Theoretical Issues in Ergonomics Science*, *4*(1-2), 113–131.

Goel, N., Basner, M., Rao, H., & Dinges, D. F. (2013). Circadian rhythms, sleep deprivation, and human performance. *Progress in Molecular Biology and Translational Science*, *119*, 155–190. doi: 10.1016/B978-0-12-396971-2.00007-5

Goodman, P. S., & Shah, S. (1992). Familiarity and work group outcomes.

Green, C. S., & Bavelier, D. (2003). Action video game modifies visual selective attention. *Nature*, *423*(6939), 534–537.

Groves, M., O'Rourke, P., & Alexander, H. (2003, 1). The clinical reasoning characteristics of diagnostic experts. *Medical Teacher*, *25*(3), 308–313. doi: 10.1080/0142159031000100427

Gutwin, C., & Greenberg, S. (1996). The effects of workspace awareness support on the usability of real-time distributed groupware. *Proceedings of the 1996 ACM annual conference on Human Factors in Computing Systems - CHI '96*, *6*(3), 511–518. doi: 10.1145/345190.345222

Gutwin, C., & Greenberg, S. (1999). A framework of awareness for small groups in shared-workspace groupware. *Computer-Supported Cooperative Work*, *3-4*, 411–446.

Gutzwiller, R. S., Wickens, C. D., & Clegg, B. A. (2014). Workload overload modeling: An experiment with matb ii to inform a computational model of task management. In (Vol. 58, pp. 849–853).

Gutzwiller, R. S., Wickens, C. D., & Clegg, B. A. (2015). The role of individual differences in executive attentional networks and switching choices in multi-task management. In (Vol. 59, pp. 632–636).

Gutzwiller, R. S., Wickens, C. D., & Clegg, B. A. (2016). The role of time on task in multi-task management. *Journal of Applied Research in Memory and Cognition*, *5*(2), 176–184.

Guzzo, R. A., Jette, R. D., & Katzell, R. A. (1985). The effects of psychologically based intervention programs on worker productivity: A meta-analysis. *Personnel psychology*, *38*(2).

H., B. G., S., J. B., & R., M. U. R. (2012, 6 1). Multinomial logistic regression model for single-vehicle and multivehicle collisions on urban u.s. highways in arkansas. *Journal of Transportation Engineering*, *138*(6), 786–797. doi: 10.1061/(ASCE)TE.1943-5436.0000370

Hamari, J., Koivisto, J., & Sarsa, H. (2014). Does gamification work?–a literature review of empirical studies on gamification. In (pp. 3025–3034).

Hamari, J., & Tuunanen, J. (2014). Player types: A meta-synthesis. *Transactions of the Digital Games Research Association*, *1*(2).

Hansen, L. K., & Salamon, P. (1990). Neural network ensembles. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *12*(10), 993–1001. doi: 10.1109/34.58871

Hart, S. G., & Staveland, L. E. (1988). Development of nasa-tlx (task load index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139–183). Elsevier. Retrieved from https://linkinghub.elsevier.com/retrieve/pii/S0166411508623869 (DOI: 10.1016/S0166-4115(08)62386-9)

Hasan, M., Choi, J., Neumann, J., Roy-Chowdhury, A. K., & Davis, L. S. (2016). Learning temporal regularity in video sequences.

doi: 10.1109/CVPR.2016.86

Hellervik, L. W., Hazucha, J. F., & Schneider, R. J. (1992). Behavior change: Models, methods, and a review of evidence.

Ho, T. K. (1995). Random decision forests. In (Vol. 1, p. 278282). IEEE.

Hoffman, R. R. (1998). How can expertise be defined? implications of research from cognitive psychology. In *Exploring expertise* (pp. 81–100). Springer.

Holding, D. H. (1979). The evaluation of chess positions. *Simulation Gaming*, *10*(2), 207–221. doi: 10.1177/104687817901000205

Horberry, T., Anderson, J., Regan, M. A., Triggs, T. J., & Brown, J. (2006). Driver distraction: The effects of concurrent in-vehicle tasks, road environment complexity and age on driving performance. *Accident Analysis Prevention*, *38*(1), 185–191.

Hosmer, D. W., & Lemeshow, S. (2000). *Applied logistic regression: Hosmer/applied logistic regression.* Hoboken, NJ, USA: John Wiley Sons, Inc. Retrieved from http://doi.wiley.com/10.1002/0471722146 (DOI: 10.1002/0471722146)

Hossain, M., Wright, S., & Petersen, L. A. (2002, 4). Comparing performance of multinomial logistic regression and discriminant analysis for monitoring access to care for acute myocardial infarction. *Journal of Clinical Epidemiology*, *55*(4), 400–406. doi: 10.1016/S0895-4356(01)00505-4

Huizinga, J. (2007). Homo ludens: A study of the play-element in culture. *European Early Childhood Education Research Journal*, *19*(2), 1–24. doi: 10.1177/0907568202009004005

Hwang, M. I. (1994). Decision making under time pressure: a model for information systems research. *Information Management*, *27*(4), 197–203.

Jex, S. M. (1998). *Stress and job performance: Theory, research, and implications for managerial practice.* Thousand Oaks, CA: Sage Publications Ltd.

Johnson, D., Nacke, L. E., & Wyeth, P. (2015). All about that base: Differing player experiences in video game genres and the unique case of moba games. In (pp. 2265–2274). doi: 10.1145/2702123.2702447

Jordan, C. S. (1992). Experimental study of the effects of an instantaneous self assessment workload recorder on task performance. *Report No. DRA/TM (CAD5)/92011. Farnborough: Defence Evaluation Research Agency*.

Juul, J. (2002). The game, the player, the world: Looking for a heart of gameness. *Proceedings at the Level Up: Digital Games Research Conference*, 30–45. doi: 10.3200/JOEE.39.2.47-58

Kallio, K. P., Myr, F., & Kaipainen, K. (2011). At least nine ways to play: Approaching gamer mentalities. *Games and Culture*, *6*(4), 327–353. doi: 10.1177/1555412010391089

Karia, R. M., Ghuntla, T. P., Mehta, H. B., Gokhale, P. A., & Shah, C. J. (2012). Effect of gender difference on visual reaction time: A study on medical students of bhavnagar region. *IOSR Journal of Pharmacy*, *2*(3), 452–454.

Kirsh, D. (2000). A few thoughts on cognitive overload. *Intellectica*, *1*(30), 19–51. doi: 10.1128/AAC.03728-14

Kool, W., McGuire, J. T., Rosen, Z. B., & Botvinick, M. M. (2010). Decision making and the avoidance of cognitive demand. *Journal of Experimental Psychology: General*, *139*(4), 665.

Krulak, D. C. (2004). Human factors in maintenance: impact on aircraft mishap frequency and severity. *Aviation, space, and environmental medicine*, *75*(5), 429–432.

Kuhn, M. (2008). Building predictive models in r using the caret package. *Journal of statistical software*, *28*(5), 1–26.

La Guardia, J. G., Ryan, R. M., Couchman, C. E., & Deci, E. L. (2000). Within-person variation in security of attachment: A self-determination theory perspective on attachment, need fulfillment, and well-being. *Journal of Personality and Social Psychology*, *79*(3), 367–384. doi: 10.1037/0022-3514.79.3.367

Latham, A. J., Patston, L. L. M., & Tippett, L. J. (2013). Just how expert are "expert" video-game players? assessing the experience and expertise of video-game players across "action" video-game genres. *Frontiers in Psychology*, *4*(DEC), 1–3. doi: 10.3389/fpsyg.2013.00941

Latham, G. P., & Locke, E. a. (1975). Increasing productivity and decreasing time limits: A field replication of parkinson's law. *Journal of Applied Psychology*, *60*, 524–526. doi: 10.1037/h0076916

Lawton, R. (1998). Not working to rule: understanding procedural violations at work. *Safety science*, *28*(2), 77–95.

Lee, J. D., Young, K. L., & Regan, M. A. (2008). Defining driver distraction. *Driver distraction: Theory, effects, and mitigation*, *13*(4), 31–40.

Lewis, D. (1998). *Naive(bayes)at forty: The independence assumption in information retrieval* (Vol. 1398).

Littlepage, G., Robison, W., & Reddington, K. (1997, 2). Effects of task experience and group experience on group performance, member ability, and recognition of expertise. *Organizational Behavior and Human Decision Processes*, *69*(2), 133–147. doi: 10.1006/obhd.1997.2677

Liu, H., & Cocea, M. (2017, 12 1). Semi-random partitioning of data into training and test sets in granular computing context. *Granular Computing*, *2*(4), 357–386. doi: 10.1007/s41066-017-0049-2

Lu, H., Setiono, R., & Liu, H. (1996). Effective data mining using neural networks. *IEEE transactions on knowledge and data engineering*, *8*(6), 957–961.

Malone, T. W. (1982). Heuristics for designing enjoyable user interfaces. In (pp. 63–68). doi: 10.1145/800049.801756

Mandell, D. S., Walrath, C. M., Manteuffel, B., Sgro, G., & Pinto-Martin, J. A. (2005, 12). The prevalence and correlates of abuse among children with autism served in comprehensive community-based mental health settings. *Child Abuse Neglect*, *29*(12), 1359–1372. doi: 10.1016/j.chiabu.2005.06.006

Marsh, J., & Jones, D. (2010). Cross-modal distraction by background speech: What role for meaning? *Noise and Health*, *12*(49), 210. doi: 10.4103/1463-1741.70499

Marshall, S. P. (2002). The index of cognitive activity: measuring cognitive workload. *Proceedings of the IEEE 7th Conference on Human Factors and Power Plants*, 5–9. doi: 10.1109/HFPP.2002.1042860

Mayer, R. E., & Moreno, R. (2003). Nine ways to reduce cognitive load in multimedia learning. *Educational Psychologist*, *38*(1), 43–52. doi: 10.1207/S15326985EP3801$_6$

McAfee, A., Brynjolfsson, E., Davenport, T. H., et al. (2012). Big data: the management revolution. *Harvard business review*, *90*(10), 60–68.

McDaniel, R., Lindgren, R., & Friskics, J. (2012). Using badges for shaping interactions in online learning environments.. doi: 10.1109/IPCC.2012.6408619

McGonigal, J. (2011). *Reality is broken: Why games make us better and how they can change the world*. Penguin.

McLachlan, G., Do, K.-A., & Ambroise, C. (2005). *Analyzing microarray gene expression data* (Vol. 422). John Wiley Sons.

Meshkati, N., Hancock, P., Rahimi, M., & M. Dawes, S. (1995). *Techniques in mental workload assessment*.

Miller, C. F. (1993, 12). Actual experience, potential experience or age, and labor force participation by married women. *Atlantic Economic Journal*, *21*(4), 60–66. doi: 10.1007/BF02302329

Miller, K. A., Deci, E. L., & Ryan, R. M. (1988). Intrinsic motivation and self-determination in human behavior. *Contemporary Sociology*, *17*(2), 253. doi: 10.2307/2070638

Mirvis, P. H., Csikszentmihalyi, M., & Csikzentmihaly, M. (1991). *Flow: The psychology of optimal experience.* (Vol. 16) (No. 3). doi: 10.5465/AMR.1991.4279513

Mohri, M., Rostamizadeh, A., & Talwalkar, A. (2013). Foundations of machine learning. *Journal of Chemical Information and Modeling*, *53*(9), 1689–1699. doi: 10.1017/CBO9781107415324.004

Nelson, M. J. (2012). Soviet and american precursors to the gamification of work. In (p. 23). doi: 10.1145/2393132.2393138

Neufeld, V. R., Norman, G. R., Feightner, J. W., & Barrows, H. S. (1981, 9). Clinical problem-solving by medical students: a cross-sectional and longitudinal analysis. *Medical Education*, *15*(5), 315–322. doi: 10.1111/j.1365-2923.1981.tb02495.x

Newheiser, M. (2009). *Playing fair: A look at competition in gaming.* Retrieved from `http://www.strangehorizons.com/2009/20090309/newheiser-a.shtml`

Olson, G. M., & Olson, J. S. (2000). Distance matters. *Human-Computer Interaction*, *15*(2), 139–178. doi: 10.1207/S15327051HCI1523$_4$

Pal, M. (2005, 1). Random forest classifier for remote sensing classification. *International Journal of Remote Sensing*, *26*(1), 217–222. doi: 10.1080/01431160412331269698

Pashler, H. (1994). Dual-task interference in simple tasks: Data and theory. *Psychological Bulletin*, *116*(2), 220–244. doi: 10.1037/0033-2909.116.2.220

Passos, E. B., Medeiros, D. B., Neto, P. A. S., & Clua, E. W. G. (2011). Turning real-world software development into a game. In (pp. 260–269). doi: 10.1109/SBGAMES.2011.32

Patten, C. J. D., Kircher, A., stlund, J., & Nilsson, L. (2004). Using mobile telephones: cognitive workload and attention resource allocation. *Accident analysis prevention*, *36*(3), 341–350.

Patten, C. J. D., Kircher, A., stlund, J., Nilsson, L., & Svenson, O. (2006). Driver experience and cognitive workload in different traffic environments. *Accident Analysis Prevention*, *38*(5), 887–894.

Patterson, J. M., & Shappell, S. A. (2010). Operator error and system deficiencies: analysis of 508 mining incidents and accidents from queensland, australia using hfacs. *Accident Analysis Prevention*, *42*(4), 1379–1385.

Pauley, K., O'Hare, D., & Wiggins, M. (2009). Measuring expertise in weather-related aeronautical risk perception: the validity of the cochran–weiss–shanteau (cws) index. *The International Journal of Aviation Psychology*, *19*(3), 201–216.

Peck, R. C. (1993). The identification of multiple accident correlates in high risk drivers with specific emphasis on the role of age, experience and prior traffic violation frequency. *Alcohol, Drugs Driving*, *9*(3-4), 145–166.

Peng, W., & Hsieh, G. (2012). The influence of competition, cooperation, and player relationship in a motor performance centered computer game. *Computers in Human Behavior*, *28*(6), 2100–2106.

Peters, L. H., O'Connor, E. J., Pooyan, A., & Quick, J. C. (1984). Research note: The relationship between time pressure and performance: A field test of parkinson's law. *Journal of Organizational Behavior*, *5*(4), 293–299.

Polich, J. (2007). Updating p300: An integrative theory of p3a and p3b. *Clinical Neurophysiology*, *118*(10), 2128–2148. doi: 10.1016/j.clinph.2007.04.019

Rashid, H. S., Place, C. S., & Braithwaite, G. R. (2014, 2). Eradicating root causes of aviation maintenance errors: Introducing the ammp. *Cogn. Technol. Work*, *16*(1), 71–90. doi: 10.1007/s10111-012-0245-4

Reason, J. (1990). *Human error*. Cambridge university press.

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. In *Advances in psychology* (Vol. 52, pp. 185–218). Elsevier. Retrieved from `https://linkinghub.elsevier.com/retrieve/pii/S0166411508623870` (DOI: 10.1016/S0166-4115(08)62387-0)

Rennie, J. D. M., Shih, L., Teevan, J., & Karger, D. R. (2003). Tackling the poor assumptions of naive bayes text classifiers. *Proceedings of the Twentieth International Conference on Machine Learning (ICML)-2003)*, *20*(1973), 616–623. doi: 10.1186/1477-3155-8-16

Ricca, F., Di Penta, M., Torchiano, M., Tonella, P., & Ceccato, M. (2007). The role of experience and ability in comprehension tasks supported by uml stereotypes. In (p. 375384). IEEE.

Riley, J. M., & Endsley, M. R. (2005). Situation awareness in hri with collaborating remotely piloted vehicles. In (Vol. 49, pp. 407–411).

Robson, K., Plangger, K., Kietzmann, J. H., McCarthy, I., & Pitt, L. (2015). Is it all a game? understanding the principles of gamification. *Business Horizons*, *58*(4), 411–420. doi: 10.1016/j.bushor.2015.03.006

Rodriguez-Galiano, V., Chica-Olmo, M., Abarca-Hernandez, F., Atkinson, P., & Jeganathan, C. (2012, 6). Random forest classification of mediterranean land cover using multi-seasonal imagery and multi-seasonal texture. *Remote Sensing of Environment*, *121*, 93–107. doi: 10.1016/j.rse.2011.12.003

Roscoe, A. H., & Ellis, G. A. (1990). *A subjective rating scale for assessing pilot workload in flight: A decade of practical use* (Tech. Rep.).

Rozado, D. (2013). Mouse and keyboard cursor warping to accelerate and reduce the effort of routine hci input tasks. *IEEE Transactions on Human-Machine Systems*, *43*(5), 487–493.

Rozado, D. (2013, Sep.). Mouse and keyboard cursor warping to accelerate and reduce the effort of routine hci input tasks. *IEEE Transactions on Human-Machine Systems*, *43*(5), 487-493. doi: 10.1109/THMS.2013.2281852

Ruff, H. A., Calhoun, G. L., Draper, M. H., Fontejon, V. J., & Guilfoos, B. J. (2004). *Exploring automation issues in supervisory control of multiple uavs* (Tech. Rep.).

Russell, S., & Norvig, P. (2013). *Artificial intelligence a modern approach*. doi: 10.1017/S0269888900007724

Ryan, J. P., Hernandez, P. M., & Herz, D. (2007, 6 1). Developmental trajectories of offending for male adolescents leaving foster care. *Social Work Research*, *31*(2), 83–93. doi: 10.1093/swr/31.2.83

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67. doi: 10.1006/ceps.1999.1020

Ryan, R. M., Rigby, C. S., & Przybylski, A. (2006). The motivational pull of video games: A self-determination theory approach. *Motivation and Emotion*, *30*(4), 347–363. doi: 10.1007/s11031-006-9051-8

Santiago-espada, Y., Myer, R. R., Latorella, K. A., & Comstock, J. R. (2011). *The multi-attribute task battery ii (matb-ii) software for human performance and workload research : A user s guide nasa/tm2011-217164* (No. July).

Schapire, R. E. (1990). The strength of weak learnability. *Machine Learning*, *5*(2), 197–227. doi: 10.1023/A:1022648800760

Schmidt, K. (2002). The problem withawareness': introductory remarks onawareness in cscw'. *Computer Supported Cooperative Work (CSCW)*, *11*(3), 285–298.

Seaborn, K., & Fels, D. I. (2015). Gamification in theory and action: A survey. *International Journal of Human-Computer Studies*, *74*, 14–31.

Sharples, S., & Houghton, R. J. (2017, 2). The field becomes the laboratory? the impact of the contextual digital footprint on the discipline of e/hf. *Ergonomics*, *60*(2), 270–283. doi: 10.1080/00140139.2016.1151946

Shawn Green, C., Sugarman, M. A., Medford, K., Klobusicky, E., & Bavelier, D. (2012). The effect of action video game experience on task-switching. *Computers in Human Behavior*, *28*(3), 984–994. doi: 10.1016/j.chb.2011.12.020

Skinner, B. F. (1938). The behavior of organisms: An experimental analysis. *The Psychological Record*, 486. doi: 10.1037/h0052216

Stevens, J. (1992). *Applied multivariate statistics for the social sciences, 2nd ed.* Hillsdale, NJ, US: Lawrence Erlbaum Associates, Inc.

Strobach, T., Frensch, P. A., & Schubert, T. (2012). Video game practice optimizes executive control skills in dual-task and task switching situations. *Acta Psychologica*, *140*(1), 13–24. doi: 10.1016/j.actpsy.2012.02.001

Sweller, J. (1988). Cognitive load during problem solving: Effects on learning. *Cognitive Science*, *12*(2), 257–285. doi: 10.1016/0364-0213(88)90023-7

Tattersall, A. J., & Foord, P. S. (1996). An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, *39*(5), 740–748.

Tsang, P. S., & Velazquez, V. L. (1996, 3). Diagnosticity and multidimensional subjective workload ratings. *Ergonomics*, *39*(3), 358–381. doi: 10.1080/00140139608964470

Tuv, E. (2006). Ensemble learning. *Feature Extraction, Foundations and Applications*, *207*, 187–204. doi: 10.1007/978-3-540-35488-8_8

van der Linden, D., Frese, M., & Meijman, T. F. (2003). Mental fatigue and the control of cognitive processes: Effects on perseveration and planning. *Acta Psychologica*, *113*(1), 45–65. doi: 10.1016/S0001-6918(02)00150-6

Van Essen, D. C. (2012, 8). Cortical cartography and caret software. *NeuroImage*, *62*(2), 757–764. doi: 10.1016/j.neuroimage.2011.10.077

Walker, G., & Strathie, A. (2016, 3). Big data and ergonomics methods: A new paradigm for tackling strategic transport safety risks. *Applied Ergonomics*, *53*, 298–311. doi: 10.1016/j.apergo.2015.09.008

Warm, J. S., Matthews, G., & Finomore Jr, V. S. (2017). Vigilance, workload, and stress. In *Performance under stress* (pp. 131–158). CRC Press.

Warm, J. S., Parasuraman, R., & Matthews, G. (2008). Vigilance requires hard mental work and is stressful. *Human factors*, *50*(3), 433–441.

Weiss, D. J., & Shanteau, J. (2003). Empirical assessment of expertise. *Human Factors*, *45*(1), 104–116.

Werner, S., & Thies, B. (2000). Is "change blindness" attenuated by domain-specific expertise? an expert-novices comparison of change detection in football images. *Visual Cognition*, *7*(1-3), 163–173. doi: 10.1080/135062800394748

Wettschereck, D., Aha, D. W., & Mohri, T. (1997). A review and empirical evaluation of feature weighting methods for a class of lazy learning algorithms. In D. W. Aha (Ed.), *Lazy learning* (pp. 273–314). Dordrecht: Springer Netherlands. doi: 10.1007/978-94-017-2053-$3_1$1

Wickens, C. D. (1986). The effects of control dynamics on performance.

Witteman, C. L., & Van Den Bercken, J. H. (2007). Intermediate effects in psychodiagnostic classification. *European Journal of Psychological Assessment*, *23*(1), 56–61. doi: 10.1027/1015-5759.23.1.56

Witteman, C. L. M., Weiss, D. J., & Metzmacher, M. (2012). Assessing diagnostic expertise of counselors using the cochran–weiss–shanteau (cws) index. *Journal of Counseling Development*, *90*(1), 30–34.

Wofford, J. C., Goodwin, V. L., & Premack, S. (1992). Meta-analysis of the antecedents of personal goal level and of the antecedents and consequences of goal commitment. *Journal of Management*, *18*(3), 595–615.

Zell, A. (1994). *Simulation neuronaler netze* (Vol. 1). Addison-Wesley Bonn.

Zhang, T., & Oles, F. J. (2001). Text categorization based on regularized linear classification methods. *Information Retrieval*, *4*(1), 5–31. doi: 10.1023/A:1011441423217

Zur, B. H., & Breznitz, S. J. (1981). The effect of time pressure on risky choice behavior. *Acta Psychologica*, *47*(2), 89–104.

APPENDICES

A. EXTRA FIGURES

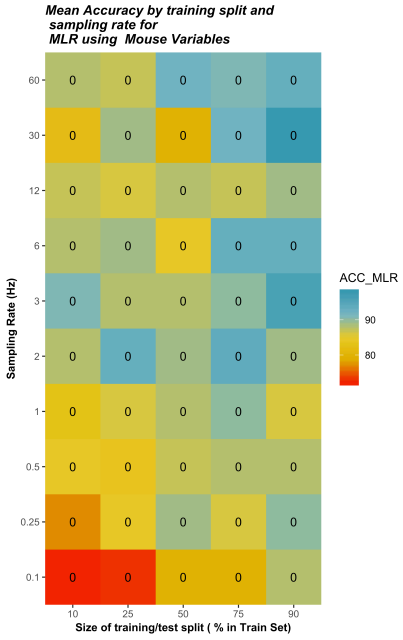
This chapter contains graphical representations of the accuracy of task difficulty and response delay models. There are additional graphical representations of the difference between the upper and lower confidence intervals. The subsections are organized by model type and are ordered by the type of variables used in the prediction.

A.1 Difficulty

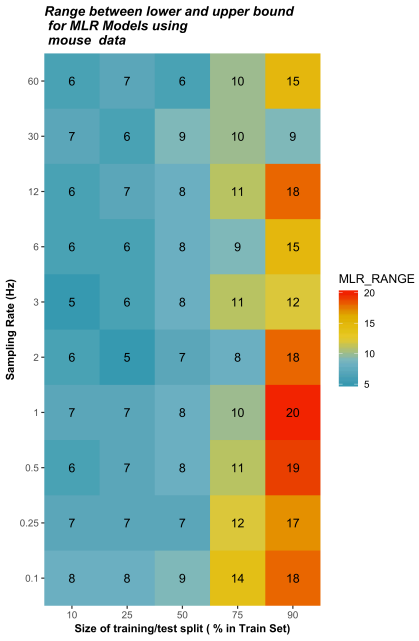
The following section contains graphical representations of the accuracy of task difficulty prediction models and the confidence intervals of the predictions. The section is structured by model class and further subdivided by predictor variables. Graphical representations of experience only prediction models are not included because none of the models were statistically significant.

A.1.1 Multinomial Logistic Regression

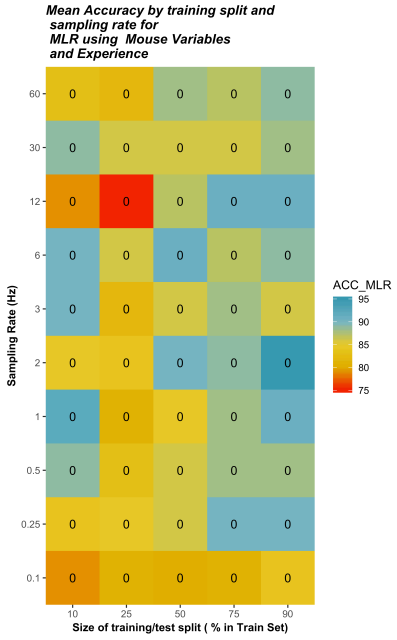
This section contains graphical representations of the accuracy of multinomial logistic regression models to predict task difficulty. The color scale reads from red to blue with redder shades being less accurate predictions.



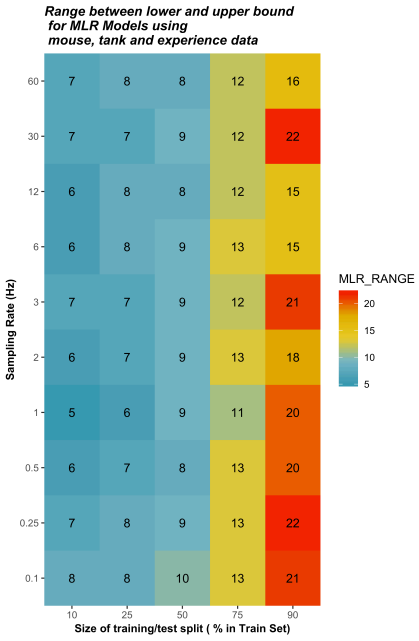
(a) Accuracy of Multinomial Logistic Regression Models using Mouse Data Only



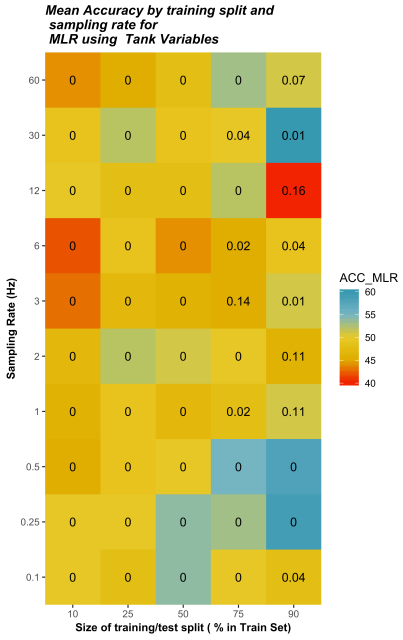
(b) Confidence interval of Multinomial Logistic Regression Models using Mouse Data Only



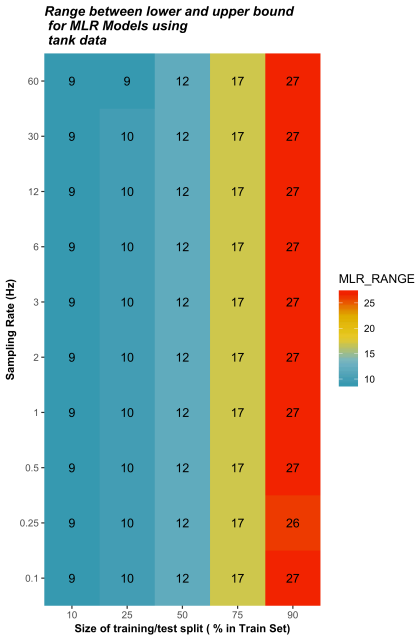
(c) Accuracy of Multinomial Logistic Regression Models using Mouse and Experience Data



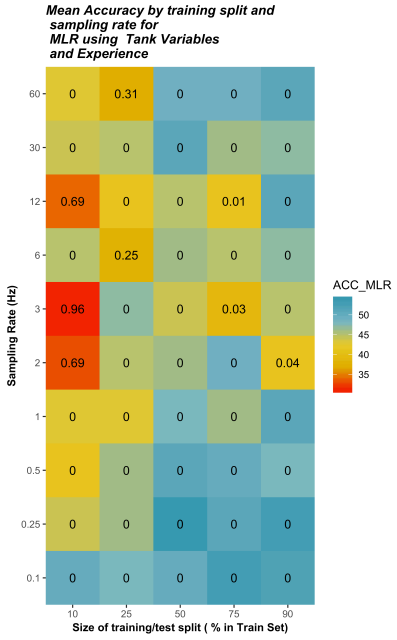
(d) Confidence interval of Multinomial Logistic Regression Models using Mouse and Experience Data



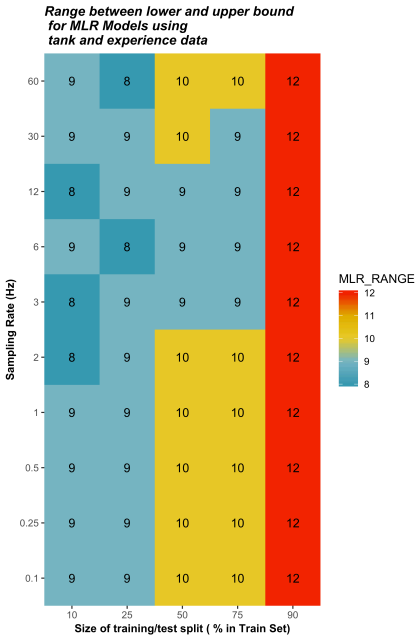
(e) Accuracy of Multinomial Logistic Regression Models using Tank Data



(f) Confidence interval of Multinomial Logistic Regression Models using Tank Data



(g) Accuracy of Multinomial Logistic Regression Models using Tank and Experience Data



(h) Confidence interval of Multinomial Logistic Regression Models using Tank and Experience Data

Multinomial Logistic Regression: Mouse Data

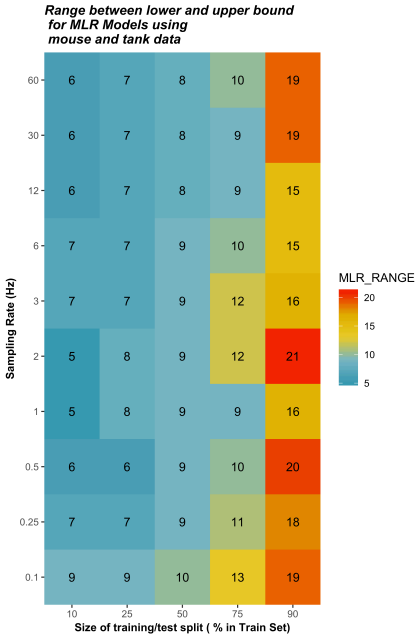
Multinomial Logistic Regression: Mouse Data and Experience

Multinomial Logistic Regression: Tank Data

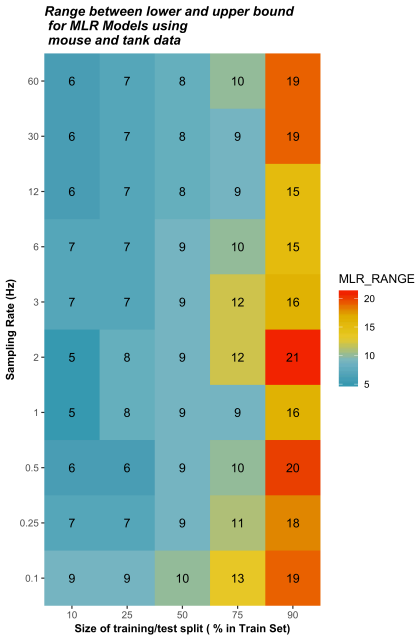
Multinomial Logistic Regression: Tank Data and Experience

Multinomial Logistic Regression: Mouse and Tank Data

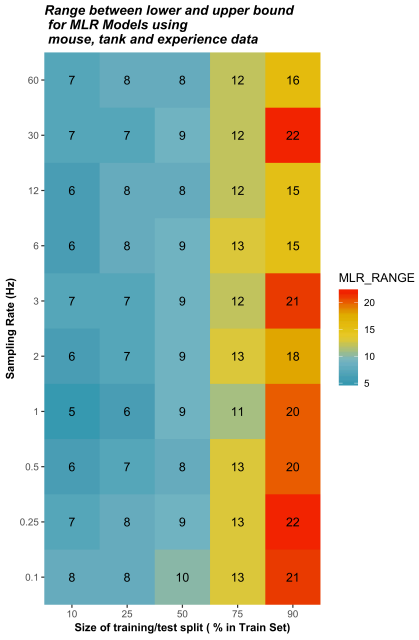
Multinomial Logistic Regression: Mouse, Tank, and Experience Data



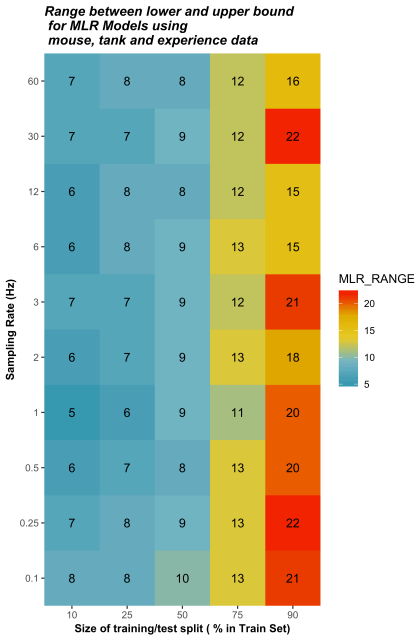
(i) Accuracy of Multinomial Logistic Regression Models using Mouse and Tank Data



(j) Confidence interval of Multinomial Logistic Regression Models using Mouse and Tank Data



(k) Accuracy of Multinomial Logistic Regression Models using Mouse, Tank and Experience Data



(l) Confidence Interval of Multinomial Logistic Regression Models using Mouse, Tank and Experience Data

A.1.2 Neural Networks

This section contains graphical representations of the accuracy of neural network models to predict task difficulty. The color scale reads from red to blue with redder shades being less accurate predictions.

Neural Network: Mouse Data

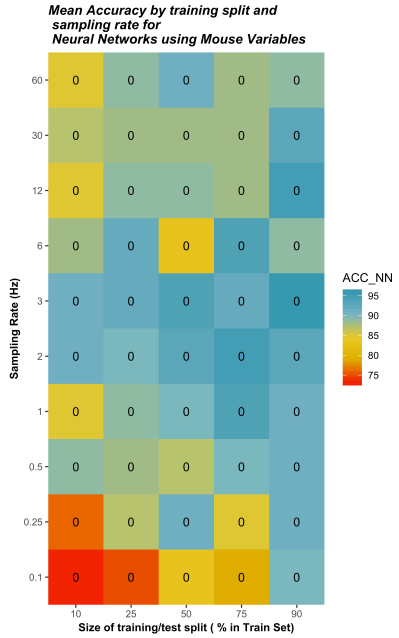
Neural Network: Mouse and Experience Data

Neural Network: Tank Data

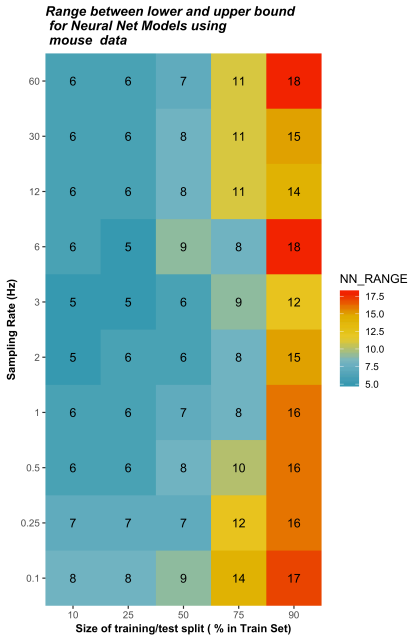
Neural Network: Tank and Experience Data

Neural Network: Mouse and Tank Data

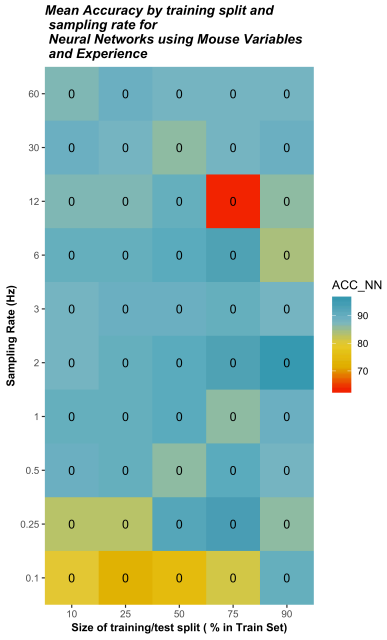
Neural Network: Mouse, Tank , and Experience Data



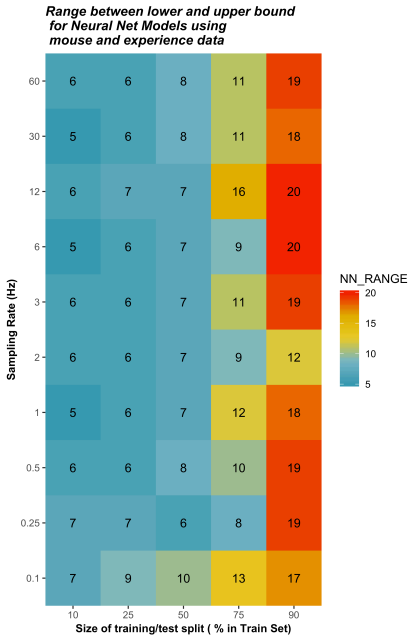
(m) Accuracy of Neural Network Models using Mouse Data



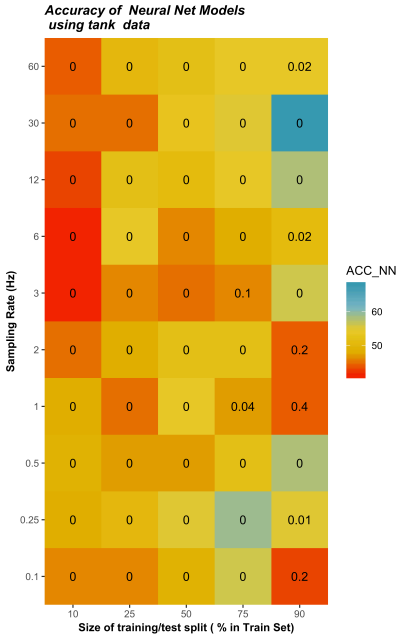
(n) Confidence Interval of Neural Network Models using Mouse Data



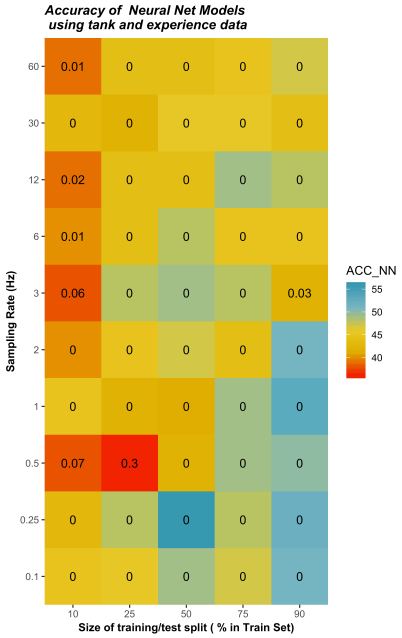
(o) Accuracy of Neural Network Models using Mouse and Experience Data



(p) Confidence Interval of Neural Network Models using Mouse and Experience Data



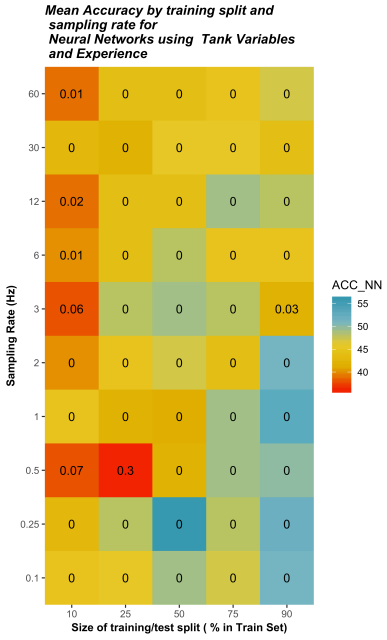
(q) Accuracy of Neural Network Models using Tank Data



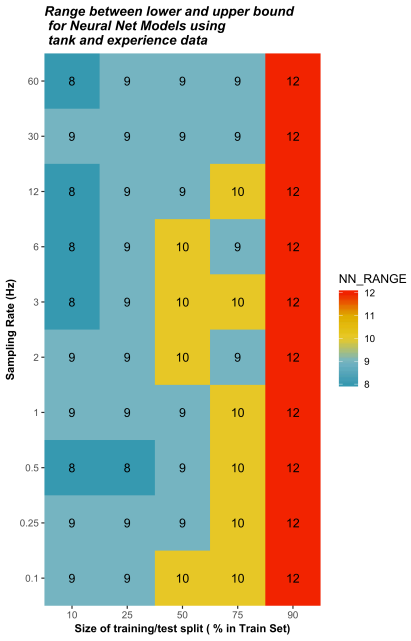
(r) Confidence Interval of Neural Network Models using Tank Data

A.1.3 Random Forests

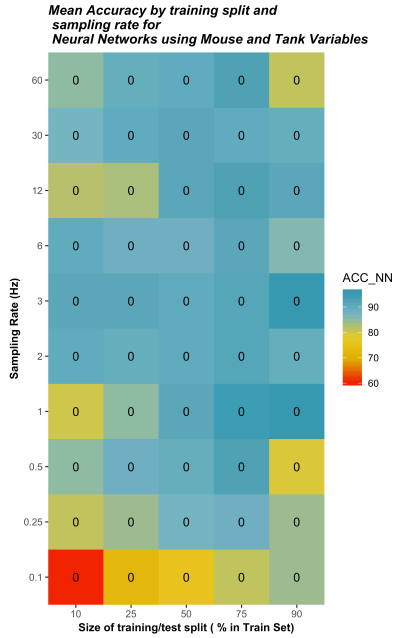
This section contains graphical representations of the accuracy of random forest models to predict task difficulty. The color scale reads from red to blue with redder shades being less accurate predictions.



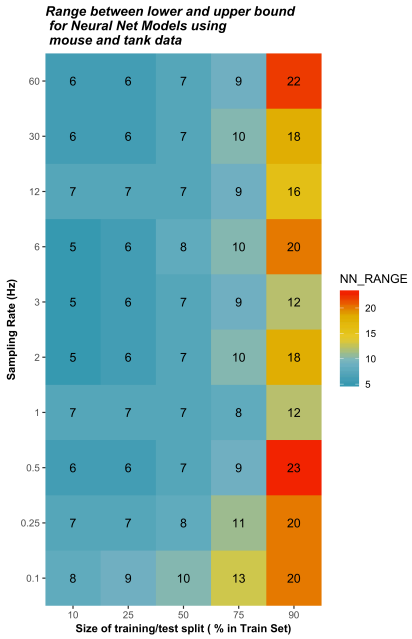
(s) Accuracy of Neural Network Models using Tank and Experience Data



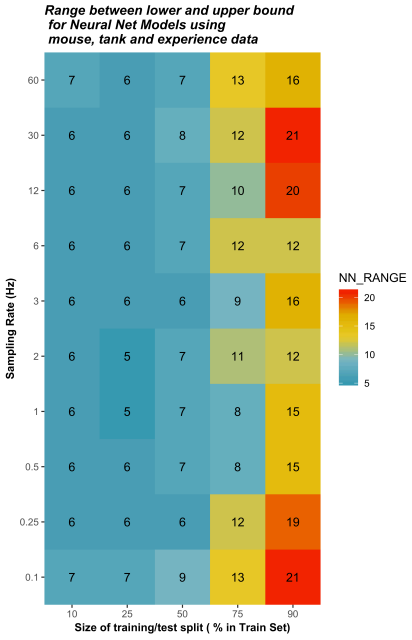
(t) Confidence Interval of Neural Network Models using Tank and Experience Data



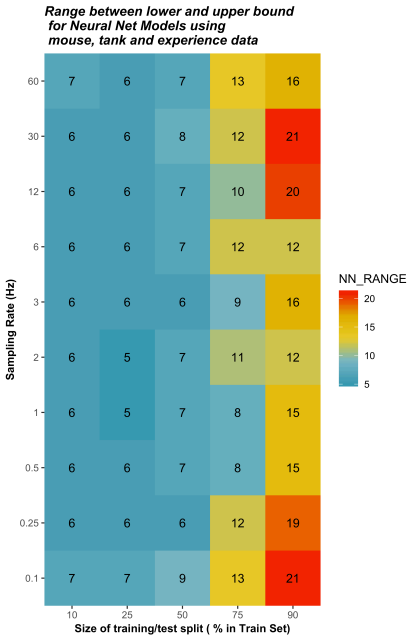
(u) Accuracy of Neural Network Models using Mouse and Tank Data



(v) Confidence Interval of Neural Network Models using Mouse and Tank Data

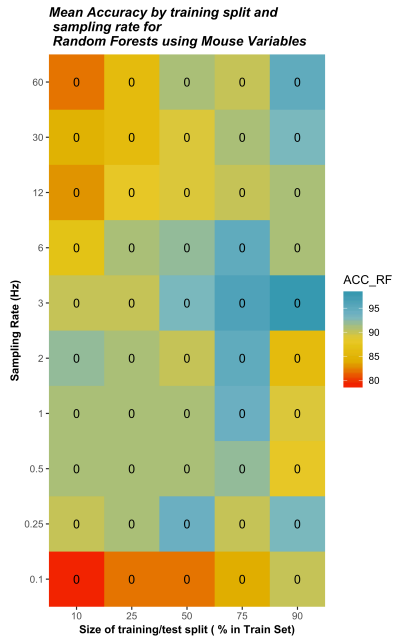


(w) Accuracy of Neural Network Models using Mouse, Tank and Experience Data

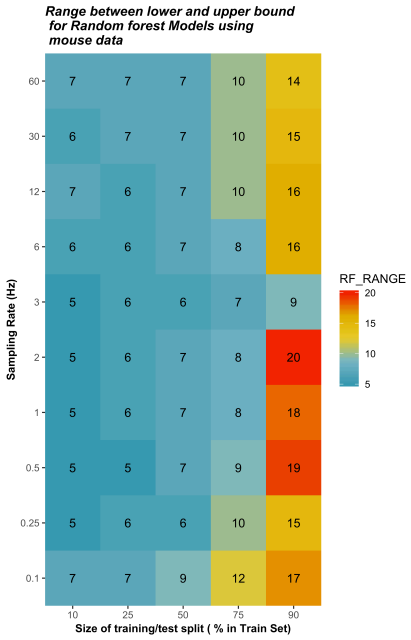


(x) Confidence Interval of Neural Network Models using Mouse, Tank and Experience Data

Random Forest: Mouse Data

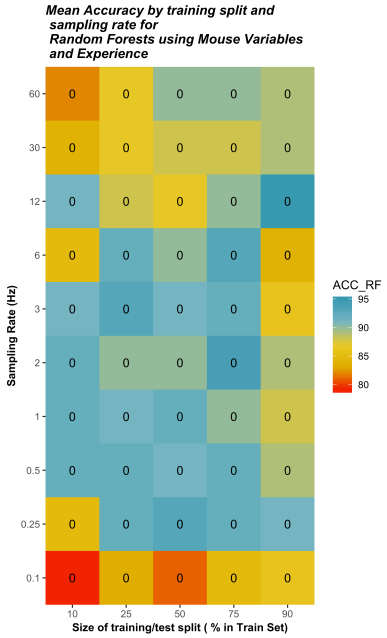


(y) Accuracy of Random Forest Models using Mouse Data

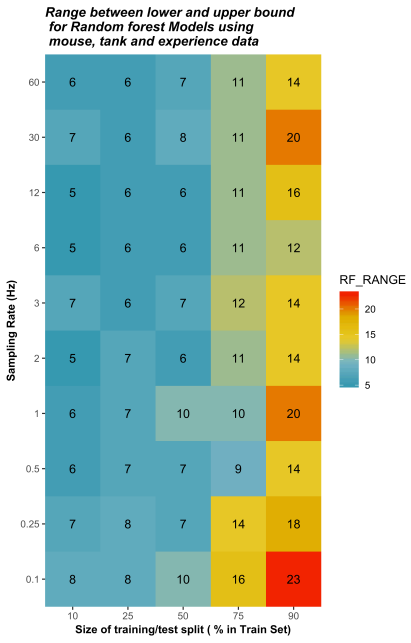


(z) Confidence Interval of Random Forest Models using Mouse Data

Random Forest: Mouse and Experience Data

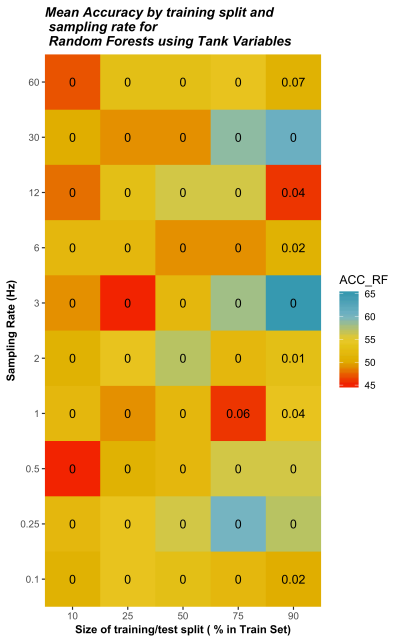


() Accuracy of Random Forest Models using Mouse and Experience Data

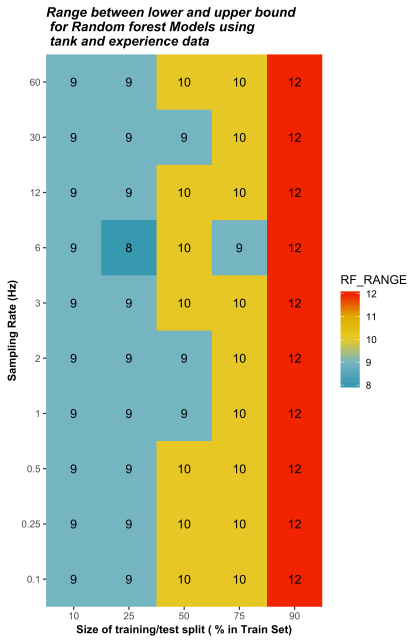


() Confidence Interval of Random Forest Models using Mouse and Experience Data

Random Forest: Tank Data

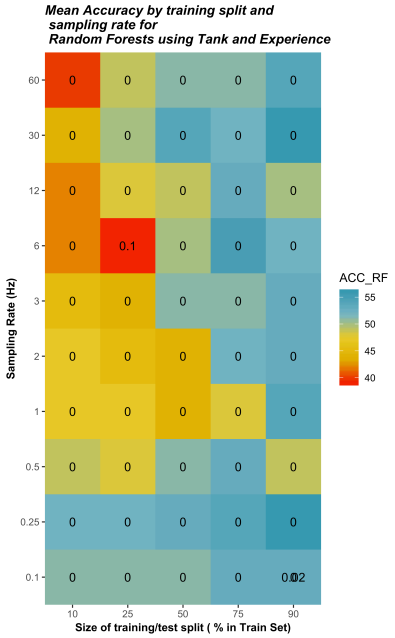


() Accuracy of Random Forest Models using Tank Data

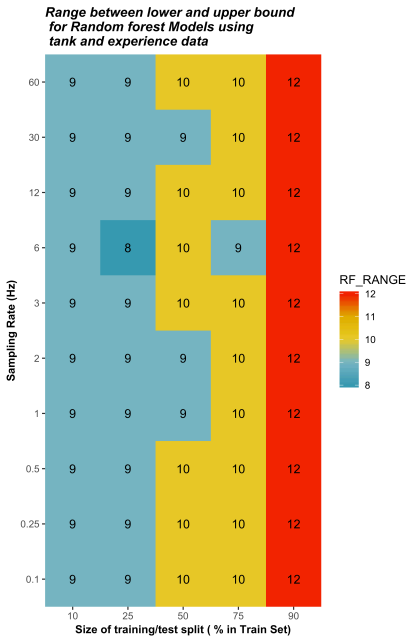


() Confidence Interval of Random Forest Models using Tank Data

Random Forest: Tank and Experience Data

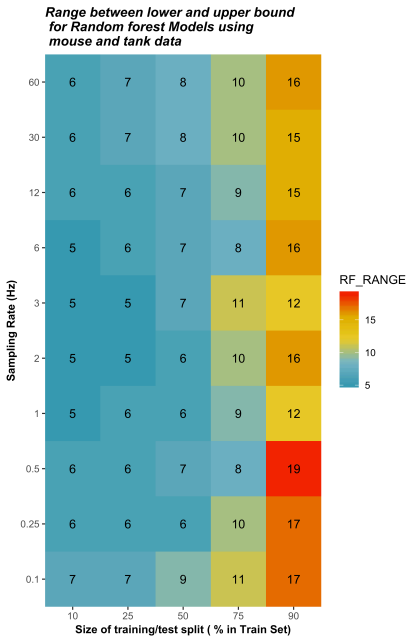


() Accuracy of Random Forest Models using Tank and Experience Data

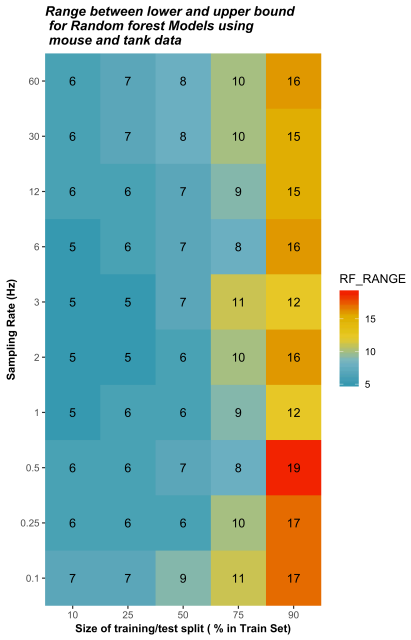


() Confidence Interval of Random Forest Models using Tank and Experience Data

Random Forest: Mouse and Tank Data

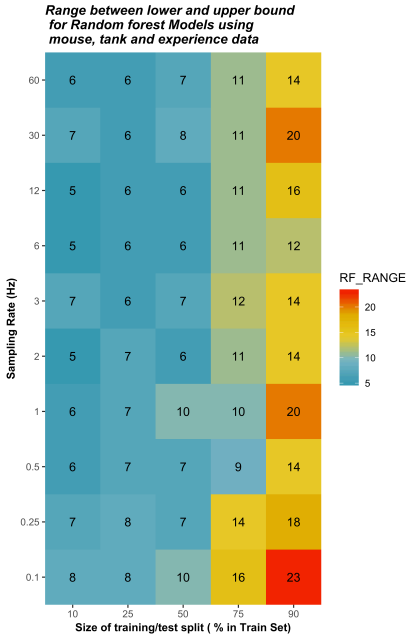


() Accuracy of Random Forest Models using Mouse and Tank Data

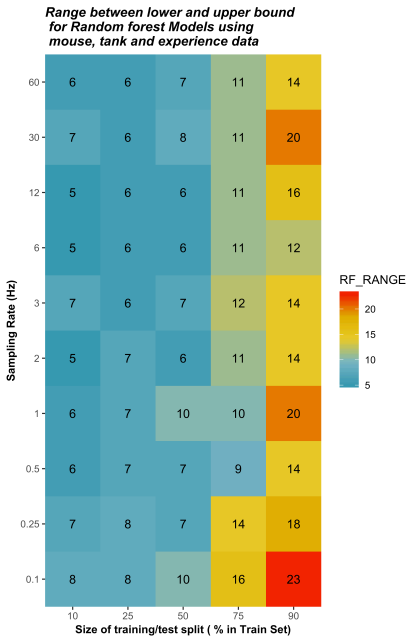


() Confidence Interval of Random Forest Models using Mouse and Tank Data

Random Forest: Mouse , Tank and Experience Data



() Accuracy of Random Forest Models using Mouse, Tank and Experience Data



() Confidence Interval of Random Forest Models using Mouse, Tank and Experience Data

A.2 Delay

The section is structured by model class and further subdivided by predictor variables. Graphical representations of experience only prediction models are not included because none of the models were statistically significant.

A.2.1 General Linear Model

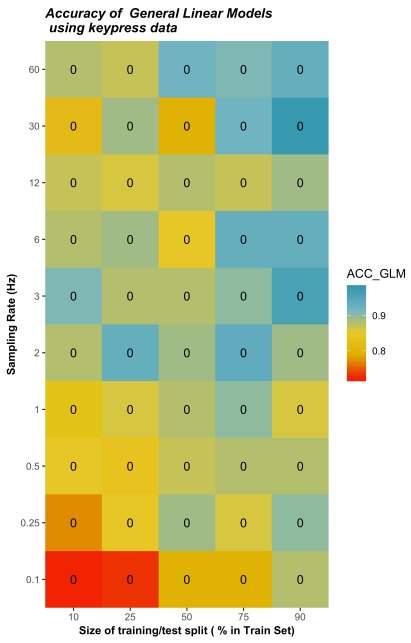
This section contains graphical representations of the accuracy of general linear models to predict response delay. The color scale reads from red to blue with redder shades being less accurate predictions.

General Linear Model : Keypress Data

General Linear Model : Keypress and Experience Data

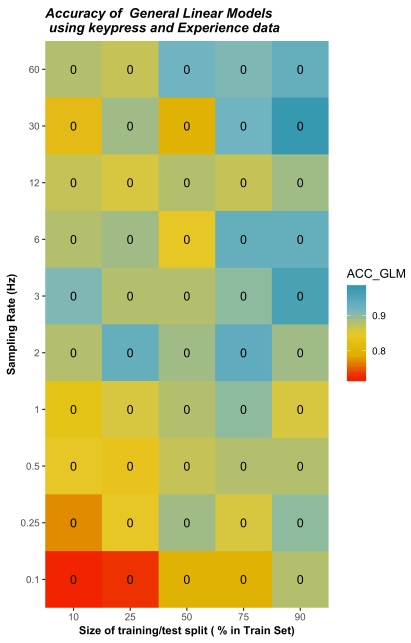
General Linear Model : Keypress and Interaction Time Data

General Linear Model : Keypress, Interaction Time and Experience Data

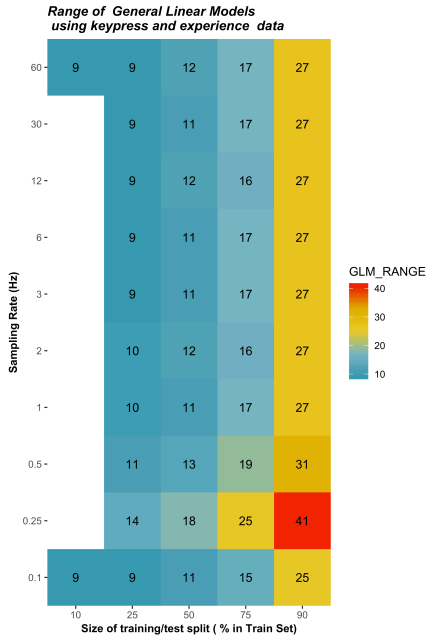


() Accuracy of General Linear Models using Keypress Data

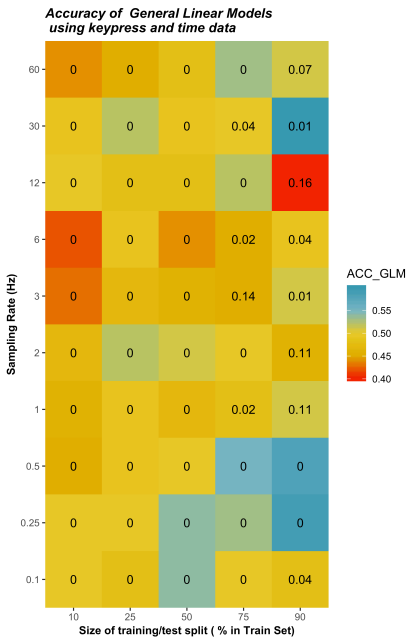
() Confidence Interval of General Linear Models using Keypress Data



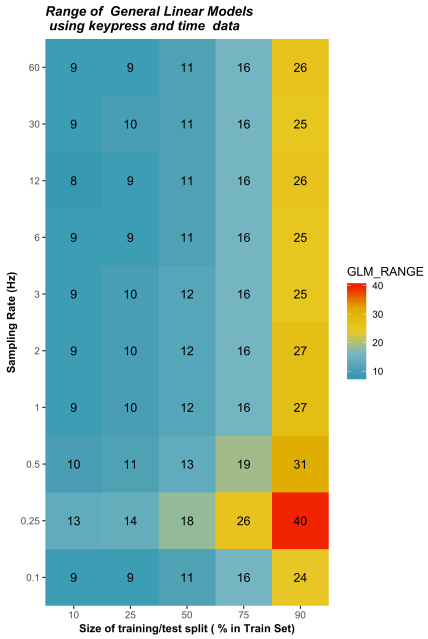
() Accuracy of General Linear Models using Keypress and Experience Data



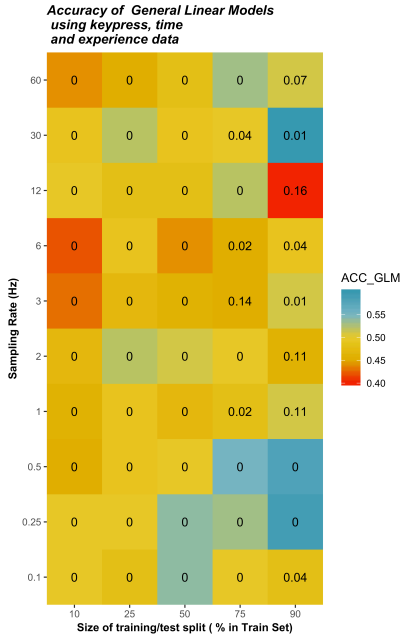
() Confidence Interval of General Linear Models using Keypress and Experience Data



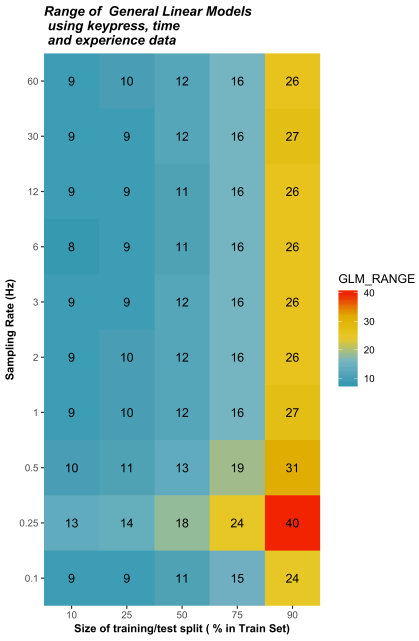
() Accuracy of General Linear Models using Keypress and Interaction Time Data



() Confidence Interval of General Linear Models using Keypress and Time Data



() Accuracy of General Linear Models using Keypress , Interaction Time and Experience Data



() Confidence Interval of General Linear Models using Keypress ,Interaction Time and Experience Data

A.2.2 Neural Networks

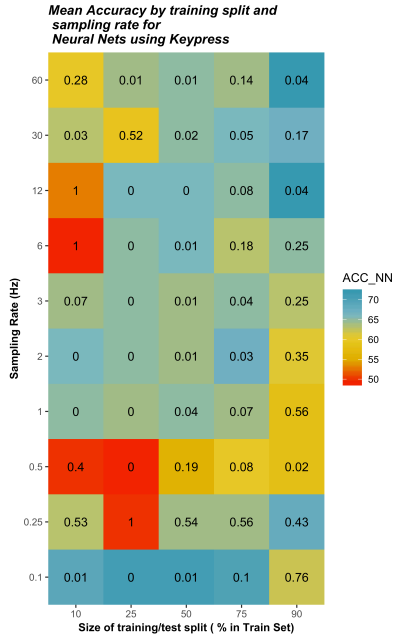
This section contains graphical representations of the accuracy of neural network models to predict response delay. The color scale reads from red to blue with redder shades being less accurate predictions.

Neural Network : Keypress Data

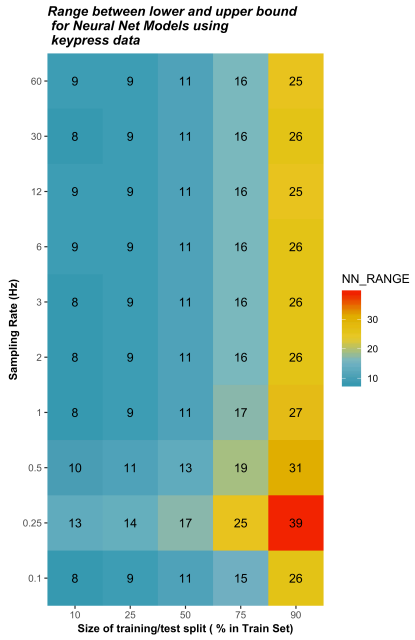
Neural Network : Keypress and Experience Data

Neural Network : Keypress and Interaction Time Data

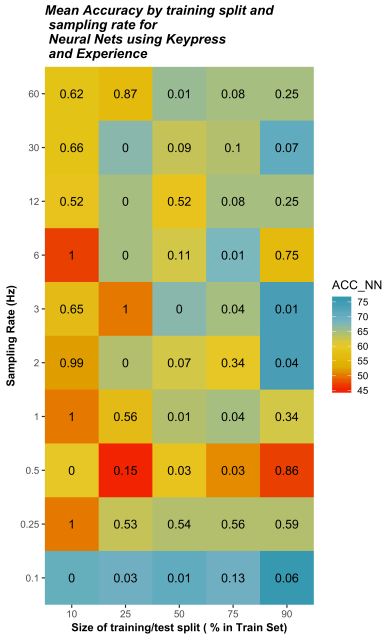
Neural Network : Keypress, Interaction Time and Experience Data



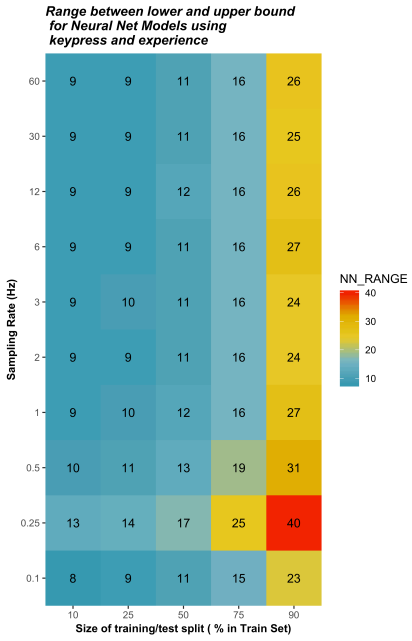
() Accuracy of Neural Network Models using Keypress Data



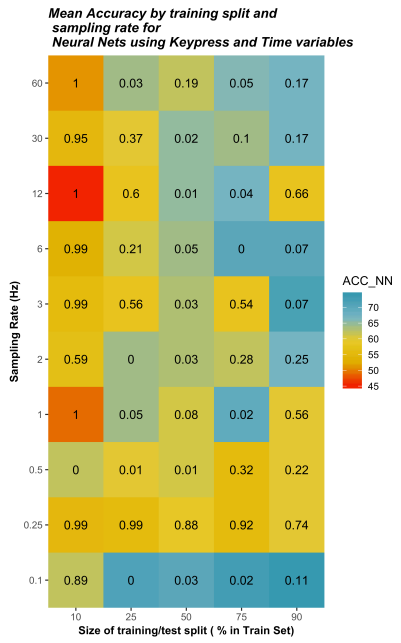
() Confidence Interval of Neural Network Models using Keypress Data



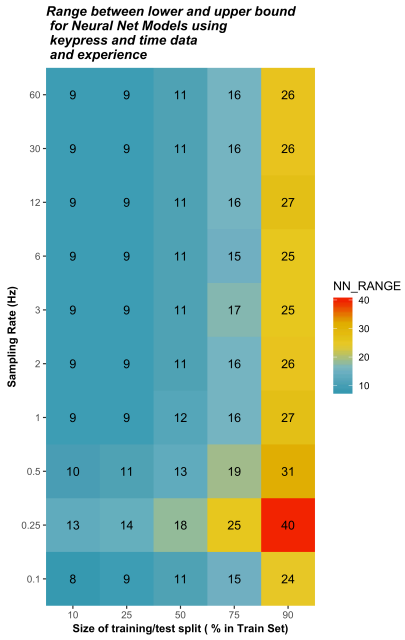
() Accuracy of Neural Network Models using Keypress and Experience Data



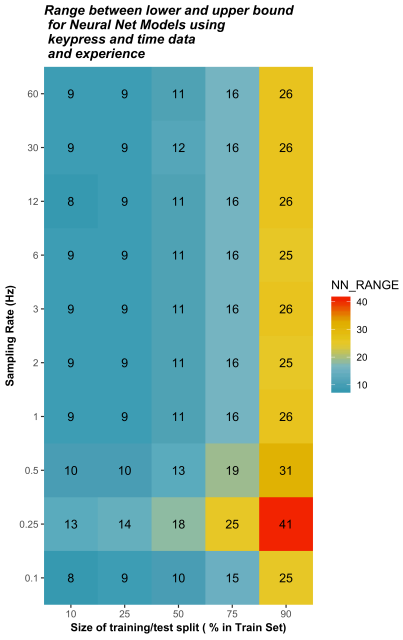
() Confidence Interval of Neural Network Models using Keypress and Experience Data



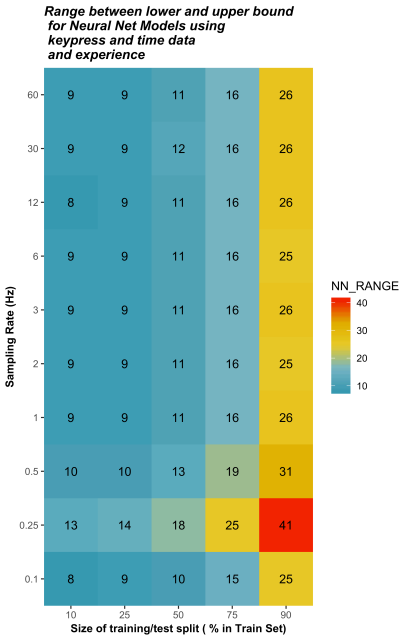
() Accuracy of Neural Network Models using Keypress and Interaction Time Data



() Confidence Interval of Neural Network Models using Keypress and Interaction Time Data



() Accuracy of Neural Network Models using Keypress, Interaction Time , Experience Data



() Confidence Interval of Neural Network Models using Keypress, Interaction Time , Experience Data

A.2.3 Random Forests

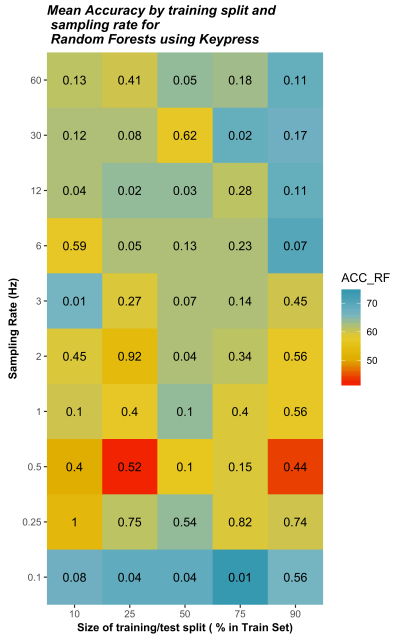
This section contains graphical representations of the accuracy of random forest models to predict response delay. The color scale reads from red to blue with redder shades being less accurate predictions.

Random Forest : Keypress Data

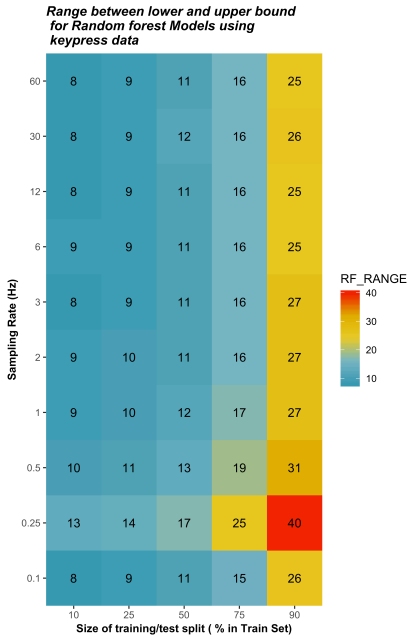
Random Forest : Keypress and Experience Data

Random Forest : Keypress and Interaction Time Data

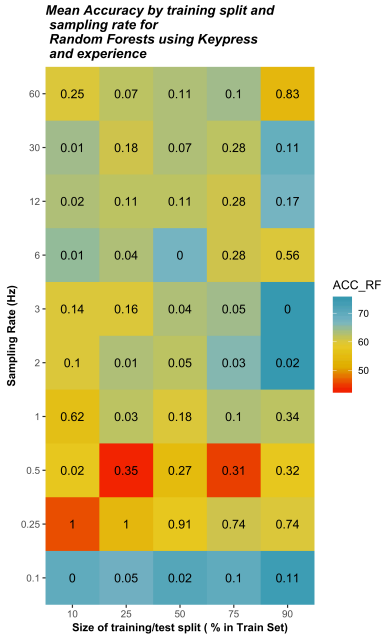
Random Forest : Keypress, Interaction Time Data



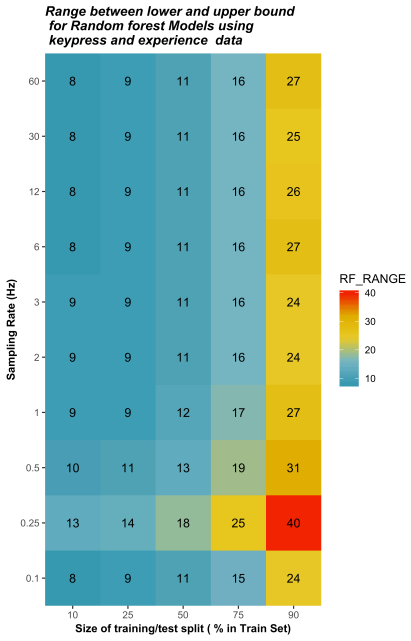
() Accuracy of Random Forests Models using Keypress Data



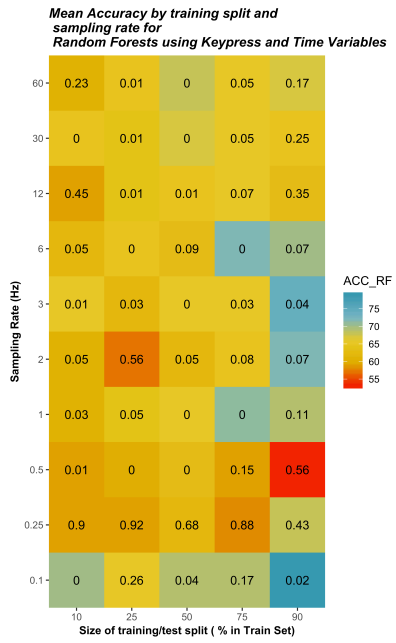
() Confidence Interval of Random Forests Models using Keypress Data



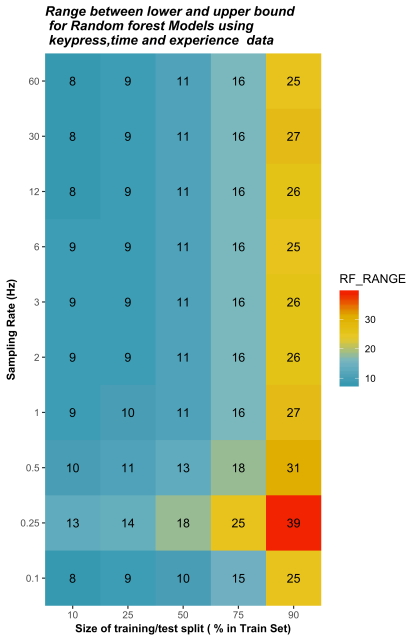
() Accuracy of Random Forests Models using Keypress and Experience Data



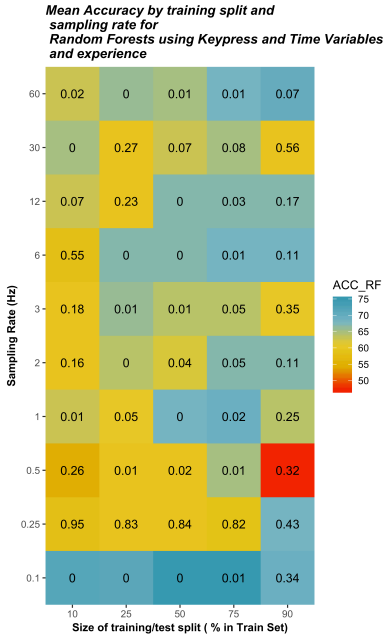
() Confidence Interval of Random Forests Models using Keypress and Experience Data



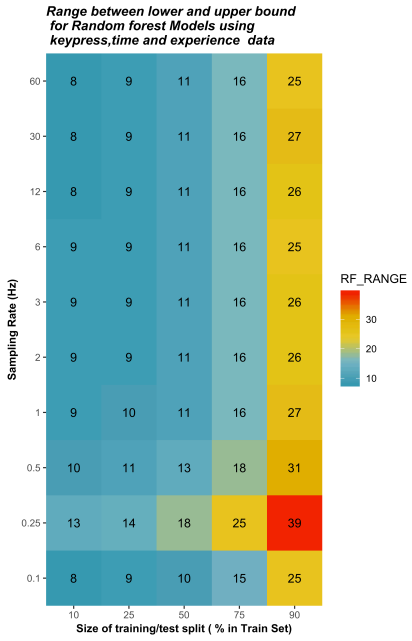
() Accuracy of Random Forests Models using Keypress and Interaction Time Data



() Confidence Interval of Random Forests Models using Keypress and Interaction Time Data



() Accuracy of Random Forests Models using Keypress, Interaction Time and Experience Data



() Confidence Interval of Random Forests Models using Keypress, Interaction Time and Experience Data

B. EXTRA TABLES

This chapter contains tables showing the Accuracy of task difficulty and response delay models. There is information on the difference between the upper and lower confidence intervals. The subsections are organized by model type and are ordered by the type of variables used in the prediction.

How to Read the Tables The tables in this section are organized in the following maner. The first column is the percent of total data used to train a model. The next column is the sampling rate in hertz. Four values are presented: the lower bound of the confidence interval, the mean predictive Accuracy , the upper bound of the confidence interval and the p-value of the model. P-values less than 0.001 are represented as 0.00.

B.1 Difficulty Predictions

This section contains tables showing the Accuracy of difficulty prediction models that have been further subdivided by predictor variables

B.1.1 Multi-nomial Logistic Regression

Multi-nomial Logistic Regression: Experience

Table B.1.: Multinomial Logistic Regression using Experience Only

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
10%	60Hz	26%	30%	34%	0.98
	30Hz	29%	33%	37%	0.69
	12Hz	29%	33%	38%	0.69
	6Hz	27%	31%	35%	0.98
	3Hz	29%	33%	37%	0.83
	2Hz	31%	35%	39%	0.41
	1Hz	28%	32%	37%	0.78
	.5Hz	30%	34%	39%	0.62
	.25Hz	27%	31%	35%	0.94
	.1Hz	27%	31%	35%	0.99
25%	60Hz	29%	33%	38%	0.74
	30Hz	32%	36%	41%	0.48
	12Hz	28%	32%	37%	0.88
	6Hz	29%	33%	38%	0.89
	3Hz	29%	33%	37%	0.93
	2Hz	30%	35%	40%	0.40
	1Hz	28%	33%	37%	0.85
	.5Hz	29%	34%	38%	0.60
	.25Hz	28%	32%	37%	0.95
	.1Hz	28%	32%	37%	0.80
50%	60Hz	25%	30%	35%	0.98
	30Hz	31%	37%	43%	0.52

continued on next page

Table B.1.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	25%	31%	36%	0.95
	6Hz	26%	32%	37%	0.94
	3Hz	31%	36%	42%	0.78
	2Hz	24%	29%	35%	0.99
	1Hz	33%	38%	44%	0.57
	.5Hz	34%	39%	45%	0.38
	.25Hz	25%	30%	36%	0.95
	.1Hz	27%	33%	38%	0.79
75%	60Hz	25%	33%	41%	0.67
	30Hz	26%	34%	42%	0.60
	12Hz	25%	33%	41%	0.90
	6Hz	23%	31%	39%	0.87
	3Hz	25%	32%	40%	0.93
	2Hz	24%	31%	39%	0.95
	1Hz	26%	33%	42%	0.60
	.5Hz	27%	35%	43%	0.60
	.25Hz	22%	29%	37%	0.97
	.1Hz	25%	33%	41%	0.73
90%	60Hz	18%	30%	43%	0.96
	30Hz	17%	28%	42%	1.00
	12Hz	17%	28%	42%	0.94
	6Hz	36%	49%	63%	0.55
	3Hz	23%	35%	49%	0.83
	2Hz	36%	49%	63%	0.55
	1Hz	16%	26%	40%	0.99

continued on next page

Table B.1.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	18%	30%	43%	0.98
	.25Hz	26%	38%	52%	0.55
	.1Hz	14%	24%	37%	1.00

Multi-nomial Logistic Regression: Mouse Data

Table B.2.: Multinomial Logistic Regression using Mouse Data Only

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
10%	60Hz	85%	88%	90%	$p \leq 0.001$
	30Hz	90%	82%	85%	$p \leq 0.001$
	12Hz	91%	87%	90%	$p \leq 0.001$
	6Hz	93%	88%	91%	$p \leq 0.001$
	3Hz	91%	91%	93%	$p \leq 0.001$
	2Hz	86%	88%	91%	$p \leq 0.001$
	1Hz	88%	83%	86%	$p \leq 0.001$
	.5Hz	81%	85%	88%	$p \leq 0.001$
	.25Hz	75%	77%	81%	$p \leq 0.001$
	.1Hz	90%	72%	75%	$p \leq 0.001$
25%	60Hz	92%	87%	90%	$p \leq 0.001$
	30Hz	89%	89%	92%	$p \leq 0.001$
	12Hz	92%	86%	89%	$p \leq 0.001$
	6Hz	90%	89%	92%	$p \leq 0.001$
	3Hz	95%	88%	90%	$p \leq 0.001$
	2Hz	89%	93%	95%	$p \leq 0.001$
	1Hz	87%	86%	89%	$p \leq 0.001$
	.5Hz	88%	84%	87%	$p \leq 0.001$
	.25Hz	77%	85%	88%	$p \leq 0.001$
	.1Hz	95%	73%	77%	$p \leq 0.001$
50%	60Hz	84%	92%	95%	$p \leq 0.001$
	30Hz	91%	80%	84%	$p \leq 0.001$

continued on next page

Table B.2.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	89%	88%	91%	$p \leq 0.001$
	6Hz	92%	85%	89%	$p \leq 0.001$
	3Hz	92%	88%	92%	$p \leq 0.001$
	2Hz	92%	89%	92%	$p \leq 0.001$
	1Hz	91%	88%	92%	$p \leq 0.001$
	.5Hz	92%	87%	91%	$p \leq 0.001$
	.25Hz	84%	89%	92%	$p \leq 0.001$
	.1Hz	95%	80%	84%	$p \leq 0.001$
75%	60Hz	96%	91%	95%	$p \leq 0.001$
	30Hz	92%	92%	96%	$p \leq 0.001$
	12Hz	96%	87%	92%	$p \leq 0.001$
	6Hz	94%	93%	96%	$p \leq 0.001$
	3Hz	97%	90%	94%	$p \leq 0.001$
	2Hz	94%	94%	97%	$p \leq 0.001$
	1Hz	93%	90%	94%	$p \leq 0.001$
	.5Hz	91%	88%	93%	$p \leq 0.001$
	.25Hz	86%	86%	91%	$p \leq 0.001$
	.1Hz	98%	80%	86%	$p \leq 0.001$
90%	60Hz	100%	93%	98%	$p \leq 0.001$
	30Hz	96%	98%	100%	$p \leq 0.001$
	12Hz	98%	89%	96%	$p \leq 0.001$
	6Hz	100%	93%	98%	$p \leq 0.001$
	3Hz	96%	96%	100%	$p \leq 0.001$
	2Hz	94%	89%	96%	$p \leq 0.001$
	1Hz	95%	86%	94%	$p \leq 0.001$

continued on next page

Table B.2.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	96%	88%	95%	$p \leq 0.001$
	.25Hz	95%	90%	96%	$p \leq 0.001$
	.1Hz	80%	88%	95%	0.00

Multi-nomial Logistic Regression: Mouse Data and Experience

Table B.3.: Multinomial Logistic Regression Mouse and Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
10%	60Hz	92%	83%	86%	$p \leq 0.001$
	30Hz	82%	89%	92%	$p \leq 0.001$
	12Hz	92%	79%	82%	$p \leq 0.001$
	6Hz	92%	90%	92%	$p \leq 0.001$
	3Hz	88%	90%	92%	$p \leq 0.001$
	2Hz	94%	85%	88%	$p \leq 0.001$
	1Hz	92%	92%	94%	$p \leq 0.001$
	.5Hz	87%	89%	92%	$p \leq 0.001$
	.25Hz	82%	84%	87%	$p \leq 0.001$
	.1Hz	85%	79%	82%	$p \leq 0.001$
25%	60Hz	89%	82%	85%	$p \leq 0.001$
	30Hz	79%	86%	89%	$p \leq 0.001$
	12Hz	89%	75%	79%	$p \leq 0.001$
	6Hz	85%	86%	89%	$p \leq 0.001$
	3Hz	87%	82%	85%	$p \leq 0.001$
	2Hz	85%	84%	87%	$p \leq 0.001$
	1Hz	86%	81%	85%	$p \leq 0.001$
	.5Hz	88%	83%	86%	$p \leq 0.001$
	.25Hz	84%	85%	88%	$p \leq 0.001$
	.1Hz	92%	81%	84%	$p \leq 0.001$
50%	60Hz	90%	88%	92%	$p \leq 0.001$
	30Hz	91%	86%	90%	$p \leq 0.001$

continued on next page

Table B.3.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	94%	87%	91%	$p \leq 0.001$
	6Hz	90%	91%	94%	$p \leq 0.001$
	3Hz	94%	86%	90%	$p \leq 0.001$
	2Hz	89%	90%	94%	$p \leq 0.001$
	1Hz	90%	85%	89%	$p \leq 0.001$
	.5Hz	90%	86%	90%	$p \leq 0.001$
	.25Hz	84%	86%	90%	$p \leq 0.001$
	.1Hz	92%	80%	84%	$p \leq 0.001$
75%	60Hz	91%	87%	92%	$p \leq 0.001$
	30Hz	95%	86%	91%	$p \leq 0.001$
	12Hz	92%	91%	95%	$p \leq 0.001$
	6Hz	93%	87%	92%	$p \leq 0.001$
	3Hz	94%	88%	93%	$p \leq 0.001$
	2Hz	93%	89%	94%	$p \leq 0.001$
	1Hz	93%	88%	93%	$p \leq 0.001$
	.5Hz	95%	88%	93%	$p \leq 0.001$
	.25Hz	87%	90%	95%	$p \leq 0.001$
	.1Hz	96%	81%	87%	$p \leq 0.001$
90%	60Hz	95%	89%	96%	$p \leq 0.001$
	30Hz	97%	88%	95%	$p \leq 0.001$
	12Hz	96%	91%	97%	$p \leq 0.001$
	6Hz	94%	89%	96%	$p \leq 0.001$
	3Hz	99%	86%	94%	$p \leq 0.001$
	2Hz	97%	95%	99%	$p \leq 0.001$
	1Hz	95%	91%	97%	$p \leq 0.001$

continued on next page

Table B.3.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	96%	88%	95%	$p \leq 0.001$
	.25Hz	93%	90%	96%	$p \leq 0.001$
	.1Hz	0%	84%	93%	$p \leq 0.001$

Multi-nomial Logistic Regression: Tank Data

Table B.4.: Multinomial Logistic Regression using Tank Variables

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
10%	60Hz	53%	44%	49%	$p \leq 0.001$
	30Hz	55%	49%	53%	$p \leq 0.001$
	12Hz	47%	50%	55%	$p \leq 0.001$
	6Hz	47%	42%	47%	$p \leq 0.001$
	3Hz	52%	43%	47%	$p \leq 0.001$
	2Hz	50%	47%	52%	$p \leq 0.001$
	1Hz	49%	46%	50%	$p \leq 0.001$
	.5Hz	55%	45%	49%	$p \leq 0.001$
	.25Hz	54%	50%	55%	$p \leq 0.001$
	.1Hz	50%	50%	54%	$p \leq 0.001$
25%	60Hz	56%	45%	50%	$p \leq 0.001$
	30Hz	52%	52%	56%	$p \leq 0.001$
	12Hz	54%	48%	52%	$p \leq 0.001$
	6Hz	51%	49%	54%	$p \leq 0.001$
	3Hz	57%	47%	51%	$p \leq 0.001$
	2Hz	54%	52%	57%	$p \leq 0.001$
	1Hz	54%	49%	54%	$p \leq 0.001$
	.5Hz	55%	49%	54%	$p \leq 0.001$
	.25Hz	53%	50%	55%	$p \leq 0.001$
	.1Hz	54%	48%	53%	$p \leq 0.001$
50%	60Hz	55%	48%	54%	$p \leq 0.001$
	30Hz	54%	49%	55%	$p \leq 0.001$

continued on next page

Table B.4.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	50%	48%	54%	$p \leq 0.001$
	6Hz	52%	44%	50%	$p \leq 0.001$
	3Hz	57%	46%	52%	$p \leq 0.001$
	2Hz	53%	51%	57%	$p \leq 0.001$
	1Hz	56%	47%	53%	$p \leq 0.001$
	.5Hz	60%	50%	56%	$p \leq 0.001$
	.25Hz	60%	54%	60%	$p \leq 0.001$
	.1Hz	61%	54%	60%	$p \leq 0.001$
75%	60Hz	58%	53%	61%	$p \leq 0.001$
	30Hz	61%	50%	58%	0.04
	12Hz	53%	52%	61%	$p \leq 0.001$
	6Hz	54%	45%	53%	0.02
	3Hz	59%	45%	54%	0.14
	2Hz	57%	50%	59%	$p \leq 0.001$
	1Hz	64%	48%	57%	0.02
	.5Hz	61%	55%	64%	$p \leq 0.001$
	.25Hz	58%	53%	61%	$p \leq 0.001$
	.1Hz	64%	50%	58%	$p \leq 0.001$
90%	60Hz	72%	51%	64%	0.07
	30Hz	54%	60%	72%	0.01
	12Hz	63%	40%	54%	0.16
	6Hz	64%	49%	63%	0.04
	3Hz	59%	51%	64%	0.01
	2Hz	64%	46%	59%	0.11
	1Hz	71%	51%	64%	0.11

continued on next page

Table B.4.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	71%	58%	71%	$p \leq 0.001$
	.25Hz	62%	59%	71%	$p \leq 0.001$
	.1Hz	0%	48%	62%	0.04

Multi-nomial Logistic Regression: Tank Data and Experience

Table B.5.: Multinomial Logistic Regression using Tank Variables and Experience

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
10%	60Hz	48%	43%	47%	$p \leq 0.001$
	30Hz	38%	44%	48%	$p \leq 0.001$
	12Hz	50%	34%	38%	0.69
	6Hz	35%	45%	50%	$p \leq 0.001$
	3Hz	38%	31%	35%	0.96
	2Hz	47%	33%	38%	0.69
	1Hz	45%	43%	47%	$p \leq 0.001$
	.5Hz	48%	41%	45%	$p \leq 0.001$
	.25Hz	54%	44%	48%	$p \leq 0.001$
	.1Hz	41%	50%	54%	$p \leq 0.001$
25%	60Hz	50%	37%	41%	0.31
	30Hz	45%	45%	50%	$p \leq 0.001$
	12Hz	42%	41%	45%	$p \leq 0.001$
	6Hz	51%	38%	42%	0.25
	3Hz	49%	47%	51%	$p \leq 0.001$
	2Hz	47%	45%	49%	$p \leq 0.001$
	1Hz	50%	43%	47%	$p \leq 0.001$
	.5Hz	50%	46%	50%	$p \leq 0.001$
	.25Hz	52%	46%	50%	$p \leq 0.001$
	.1Hz	54%	48%	52%	$p \leq 0.001$
50%	60Hz	56%	49%	54%	$p \leq 0.001$
	30Hz	50%	51%	56%	$p \leq 0.001$

continued on next page

Table B.5.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	51%	45%	50%	$p \leq 0.001$
	6Hz	49%	46%	51%	$p \leq 0.001$
	3Hz	51%	44%	49%	$p \leq 0.001$
	2Hz	53%	46%	51%	$p \leq 0.001$
	1Hz	56%	48%	53%	$p \leq 0.001$
	.5Hz	59%	51%	56%	$p \leq 0.001$
	.25Hz	55%	54%	59%	$p \leq 0.001$
	.1Hz	54%	50%	55%	$p \leq 0.001$
75%	60Hz	50%	49%	54%	$p \leq 0.001$
	30Hz	46%	46%	50%	$p \leq 0.001$
	12Hz	50%	41%	46%	0.01
	6Hz	44%	45%	50%	$p \leq 0.001$
	3Hz	54%	39%	44%	0.03
	2Hz	51%	49%	54%	$p \leq 0.001$
	1Hz	55%	46%	51%	$p \leq 0.001$
	.5Hz	56%	50%	55%	$p \leq 0.001$
	.25Hz	58%	51%	56%	$p \leq 0.001$
	.1Hz	57%	53%	58%	$p \leq 0.001$
90%	60Hz	53%	51%	57%	$p \leq 0.001$
	30Hz	57%	47%	53%	$p \leq 0.001$
	12Hz	53%	51%	57%	$p \leq 0.001$
	6Hz	50%	47%	53%	$p \leq 0.001$
	3Hz	47%	45%	50%	$p \leq 0.001$
	2Hz	57%	41%	47%	0.04
	1Hz	54%	51%	57%	$p \leq 0.001$

continued on next page

Table B.5.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	59%	48%	54%	$p \leq 0.001$
	.25Hz	58%	53%	59%	$p \leq 0.001$
	.1Hz	0%	52%	58%	0.00

Multi-nomial Logistic Regression: Mouse and Tank Data

Table B.6.: Multinomial Logistic Regression using Tank and Mouse Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
10%	60Hz	90%	86%	89%	$p \leq 0.001$
	30Hz	91%	87%	90%	$p \leq 0.001$
	12Hz	82%	88%	91%	$p \leq 0.001$
	6Hz	84%	79%	82%	$p \leq 0.001$
	3Hz	92%	81%	84%	$p \leq 0.001$
	2Hz	94%	90%	92%	$p \leq 0.001$
	1Hz	90%	92%	94%	$p \leq 0.001$
	.5Hz	82%	88%	90%	$p \leq 0.001$
	.25Hz	63%	79%	82%	$p \leq 0.001$
	.1Hz	87%	58%	63%	$p \leq 0.001$
25%	60Hz	86%	84%	87%	$p \leq 0.001$
	30Hz	87%	82%	86%	$p \leq 0.001$
	12Hz	88%	83%	87%	$p \leq 0.001$
	6Hz	87%	85%	88%	$p \leq 0.001$
	3Hz	82%	84%	87%	$p \leq 0.001$
	2Hz	83%	78%	82%	$p \leq 0.001$
	1Hz	90%	79%	83%	$p \leq 0.001$
	.5Hz	86%	88%	90%	$p \leq 0.001$
	.25Hz	76%	83%	86%	$p \leq 0.001$
	.1Hz	91%	72%	76%	$p \leq 0.001$
50%	60Hz	90%	87%	91%	$p \leq 0.001$
	30Hz	92%	86%	90%	$p \leq 0.001$

continued on next page

Table B.6.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	87%	88%	92%	$p \leq 0.001$
	6Hz	88%	83%	87%	$p \leq 0.001$
	3Hz	87%	84%	88%	$p \leq 0.001$
	2Hz	89%	83%	87%	$p \leq 0.001$
	1Hz	88%	85%	89%	$p \leq 0.001$
	.5Hz	87%	84%	88%	$p \leq 0.001$
	.25Hz	81%	83%	87%	$p \leq 0.001$
	.1Hz	95%	76%	81%	$p \leq 0.001$
75%	60Hz	96%	91%	95%	$p \leq 0.001$
	30Hz	97%	93%	96%	$p \leq 0.001$
	12Hz	95%	93%	97%	$p \leq 0.001$
	6Hz	91%	91%	95%	$p \leq 0.001$
	3Hz	91%	86%	91%	$p \leq 0.001$
	2Hz	97%	86%	91%	$p \leq 0.001$
	1Hz	95%	93%	97%	$p \leq 0.001$
	.5Hz	93%	91%	95%	$p \leq 0.001$
	.25Hz	89%	88%	93%	$p \leq 0.001$
	.1Hz	95%	83%	89%	$p \leq 0.001$
90%	60Hz	95%	88%	95%	$p \leq 0.001$
	30Hz	98%	88%	95%	$p \leq 0.001$
	12Hz	98%	93%	98%	$p \leq 0.001$
	6Hz	97%	93%	98%	$p \leq 0.001$
	3Hz	91%	91%	97%	$p \leq 0.001$
	2Hz	97%	82%	91%	$p \leq 0.001$
	1Hz	93%	91%	97%	$p \leq 0.001$

continued on next page

Table B.6.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	95%	84%	93%	$p \leq 0.001$
	.25Hz	94%	88%	95%	$p \leq 0.001$
	.1Hz	0%	86%	94%	0.00

Multi-nomial Logistic Regression: Mouse. Tank. and Experience Data

Table B.7.: Multinomial Logistic Regression using Tank, Mouse and Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
10%	60Hz	84%	78%	81%	$p \leq 0.001$
	30Hz	90%	81%	84%	$p \leq 0.001$
	12Hz	88%	87%	90%	$p \leq 0.001$
	6Hz	85%	85%	88%	$p \leq 0.001$
	3Hz	88%	82%	85%	$p \leq 0.001$
	2Hz	92%	85%	88%	$p \leq 0.001$
	1Hz	89%	89%	92%	$p \leq 0.001$
	.5Hz	84%	87%	89%	$p \leq 0.001$
	.25Hz	68%	81%	84%	$p \leq 0.001$
	.1Hz	84%	64%	68%	$p \leq 0.001$
25%	60Hz	88%	81%	84%	$p \leq 0.001$
	30Hz	79%	85%	88%	$p \leq 0.001$
	12Hz	84%	76%	79%	$p \leq 0.001$
	6Hz	89%	80%	84%	$p \leq 0.001$
	3Hz	89%	86%	89%	$p \leq 0.001$
	2Hz	93%	86%	89%	$p \leq 0.001$
	1Hz	88%	91%	93%	$p \leq 0.001$
	.5Hz	81%	85%	88%	$p \leq 0.001$
	.25Hz	78%	77%	81%	$p \leq 0.001$
	.1Hz	89%	74%	78%	$p \leq 0.001$
50%	60Hz	88%	85%	89%	$p \leq 0.001$
	30Hz	90%	84%	88%	$p \leq 0.001$

continued on next page

Table B.7.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	88%	87%	90%	$p \leq 0.001$
	6Hz	87%	84%	88%	$p \leq 0.001$
	3Hz	87%	83%	87%	$p \leq 0.001$
	2Hz	87%	83%	87%	$p \leq 0.001$
	1Hz	90%	83%	87%	$p \leq 0.001$
	.5Hz	88%	86%	90%	$p \leq 0.001$
	.25Hz	82%	84%	88%	$p \leq 0.001$
	.1Hz	90%	77%	82%	$p \leq 0.001$
75%	60Hz	90%	85%	90%	$p \leq 0.001$
	30Hz	90%	85%	90%	$p \leq 0.001$
	12Hz	89%	85%	90%	$p \leq 0.001$
	6Hz	90%	84%	89%	$p \leq 0.001$
	3Hz	87%	85%	90%	$p \leq 0.001$
	2Hz	93%	81%	87%	$p \leq 0.001$
	1Hz	88%	88%	93%	$p \leq 0.001$
	.5Hz	88%	82%	88%	$p \leq 0.001$
	.25Hz	89%	82%	88%	$p \leq 0.001$
	.1Hz	97%	83%	89%	$p \leq 0.001$
90%	60Hz	90%	91%	97%	$p \leq 0.001$
	30Hz	98%	81%	90%	$p \leq 0.001$
	12Hz	98%	93%	98%	$p \leq 0.001$
	6Hz	91%	93%	98%	$p \leq 0.001$
	3Hz	96%	82%	91%	$p \leq 0.001$
	2Hz	93%	89%	96%	$p \leq 0.001$
	1Hz	93%	84%	93%	$p \leq 0.001$

continued on next page

Table B.7.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	89%	84%	93%	$p \leq 0.001$
	.25Hz	91%	79%	89%	$p \leq 0.001$
	.1Hz	71%	83%	91%	0.00

Table B.8.: Multinomial Logistic Regression using Tank. Mouse and Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
10%	60Hz	84%	78%	81%	$p \leq 0.001$
	30Hz	90%	81%	84%	$p \leq 0.001$
	12Hz	88%	87%	90%	$p \leq 0.001$
	6Hz	85%	85%	88%	$p \leq 0.001$
	3Hz	88%	82%	85%	$p \leq 0.001$
	2Hz	92%	85%	88%	$p \leq 0.001$
	1Hz	89%	89%	92%	$p \leq 0.001$
	.5Hz	84%	87%	89%	$p \leq 0.001$
	.25Hz	68%	81%	84%	$p \leq 0.001$
	.1Hz	84%	64%	68%	$p \leq 0.001$
25%	60Hz	88%	81%	84%	$p \leq 0.001$
	30Hz	79%	85%	88%	$p \leq 0.001$
	12Hz	84%	76%	79%	$p \leq 0.001$
	6Hz	89%	80%	84%	$p \leq 0.001$
	3Hz	89%	86%	89%	$p \leq 0.001$
	2Hz	93%	86%	89%	$p \leq 0.001$
	1Hz	88%	91%	93%	$p \leq 0.001$

continued on next page

Table B.8.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	81%	85%	88%	$p \leq 0.001$
	.25Hz	78%	77%	81%	$p \leq 0.001$
	.1Hz	89%	74%	78%	$p \leq 0.001$
50%	60Hz	88%	85%	89%	$p \leq 0.001$
	30Hz	90%	84%	88%	$p \leq 0.001$
	12Hz	88%	87%	90%	$p \leq 0.001$
	6Hz	87%	84%	88%	$p \leq 0.001$
	3Hz	87%	83%	87%	$p \leq 0.001$
	2Hz	87%	83%	87%	$p \leq 0.001$
	1Hz	90%	83%	87%	$p \leq 0.001$
	.5Hz	88%	86%	90%	$p \leq 0.001$
	.25Hz	82%	84%	88%	$p \leq 0.001$
	.1Hz	90%	77%	82%	$p \leq 0.001$
75%	60Hz	90%	85%	90%	$p \leq 0.001$
	30Hz	90%	85%	90%	$p \leq 0.001$
	12Hz	89%	85%	90%	$p \leq 0.001$
	6Hz	90%	84%	89%	$p \leq 0.001$
	3Hz	87%	85%	90%	$p \leq 0.001$
	2Hz	93%	81%	87%	$p \leq 0.001$
	1Hz	88%	88%	93%	$p \leq 0.001$
	.5Hz	88%	82%	88%	$p \leq 0.001$
	.25Hz	89%	82%	88%	$p \leq 0.001$
	.1Hz	97%	83%	89%	$p \leq 0.001$
90%	60Hz	90%	91%	97%	$p \leq 0.001$
	30Hz	98%	81%	90%	$p \leq 0.001$

continued on next page

Table B.8.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	98%	93%	98%	$p \leq 0.001$
	6Hz	91%	93%	98%	$p \leq 0.001$
	3Hz	96%	82%	91%	$p \leq 0.001$
	2Hz	93%	89%	96%	$p \leq 0.001$
	1Hz	93%	84%	93%	$p \leq 0.001$
	.5Hz	89%	84%	93%	$p \leq 0.001$
	.25Hz	91%	79%	89%	$p \leq 0.001$
	.1Hz	71%	83%	91%	0.00

B.1.2 Neural Networks

Neural Network: Experience

Table B.9.: Neural Network using Experience Only

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	30%	34%	39%	.5
	30Hz	27%	31%	35%	.9
	12Hz	29%	33%	37%	.8
	6Hz	30%	34%	38%	.8
	3Hz	31%	35%	39%	.5
	2Hz	29%	33%	37%	.8
	1Hz	30%	34%	38%	.5
	.5Hz	31%	35%	39%	.5
	.25Hz	28%	32%	36%	.9
	.1Hz	29%	33%	38%	.9
0.25	60Hz	28%	32%	37%	.8
	30Hz	31%	36%	41%	.5
	12Hz	27%	31%	36%	.9
	6Hz	31%	36%	41%	.5
	3Hz	29%	33%	37%	.9
	2Hz	28%	33%	37%	.8
	1Hz	29%	33%	37%	.8
	.5Hz	28%	32%	37%	.8
	.25Hz	27%	31%	36%	1
	.1Hz	25%	30%	34%	1
0.50	60Hz	27%	32%	38%	.9
	30Hz	33%	38%	44%	.3
	12Hz	27%	33%	38%	.8

continued on next page

Table B.9.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	6Hz	27%	32%	38%	.9
	3Hz	33%	38%	44%	.5
	2Hz	27%	33%	38%	.9
	1Hz	30%	35%	41%	.9
	.5Hz	34%	40%	46%	.3
	.25Hz	28%	33%	39%	.7
	.1Hz	29%	34%	40%	.6
0.75	60Hz	28%	36%	44%	.3
	30Hz	26%	33%	42%	.7
	12Hz	25%	32%	40%	.9
	6Hz	27%	35%	43%	.5
	3Hz	25%	33%	41%	.9
	2Hz	23%	31%	39%	1
	1Hz	26%	34%	42%	.5
	.5Hz	30%	38%	47%	.3
	.25Hz	20%	27%	35%	1
	.1Hz	25%	33%	41%	.7
0.90	60Hz	21%	33%	47%	.9
	30Hz	16%	26%	40%	1
	12Hz	17%	28%	42%	.9
	6Hz	36%	49%	63%	.6
	3Hz	28%	40%	54%	.6
	2Hz	34%	47%	61%	.7
	1Hz	16%	26%	40%	1
	.5Hz	21%	33%	47%	.9

continued on next page

Table B.9.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.25Hz	18%	29%	43%	.9
	.1Hz	18%	29%	43%	1

Neural Network: Mouse Data

Table B.10.: Neural Network using Mouse Data Only

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	82%	85%	88%	$p \leq 0.001$
	30Hz	84%	87%	90%	$p \leq 0.001$
	12Hz	82%	85%	88%	$p \leq 0.001$
	6Hz	85%	88%	91%	$p \leq 0.001$
	3Hz	88%	91%	93%	$p \leq 0.001$
	2Hz	89%	91%	94%	$p \leq 0.001$
	1Hz	82%	85%	88%	$p \leq 0.001$
	.5Hz	86%	89%	91%	$p \leq 0.001$
	.25Hz	72%	76%	79%	$p \leq 0.001$
	.1Hz	69%	73%	77%	$p \leq 0.001$
0.25	60Hz	86%	89%	92%	$p \leq 0.001$
	30Hz	84%	88%	91%	$p \leq 0.001$
	12Hz	86%	89%	92%	$p \leq 0.001$
	6Hz	89%	92%	95%	0.00
	3Hz	89%	92%	95%	$p \leq 0.001$
	2Hz	86%	90%	92%	$p \leq 0.001$
	1Hz	85%	89%	91%	$p \leq 0.001$
	.5Hz	84%	88%	91%	$p \leq 0.001$
	.25Hz	83%	87%	90%	$p \leq 0.001$
	.1Hz	71%	75%	79%	$p \leq 0.001$
0.50	60Hz	87%	91%	94%	$p \leq 0.001$
	30Hz	84%	88%	91%	$p \leq 0.001$
	12Hz	85%	89%	92%	$p \leq 0.001$

continued on next page

Table B.10.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	6Hz	79%	83%	87%	$p \leq 0.001$
	3Hz	90%	94%	96%	$p \leq 0.001$
	2Hz	90%	93%	96%	$p \leq 0.001$
	1Hz	87%	90%	94%	$p \leq 0.001$
	.5Hz	82%	87%	90%	$p \leq 0.001$
	.25Hz	87%	91%	94%	$p \leq 0.001$
	.1Hz	78%	83%	87%	$p \leq 0.001$
0.75	60Hz	82%	88%	93%	$p \leq 0.001$
	30Hz	81%	88%	93%	$p \leq 0.001$
	12Hz	81%	88%	93%	$p \leq 0.001$
	6Hz	89%	94%	97%	$p \leq 0.001$
	3Hz	87%	92%	96%	$p \leq 0.001$
	2Hz	90%	95%	98%	$p \leq 0.001$
	1Hz	89%	94%	97%	$p \leq 0.001$
	.5Hz	84%	90%	95%	$p \leq 0.001$
	.25Hz	78%	85%	90%	$p \leq 0.001$
	.1Hz	71%	79%	85%	$p \leq 0.001$
0.90	60Hz	78%	89%	96%	$p \leq 0.001$
	30Hz	83%	93%	98%	$p \leq 0.001$
	12Hz	85%	95%	99%	$p \leq 0.001$
	6Hz	78%	89%	96%	$p \leq 0.001$
	3Hz	88%	96%	100%	$p \leq 0.001$
	2Hz	83%	93%	98%	$p \leq 0.001$
	1Hz	81%	91%	97%	$p \leq 0.001$
	.5Hz	81%	91%	97%	$p \leq 0.001$

continued on next page

Table B.10.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.25Hz	81%	91%	97%	0.00
	.1Hz	79%	90%	96%	0.00

Neural Network: Mouse and Experience Data

Table B.11.: Neural Network Mouse and Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	84%	87%	90%	$p \leq 0.001$
	30Hz	86%	89%	92%	$p \leq 0.001$
	12Hz	83%	87%	89%	$p \leq 0.001$
	6Hz	87%	90%	92%	$p \leq 0.001$
	3Hz	85%	88%	91%	$p \leq 0.001$
	2Hz	85%	88%	91%	$p \leq 0.001$
	1Hz	87%	90%	92%	$p \leq 0.001$
	.5Hz	86%	89%	91%	$p \leq 0.001$
	.25Hz	80%	83%	86%	$p \leq 0.001$
	.1Hz	76%	80%	83%	0.00
0.25	60Hz	85%	89%	92%	0.00
	30Hz	85%	88%	91%	$p \leq 0.001$
	12Hz	83%	87%	90%	$p \leq 0.001$
	6Hz	86%	90%	92%	$p \leq 0.001$
	3Hz	86%	89%	92%	0.00
	2Hz	87%	90%	93%	$p \leq 0.001$
	1Hz	86%	90%	92%	$p \leq 0.001$
	.5Hz	87%	90%	93%	$p \leq 0.001$
	.25Hz	80%	83%	87%	$p \leq 0.001$
	.1Hz	68%	73%	77%	$p \leq 0.001$
0.50	60Hz	84%	88%	92%	$p \leq 0.001$
	30Hz	82%	86%	90%	$p \leq 0.001$

continued on next page

Table B.11.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	86%	90%	93%	$p \leq 0.001$
	6Hz	87%	91%	94%	$p \leq 0.001$
	3Hz	85%	89%	93%	$p \leq 0.001$
	2Hz	87%	91%	94%	$p \leq 0.001$
	1Hz	88%	91%	94%	$p \leq 0.001$
	.5Hz	82%	86%	90%	$p \leq 0.001$
	.25Hz	88%	92%	95%	$p \leq 0.001$
	.1Hz	70%	75%	80%	$p \leq 0.001$
0.75	60Hz	82%	88%	93%	$p \leq 0.001$
	30Hz	82%	88%	93%	$p \leq 0.001$
	12Hz	55%	63%	71%	$p \leq 0.001$
	6Hz	87%	93%	96%	$p \leq 0.001$
	3Hz	84%	90%	94%	$p \leq 0.001$
	2Hz	87%	93%	96%	$p \leq 0.001$
	1Hz	80%	86%	91%	$p \leq 0.001$
	.5Hz	85%	91%	95%	$p \leq 0.001$
	.25Hz	89%	94%	97%	$p \leq 0.001$
	.1Hz	74%	81%	87%	$p \leq 0.001$
0.90	60Hz	76%	88%	95%	$p \leq 0.001$
	30Hz	78%	89%	96%	$p \leq 0.001$
	12Hz	74%	86%	94%	$p \leq 0.001$
	6Hz	72%	84%	93%	$p \leq 0.001$
	3Hz	76%	88%	95%	$p \leq 0.001$
	2Hz	88%	96%	100%	$p \leq 0.001$
	1Hz	78%	89%	96%	$p \leq 0.001$

continued on next page

Table B.11.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	76%	88%	95%	$p \leq 0.001$
	.25Hz	75%	86%	94%	$p \leq 0.001$
	.1Hz	79%	90%	96%	$p \leq 0.001$

Neural Network: Tank Data

Table B.12.: Neural Network using Tank Variables

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	40%	44%	49%	$p \leq 0.001$
	30Hz	41%	45%	50%	$p \leq 0.001$
	12Hz	38%	43%	47%	$p \leq 0.001$
	6Hz	37%	41%	46%	$p \leq 0.001$
	3Hz	37%	41%	46%	$p \leq 0.001$
	2Hz	40%	45%	49%	$p \leq 0.001$
	1Hz	44%	48%	52%	$p \leq 0.001$
	.5Hz	45%	49%	53%	$p \leq 0.001$
	.25Hz	43%	48%	52%	$p \leq 0.001$
	.1Hz	42%	46%	50%	0.00
0.25	60Hz	45%	50%	54%	$p \leq 0.001$
	30Hz	41%	45%	50%	$p \leq 0.001$
	12Hz	47%	52%	56%	$p \leq 0.001$
	6Hz	49%	54%	59%	$p \leq 0.001$
	3Hz	42%	46%	51%	$p \leq 0.001$
	2Hz	43%	48%	53%	$p \leq 0.001$
	1Hz	40%	45%	49%	$p \leq 0.001$
	.5Hz	43%	47%	52%	$p \leq 0.001$
	.25Hz	45%	50%	55%	$p \leq 0.001$
	.1Hz	41%	46%	51%	$p \leq 0.001$
0.50	60Hz	46%	52%	58%	$p \leq 0.001$
	30Hz	47%	53%	59%	$p \leq 0.001$
	12Hz	45%	51%	57%	$p \leq 0.001$

continued on next page

Table B.12.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	6Hz	40%	46%	52%	$p \leq 0.001$
	3Hz	39%	45%	51%	$p \leq 0.001$
	2Hz	46%	52%	58%	$p \leq 0.001$
	1Hz	48%	54%	60%	$p \leq 0.001$
	.5Hz	41%	47%	52%	$p \leq 0.001$
	.25Hz	50%	55%	61%	$p \leq 0.001$
	.1Hz	43%	49%	55%	$p \leq 0.001$
0.75	60Hz	45%	54%	62%	$p \leq 0.001$
	30Hz	47%	55%	63%	$p \leq 0.001$
	12Hz	46%	54%	63%	$p \leq 0.001$
	6Hz	39%	48%	56%	$p \leq 0.001$
	3Hz	38%	46%	54%	.1
	2Hz	43%	52%	60%	$p \leq 0.001$
	1Hz	39%	47%	55%	$p \leq 0.001$
	.5Hz	44%	52%	60%	$p \leq 0.001$
	.25Hz	51%	59%	67%	$p \leq 0.001$
	.1Hz	48%	56%	65%	$p \leq 0.001$
0.90	60Hz	41%	54%	68%	$p \leq 0.001$
	30Hz	55%	68%	80%	$p \leq 0.001$
	12Hz	44%	58%	71%	$p \leq 0.001$
	6Hz	37%	51%	64%	$p \leq 0.001$
	3Hz	42%	56%	69%	$p \leq 0.001$
	2Hz	31%	44%	58%	.2
	1Hz	31%	44%	58%	.4
	.5Hz	44%	58%	71%	$p \leq 0.001$

continued on next page

Table B.12.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.25Hz	42%	55%	68%	0.00
	.1Hz	30%	43%	57%	.2

Neural Network: Tank and Experience Data

Table B.13.: Neural Network Tank Variables and Experience

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	35%	39%	43%	.01
	30Hz	39%	43%	48%	$p \leq 0.001$
	12Hz	35%	39%	43%	.02
	6Hz	35%	40%	44%	$p \leq 0.001$
	3Hz	34%	38%	42%	0
	2Hz	36%	40%	44%	$p \leq 0.001$
	1Hz	41%	45%	49%	$p \leq 0.001$
	.5Hz	34%	38%	43%	0
	.25Hz	38%	43%	47%	$p \leq 0.001$
	.1Hz	40%	45%	49%	$p \leq 0.001$
0.25	60Hz	39%	44%	48%	$p \leq 0.001$
	30Hz	38%	42%	46%	$p \leq 0.001$
	12Hz	39%	44%	48%	$p \leq 0.001$
	6Hz	40%	44%	48%	$p \leq 0.001$
	3Hz	44%	48%	52%	$p \leq 0.001$
	2Hz	41%	45%	50%	$p \leq 0.001$
	1Hz	38%	42%	46%	$p \leq 0.001$
	.5Hz	32%	36%	40%	.3
	.25Hz	44%	48%	52%	$p \leq 0.001$
	.1Hz	42%	46%	50%	$p \leq 0.001$
0.50	60Hz	40%	44%	49%	$p \leq 0.001$
	30Hz	42%	46%	51%	$p \leq 0.001$

continued on next page

Table B.13.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	40%	44%	49%	$p \leq 0.001$
	6Hz	43%	48%	53%	$p \leq 0.001$
	3Hz	44%	49%	54%	$p \leq 0.001$
	2Hz	42%	47%	52%	$p \leq 0.001$
	1Hz	37%	41%	46%	0
	.5Hz	38%	42%	47%	$p \leq 0.001$
	.25Hz	51%	56%	60%	$p \leq 0.001$
	.1Hz	44%	49%	54%	$p \leq 0.001$
0.75	60Hz	40%	45%	50%	$p \leq 0.001$
	30Hz	41%	46%	50%	$p \leq 0.001$
	12Hz	45%	49%	54%	$p \leq 0.001$
	6Hz	41%	45%	50%	$p \leq 0.001$
	3Hz	44%	48%	53%	$p \leq 0.001$
	2Hz	39%	44%	49%	$p \leq 0.001$
	1Hz	45%	49%	54%	$p \leq 0.001$
	.5Hz	44%	49%	54%	$p \leq 0.001$
	.25Hz	43%	48%	52%	0
	.1Hz	42%	47%	51%	$p \leq 0.001$
0.90	60Hz	41%	47%	53%	$p \leq 0.001$
	30Hz	39%	44%	50%	$p \leq 0.001$
	12Hz	42%	48%	54%	0
	6Hz	40%	45%	51%	0
	3Hz	36%	42%	48%	$p \leq 0.001$
	2Hz	45%	51%	57%	$p \leq 0.001$
	1Hz	48%	53%	59%	0

continued on next page

Table B.13.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	44%	50%	56%	$p \leq 0.001$
	.25Hz	46%	52%	58%	$p \leq 0.001$
	.1Hz	45%	51%	57%	0

Neural Network: Mouse and Tank Data

Table B.14.: Neural Network using Tank and Mouse Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	81%	85%	88%	$p \leq 0.001$
	30Hz	84%	87%	90%	$p \leq 0.001$
	12Hz	79%	82%	85%	0
	6Hz	87%	90%	93%	$p \leq 0.001$
	3Hz	88%	91%	93%	$p \leq 0.001$
	2Hz	87%	90%	92%	$p \leq 0.001$
	1Hz	77%	80%	84%	$p \leq 0.001$
	.5Hz	82%	85%	88%	$p \leq 0.001$
	.25Hz	77%	81%	84%	$p \leq 0.001$
	.1Hz	56%	60%	65%	$p \leq 0.001$
0.25	60Hz	85%	89%	92%	$p \leq 0.001$
	30Hz	87%	90%	93%	$p \leq 0.001$
	12Hz	79%	83%	87%	$p \leq 0.001$
	6Hz	85%	88%	91%	0
	3Hz	88%	91%	94%	$p \leq 0.001$
	2Hz	86%	89%	92%	$p \leq 0.001$
	1Hz	82%	85%	89%	$p \leq 0.001$
	.5Hz	85%	88%	91%	$p \leq 0.001$
	.25Hz	80%	84%	87%	$p \leq 0.001$
	.1Hz	67%	72%	76%	$p \leq 0.001$
0.50	60Hz	86%	90%	93%	0
	30Hz	87%	91%	94%	$p \leq 0.001$
	12Hz	88%	91%	94%	$p \leq 0.001$

continued on next page

Table B.14.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	6Hz	84%	88%	92%	$p \leq 0.001$
	3Hz	86%	90%	93%	$p \leq 0.001$
	2Hz	86%	90%	93%	$p \leq 0.001$
	1Hz	87%	91%	94%	$p \leq 0.001$
	.5Hz	85%	89%	93%	$p \leq 0.001$
	.25Hz	83%	87%	91%	$p \leq 0.001$
	.1Hz	71%	76%	81%	$p \leq 0.001$
0.75	60Hz	87%	93%	96%	$p \leq 0.001$
	30Hz	85%	90%	95%	$p \leq 0.001$
	12Hz	87%	93%	96%	$p \leq 0.001$
	6Hz	85%	91%	95%	$p \leq 0.001$
	3Hz	87%	92%	96%	$p \leq 0.001$
	2Hz	86%	92%	96%	$p \leq 0.001$
	1Hz	90%	95%	98%	0
	.5Hz	88%	93%	97%	$p \leq 0.001$
	.25Hz	81%	88%	93%	$p \leq 0.001$
	.1Hz	74%	81%	87%	$p \leq 0.001$
0.90	60Hz	68%	81%	90%	$p \leq 0.001$
	30Hz	78%	89%	96%	$p \leq 0.001$
	12Hz	81%	91%	97%	$p \leq 0.001$
	6Hz	74%	86%	94%	$p \leq 0.001$
	3Hz	88%	96%	100%	$p \leq 0.001$
	2Hz	78%	89%	96%	$p \leq 0.001$
	1Hz	88%	96%	100%	$p \leq 0.001$
	.5Hz	66%	79%	89%	$p \leq 0.001$

continued on next page

Table B.14.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.25Hz	73%	84%	93%	$p \leq 0.001$
	.1Hz	73%	84%	93%	$p \leq 0.001$

Neural Network: Mouse, Tank , and Experience Data

Table B.15.: Neural Network using Tank. Mouse and Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	80%	83%	86%	$p \leq 0.001$
	30Hz	84%	87%	90%	0.00
	12Hz	85%	88%	91%	$p \leq 0.001$
	6Hz	83%	86%	89%	$p \leq 0.001$
	3Hz	81%	84%	87%	$p \leq 0.001$
	2Hz	85%	88%	91%	$p \leq 0.001$
	1Hz	82%	85%	88%	$p \leq 0.001$
	.5Hz	84%	87%	90%	$p \leq 0.001$
	.25Hz	83%	86%	89%	$p \leq 0.001$
	.1Hz	77%	80%	84%	$p \leq 0.001$
0.25	60Hz	84%	88%	91%	$p \leq 0.001$
	30Hz	84%	88%	91%	$p \leq 0.001$
	12Hz	85%	88%	91%	$p \leq 0.001$
	6Hz	86%	89%	92%	$p \leq 0.001$
	3Hz	87%	90%	93%	$p \leq 0.001$
	2Hz	88%	91%	94%	$p \leq 0.001$
	1Hz	89%	92%	94%	$p \leq 0.001$
	.5Hz	87%	90%	93%	$p \leq 0.001$
	.25Hz	85%	89%	91%	$p \leq 0.001$
	.1Hz	78%	82%	86%	$p \leq 0.001$
0.50	60Hz	87%	91%	94%	$p \leq 0.001$
	30Hz	84%	88%	92%	$p \leq 0.001$

continued on next page

Table B.15.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	88%	91%	94%	$p \leq 0.001$
	6Hz	87%	91%	94%	$p \leq 0.001$
	3Hz	88%	92%	95%	$p \leq 0.001$
	2Hz	87%	91%	94%	$p \leq 0.001$
	1Hz	88%	92%	95%	$p \leq 0.001$
	.5Hz	86%	90%	93%	$p \leq 0.001$
	.25Hz	88%	92%	95%	$p \leq 0.001$
	.1Hz	79%	84%	88%	$p \leq 0.001$
0.75	60Hz	77%	84%	89%	$p \leq 0.001$
	30Hz	79%	86%	91%	$p \leq 0.001$
	12Hz	84%	90%	94%	$p \leq 0.001$
	6Hz	80%	86%	91%	$p \leq 0.001$
	3Hz	87%	92%	96%	$p \leq 0.001$
	2Hz	83%	89%	94%	$p \leq 0.001$
	1Hz	89%	94%	97%	$p \leq 0.001$
	.5Hz	90%	95%	98%	$p \leq 0.001$
	.25Hz	80%	86%	91%	$p \leq 0.001$
	.1Hz	73%	80%	86%	$p \leq 0.001$
0.90	60Hz	81%	91%	97%	$p \leq 0.001$
	30Hz	70%	82%	91%	$p \leq 0.001$
	12Hz	74%	86%	94%	0.00
	6Hz	88%	96%	100% 0	
	3Hz	81%	91%	97%	$p \leq 0.001$
	2Hz	88%	96%	100%	$p \leq 0.001$
	1Hz	83%	93%	98%	$p \leq 0.001$

continued on next page

Table B.15.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	83%	93%	98%	$p \leq 0.001$
	.25Hz	75%	86%	94%	$p \leq 0.001$
	.1Hz	71%	83%	91%	$p \leq 0.001$

B.1.3 Random Forests

Random Forest: Experience

Table B.16.: Random Forest using Experience Only

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	30%	34%	38%	.6
	30Hz	27%	31%	35%	1
	12Hz	30%	34%	38%	.6
	6Hz	28%	32%	36%	1
	3Hz	29%	33%	37%	.8
	2Hz	30%	34%	38%	.6
	1Hz	29%	33%	37%	.7
	.5Hz	31%	35%	39%	.5
	.25Hz	29%	33%	37%	.8
	.1Hz	27%	31%	35%	1
0.25	60Hz	28%	32%	37%	.9
	30Hz	27%	31%	36%	1
	12Hz	28%	32%	37%	.9
	6Hz	28%	32%	37%	1
	3Hz	29%	33%	38%	.9
	2Hz	30%	34%	39%	.5
	1Hz	28%	32%	37%	.9
	.5Hz	29%	34%	38%	.6
	.25Hz	29%	33%	38%	.9
	.1Hz	27%	32%	36%	.9
0.50	60Hz	30%	36%	41%	.5
	30Hz	27%	32%	38%	1
	12Hz	25%	31%	36%	1

continued on next page

Table B.16.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	6Hz	25%	30%	35%	1
	3Hz	28%	34%	39%	.9
	2Hz	26%	32%	37%	.9
	1Hz	29%	34%	40%	.9
	.5Hz	28%	33%	39%	1
	.25Hz	27%	33%	38%	.8
	.1Hz	27%	32%	38%	.8
0.75	60Hz	26%	34%	42%	.5
	30Hz	26%	34%	42%	.6
	12Hz	23%	30%	38%	1
	6Hz	27%	35%	43%	.5
	3Hz	27%	35%	43%	.8
	2Hz	25%	32%	40%	.9
	1Hz	25%	33%	41%	.7
	.5Hz	28%	36%	44%	.5
	.25Hz	24%	31%	39%	.9
	.1Hz	25%	33%	41%	.7
0.90	60Hz	18%	30%	43%	1
	30Hz	14%	25%	38%	1
	12Hz	16%	26%	40%	1
	6Hz	24%	37%	51%	1
	3Hz	23%	35%	49%	.8
	2Hz	23%	35%	49%	1
	1Hz	16%	26%	40%	1
	.5Hz	18%	30%	43%	1

continued on next page

Table B.16.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.25Hz	17%	28%	41%	1
	.1Hz	21%	33%	46%	1

Random Forest: Mouse Data

Table B.17.: Random Forest using Mouse Data Only

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	79%	82%	86%	$p \leq 0.001$
	30Hz	82%	85%	88%	0
	12Hz	80%	83%	86%	$p \leq 0.001$
	6Hz	84%	87%	90%	$p \leq 0.001$
	3Hz	87%	90%	93%	$p \leq 0.001$
	2Hz	89%	92%	94%	$p \leq 0.001$
	1Hz	89%	91%	94%	$p \leq 0.001$
	.5Hz	88%	91%	93%	$p \leq 0.001$
	.25Hz	87%	90%	92%	$p \leq 0.001$
	.1Hz	75%	79%	82%	$p \leq 0.001$
0.25	60Hz	83%	86%	89%	$p \leq 0.001$
	30Hz	82%	86%	89%	$p \leq 0.001$
	12Hz	85%	88%	91%	0
	6Hz	88%	91%	93%	$p \leq 0.001$
	3Hz	87%	90%	93%	$p \leq 0.001$
	2Hz	88%	91%	94%	$p \leq 0.001$
	1Hz	88%	91%	94%	0
	.5Hz	88%	91%	94%	$p \leq 0.001$
	.25Hz	88%	91%	94%	0
	.1Hz	78%	82%	86%	$p \leq 0.001$
0.50	60Hz	87%	91%	94%	$p \leq 0.001$
	30Hz	85%	89%	92%	$p \leq 0.001$
	12Hz	85%	89%	93%	$p \leq 0.001$

continued on next page

Table B.17.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	6Hz	88%	92%	94%	$p \leq 0.001$
	3Hz	89%	93%	96%	$p \leq 0.001$
	2Hz	86%	90%	93%	0
	1Hz	87%	91%	94%	$p \leq 0.001$
	.5Hz	88%	91%	94%	$p \leq 0.001$
	.25Hz	90%	94%	96%	$p \leq 0.001$
	.1Hz	77%	82%	86%	$p \leq 0.001$
0.75	60Hz	84%	90%	94%	$p \leq 0.001$
	30Hz	85%	91%	95%	$p \leq 0.001$
	12Hz	84%	90%	94%	$p \leq 0.001$
	6Hz	90%	95%	98%	$p \leq 0.001$
	3Hz	91%	96%	98%	$p \leq 0.001$
	2Hz	90%	95%	98%	$p \leq 0.001$
	1Hz	89%	94%	97%	$p \leq 0.001$
	.5Hz	87%	92%	96%	$p \leq 0.001$
	.25Hz	85%	90%	95%	$p \leq 0.001$
	.1Hz	77%	84%	90%	$p \leq 0.001$
0.90	60Hz	85%	95%	99%	0
	30Hz	83%	93%	98%	0
	12Hz	81%	91%	97%	$p \leq 0.001$
	6Hz	81%	91%	97%	0
	3Hz	91%	98%	100%	$p \leq 0.001$
	2Hz	74%	86%	94%	$p \leq 0.001$
	1Hz	78%	89%	96%	$p \leq 0.001$
	.5Hz	76%	88%	95%	$p \leq 0.001$

continued on next page

Table B.17.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.25Hz	83%	93%	98%	$p \leq 0.001$
	.1Hz	79%	90%	96%	0

Random Forest: Mouse and Experience Data

Table B.18.: Random Forest Mouse and Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	79%	82%	85%	0
	30Hz	81%	84%	87%	$p \leq 0.001$
	12Hz	88%	91%	93%	$p \leq 0.001$
	6Hz	81%	85%	88%	$p \leq 0.001$
	3Hz	88%	91%	93%	$p \leq 0.001$
	2Hz	90%	92%	94%	$p \leq 0.001$
	1Hz	89%	92%	94%	$p \leq 0.001$
	.5Hz	90%	92%	94%	$p \leq 0.001$
	.25Hz	81%	85%	88%	$p \leq 0.001$
	.1Hz	75%	79%	82%	$p \leq 0.001$
0.25	60Hz	83%	87%	90%	$p \leq 0.001$
	30Hz	83%	87%	90%	$p \leq 0.001$
	12Hz	85%	88%	91%	$p \leq 0.001$
	6Hz	89%	92%	95%	$p \leq 0.001$
	3Hz	90%	93%	95%	$p \leq 0.001$
	2Hz	87%	90%	92%	$p \leq 0.001$
	1Hz	88%	91%	94%	$p \leq 0.001$
	.5Hz	89%	92%	94%	$p \leq 0.001$
	.25Hz	89%	92%	94%	$p \leq 0.001$
	.1Hz	79%	83%	86%	$p \leq 0.001$
0.50	60Hz	86%	90%	93%	$p \leq 0.001$
	30Hz	84%	88%	92%	$p \leq 0.001$
	12Hz	83%	87%	91%	$p \leq 0.001$

continued on next page

Table B.18.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	6Hz	86%	90%	93%	$p \leq 0.001$
	3Hz	87%	91%	94%	$p \leq 0.001$
	2Hz	86%	90%	93%	$p \leq 0.001$
	1Hz	88%	92%	95%	$p \leq 0.001$
	.5Hz	87%	91%	94%	$p \leq 0.001$
	.25Hz	89%	93%	95%	$p \leq 0.001$
	.1Hz	76%	81%	85%	$p \leq 0.001$
0.75	60Hz	84%	90%	94%	$p \leq 0.001$
	30Hz	82%	88%	93%	$p \leq 0.001$
	12Hz	84%	90%	94%	$p \leq 0.001$
	6Hz	88%	93%	97%	$p \leq 0.001$
	3Hz	87%	92%	96%	$p \leq 0.001$
	2Hz	89%	94%	97%	$p \leq 0.001$
	1Hz	84%	90%	94%	$p \leq 0.001$
	.5Hz	86%	92%	96%	$p \leq 0.001$
	.25Hz	86%	92%	96%	$p \leq 0.001$
	.1Hz	78%	85%	90%	$p \leq 0.001$
0.90	60Hz	78%	89%	96%	$p \leq 0.001$
	30Hz	78%	89%	96%	$p \leq 0.001$
	12Hz	85%	95%	99%	$p \leq 0.001$
	6Hz	72%	84%	93%	$p \leq 0.001$
	3Hz	74%	86%	94%	$p \leq 0.001$
	2Hz	78%	89%	96%	$p \leq 0.001$
	1Hz	76%	88%	95%	$p \leq 0.001$
	.5Hz	78%	89%	96%	$p \leq 0.001$

continued on next page

Table B.18.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.25Hz	81%	91%	97%	$p \leq 0.001$
	.1Hz	75%	86%	94%	$p \leq 0.001$

Random Forest: Tank Data

Table B.19.: Random Forest using Tank Variables

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	42%	47%	51%	$p \leq 0.001$
	30Hz	45%	50%	54%	$p \leq 0.001$
	12Hz	44%	48%	52%	$p \leq 0.001$
	6Hz	47%	52%	56%	$p \leq 0.001$
	3Hz	45%	49%	53%	$p \leq 0.001$
	2Hz	47%	51%	56%	$p \leq 0.001$
	1Hz	47%	52%	56%	$p \leq 0.001$
	.5Hz	41%	45%	50%	$p \leq 0.001$
	.25Hz	47%	52%	56%	$p \leq 0.001$
	.1Hz	47%	51%	56%	$p \leq 0.001$
0.25	60Hz	49%	53%	58%	$p \leq 0.001$
	30Hz	44%	49%	53%	$p \leq 0.001$
	12Hz	48%	53%	58%	$p \leq 0.001$
	6Hz	47%	52%	57%	$p \leq 0.001$
	3Hz	41%	45%	50%	$p \leq 0.001$
	2Hz	50%	54%	59%	$p \leq 0.001$
	1Hz	44%	49%	54%	$p \leq 0.001$
	.5Hz	46%	51%	56%	$p \leq 0.001$
	.25Hz	49%	54%	59%	$p \leq 0.001$
	.1Hz	49%	54%	59%	$p \leq 0.001$
0.50	60Hz	47%	53%	58%	$p \leq 0.001$
	30Hz	44%	49%	55%	$p \leq 0.001$
	12Hz	51%	56%	62%	$p \leq 0.001$

continued on next page

Table B.19.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	6Hz	44%	49%	55%	$p \leq 0.001$
	3Hz	46%	52%	58%	$p \leq 0.001$
	2Hz	51%	57%	63%	$p \leq 0.001$
	1Hz	46%	52%	58%	$p \leq 0.001$
	.5Hz	46%	52%	58%	$p \leq 0.001$
	.25Hz	50%	56%	62%	$p \leq 0.001$
	.1Hz	47%	53%	59%	$p \leq 0.001$
0.75	60Hz	46%	54%	63%	$p \leq 0.001$
	30Hz	51%	59%	67%	$p \leq 0.001$
	12Hz	48%	56%	65%	$p \leq 0.001$
	6Hz	41%	49%	57%	$p \leq 0.001$
	3Hz	49%	58%	66%	$p \leq 0.001$
	2Hz	43%	52%	60%	$p \leq 0.001$
	1Hz	38%	46%	55%	$p \leq 0.001$
	.5Hz	48%	56%	64%	$p \leq 0.001$
	.25Hz	51%	60%	68%	$p \leq 0.001$
	.1Hz	44%	52%	61%	$p \leq 0.001$
0.90	60Hz	37%	51%	64%	$p \leq 0.001$
	30Hz	48%	61%	74%	$p \leq 0.001$
	12Hz	32%	46%	59%	$p \leq 0.001$
	6Hz	37%	51%	64%	$p \leq 0.001$
	3Hz	51%	65%	77%	$p \leq 0.001$
	2Hz	39%	53%	66%	$p \leq 0.001$
	1Hz	41%	54%	68%	$p \leq 0.001$
	.5Hz	42%	56%	69%	$p \leq 0.001$

continued on next page

Table B.19.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.25Hz	43%	57%	70%	$p \leq 0.001$
	.1Hz	37%	50%	63%	$p \leq 0.001$

Random Forest: Tank Data and Experience

Table B.20.: Random Forest Tank Variables and Experience

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	36%	40%	45%	$p \leq 0.001$
	30Hz	40%	44%	49%	$p \leq 0.001$
	12Hz	38%	42%	47%	$p \leq 0.001$
	6Hz	38%	42%	47%	$p \leq 0.001$
	3Hz	41%	45%	49%	$p \leq 0.001$
	2Hz	43%	47%	51%7$p \leq 0.001$	
	1Hz	43%	47%	52%	$p \leq 0.001$
	.5Hz	45%	49%	53%	$p \leq 0.001$
	.25Hz	48%	52%	57%	$p \leq 0.001$
	.1Hz	46%	51%	55%	$p \leq 0.001$
0.25	60Hz	44%	49%	53%	$p \leq 0.001$
	30Hz	45%	50%	54%	$p \leq 0.001$
	12Hz	43%	48%	52%	$p \leq 0.001$
	6Hz	35%	39%	44%	.1
	3Hz	40%	44%	48%	$p \leq 0.001$
	2Hz	41%	45%	49%	$p \leq 0.001$
	1Hz	42%	47%	51%	$p \leq 0.001$
	.5Hz	44%	48%	53%	$p \leq 0.001$
	.25Hz	47%	52%	56%	$p \leq 0.001$
	.1Hz	47%	51%	55%	$p \leq 0.001$
0.50	60Hz	46%	51%	56%	$p \leq 0.001$
	30Hz	49%	54%	59%	$p \leq 0.001$

continued on next page

Table B.20.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	44%	49%	54%	$p \leq 0.001$
	6Hz	46%	50%	55%	$p \leq 0.001$
	3Hz	46%	51%	56%	$p \leq 0.001$
	2Hz	40%	44%	49%	$p \leq 0.001$
	1Hz	39%	44%	49%	$p \leq 0.001$
	.5Hz	46%	51%	55%	$p \leq 0.001$
	.25Hz	48%	53%	57%	$p \leq 0.001$
	.1Hz	46%	51%	55%	$p \leq 0.001$
0.75	60Hz	47%	51%	56%	$p \leq 0.001$
	30Hz	47%	52%	57%	$p \leq 0.001$
	12Hz	48%	53%	58%	$p \leq 0.001$
	6Hz	50%	55%	59%	$p \leq 0.001$
	3Hz	46%	51%	56%	$p \leq 0.001$
	2Hz	47%	52%	57%	$p \leq 0.001$
	1Hz	43%	48%	53%	$p \leq 0.001$
	.5Hz	49%	53%	58%	$p \leq 0.001$
	.25Hz	49%	54%	59%	$p \leq 0.001$
	.1Hz	48%	53%	58%	$p \leq 0.001$
0.90	60Hz	48%	54%	60%	$p \leq 0.001$
	30Hz	50%	56%	62%	$p \leq 0.001$
	12Hz	44%	50%	56%	$p \leq 0.001$
	6Hz	46%	52%	57%	$p \leq 0.001$
	3Hz	47%	53%	59%	0.00
	2Hz	47%	53%	59%	0.00
	1Hz	49%	54%	60%	$p \leq 0.001$

continued on next page

Table B.20.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	43%	49%	55%	0.00
	.25Hz	51%	56%	62%	$p \leq 0.001$
	.1Hz	48%	53%	59%	0.00

Random Forest: Mouse and Tank Data

Table B.21.: Random Forest using Tank and Mouse Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	81%	84%	87%	0.00
	30Hz	83%	86%	89%	$p \leq 0.001$
	12Hz	85%	88%	90%	$p \leq 0.001$
	6Hz	87%	90%	93%	$p \leq 0.001$
	3Hz	88%	91%	93%	$p \leq 0.001$
	2Hz	89%	91%	94%	$p \leq 0.001$
	1Hz	87%	90%	92%	$p \leq 0.001$
	.5Hz	85%	88%	91%	$p \leq 0.001$
	.25Hz	84%	88%	90%	$p \leq 0.001$
	.1Hz	73%	77%	81%	$p \leq 0.001$
0.25	60Hz	79%	83%	86%	0.00
	30Hz	82%	86%	89%	0.00
	12Hz	86%	89%	92%	$p \leq 0.001$
	6Hz	87%	90%	93%	$p \leq 0.001$
	3Hz	89%	92%	94%	$p \leq 0.001$
	2Hz	89%	92%	94%	$p \leq 0.001$
	1Hz	87%	90%	92%	$p \leq 0.001$
	.5Hz	88%	91%	93%	0.00
	.25Hz	87%	90%	93%	0.00
	.1Hz	78%	82%	86%	$p \leq 0.001$
0.50	60Hz	81%	86%	90%	$p \leq 0.001$
	30Hz	84%	88%	91%	$p \leq 0.001$
	12Hz	88%	92%	95%	$p \leq 0.001$

continued on next page

Table B.21.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	6Hz	86%	90%	93%	$p \leq 0.001$
	3Hz	87%	90%	94%	$p \leq 0.001$
	2Hz	88%	92%	95%	$p \leq 0.001$
	1Hz	90%	94%	96%	0.00
	.5Hz	88%	91%	94%	$p \leq 0.001$
	.25Hz	88%	92%	95%	$p \leq 0.001$
	.1Hz	77%	82%	86%	$p \leq 0.001$
0.75	60Hz	84%	90%	94%	$p \leq 0.001$
	30Hz	86%	92%	96%	$p \leq 0.001$
	12Hz	87%	93%	96%	$p \leq 0.001$
	6Hz	89%	94%	97%	$p \leq 0.001$
	3Hz	82%	88%	93%	$p \leq 0.001$
	2Hz	85%	91%	95%	$p \leq 0.001$
	1Hz	88%	93%	97%	$p \leq 0.001$
	.5Hz	90%	95%	98%	$p \leq 0.001$
	.25Hz	85%	91%	95%	$p \leq 0.001$
	.1Hz	81%	87%	92%	$p \leq 0.001$
0.90	60Hz	81%	91%	97%	$p \leq 0.001$
	30Hz	83%	93%	98%	$p \leq 0.001$
	12Hz	83%	93%	98%	0.00
	6Hz	81%	91%	97%	$p \leq 0.001$
	3Hz	88%	96%	100%	$p \leq 0.001$
	2Hz	81%	91%	97%	$p \leq 0.001$
	1Hz	88%	96%	100%	$p \leq 0.001$
	.5Hz	76%	88%	95%	$p \leq 0.001$

continued on next page

Table B.21.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.25Hz	79%	90%	96%	$p \leq 0.001$
	.1Hz	79%	90%	96%	$p \leq 0.001$

Random Forest: Mouse, Tank and Experience Data

Table B.22.: Random Forest using Tank. Mouse and Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	85%	88%	91%	$p \leq 0.001$
	30Hz	79%	82%	85%	$p \leq 0.001$
	12Hz	87%	90%	92%	$p \leq 0.001$
	6Hz	87%	90%	92%	$p \leq 0.001$
	3Hz	77%	80%	84%	$p \leq 0.001$
	2Hz	86%	89%	92%	$p \leq 0.001$
	1Hz	86%	89%	92%	$p \leq 0.001$
	.5Hz	83%	86%	89%	$p \leq 0.001$
	.25Hz	80%	83%	86%	$p \leq 0.001$
	.1Hz	59%	63%	67%	$p \leq 0.001$
0.25	60Hz	84%	88%	91%	$p \leq 0.001$
	30Hz	85%	88%	91%	$p \leq 0.001$
	12Hz	86%	90%	92%	$p \leq 0.001$
	6Hz	87%	90%	93%	$p \leq 0.001$
	3Hz	86%	89%	92%	$p \leq 0.001$
	2Hz	79%	83%	87%	$p \leq 0.001$
	1Hz	82%	86%	89%	$p \leq 0.001$
	.5Hz	80%	84%	87%	$p \leq 0.001$
	.25Hz	77%	81%	85%	$p \leq 0.001$
	.1Hz	69%	73%	78%	$p \leq 0.001$
0.50	60Hz	88%	92%	95%	$p \leq 0.001$
	30Hz	84%	88%	92%	$p \leq 0.001$

continued on next page

Table B.22.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	89%	93%	96%	$p \leq 0.001$
	6Hz	89%	93%	95%	$p \leq 0.001$
	3Hz	86%	90%	93%	$p \leq 0.001$
	2Hz	90%	93%	96%	$p \leq 0.001$
	1Hz	69%	74%	79%	$p \leq 0.001$
	.5Hz	85%	89%	92%	$p \leq 0.001$
	.25Hz	85%	89%	92%	$p \leq 0.001$
	.1Hz	68%	74%	79%	$p \leq 0.001$
0.75	60Hz	82%	88%	93%	$p \leq 0.001$
	30Hz	83%	89%	94%	$p \leq 0.001$
	12Hz	82%	88%	93%	$p \leq 0.001$
	6Hz	82%	88%	93%	$p \leq 0.001$
	3Hz	80%	87%	92%	$p \leq 0.001$
	2Hz	81%	87%	92%	$p \leq 0.001$
	1Hz	84%	90%	94%	$p \leq 0.001$
	.5Hz	89%	94%	97%	$p \leq 0.001$
	.25Hz	70%	78%	84%	$p \leq 0.001$
	.1Hz	59%	67%	75%	$p \leq 0.001$
0.90	60Hz	85%	95%	99%	$p \leq 0.001$
	30Hz	74%	86%	94%	$p \leq 0.001$
	12Hz	81%	91%	97%	$p \leq 0.001$
	6Hz	88%	96%	100%	$p \leq 0.001$
	3Hz	85%	95%	99%	$p \leq 0.001$
	2Hz	85%	95%	99%	$p \leq 0.001$
	1Hz	74%	86%	94%	$p \leq 0.001$

continued on next page

Table B.22.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	85%	95%	99%	$p \leq 0.001$
	.25Hz	77%	88%	95%	$p \leq 0.001$
	.1Hz	65%	78%	87%	$p \leq 0.001$

B.2 Delay Prediction

This section contains tables showing the Accuracyuracy of delay prediction models and is structured by model class and further sUpper Bounddivided by predictor variables

B.2.1 General Linear Model

General Linear Model : Experience Data

Table B.23.: General Linear Model using Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
10%	60Hz	55%	59%	63%	0.5
	30Hz	55%	60%	64%	0.5
	12Hz	55%	59%	63%	0.4
	6Hz	52%	57%	61%	0.8
	3Hz	53%	57%	61%	0.8
	2Hz	49%	53%	58%	1.0
	1Hz	52%	56%	61%	0.6
	.5Hz	36%	41%	46%	0.7
	.25Hz	49%	56%	62%	1.0
	.1Hz	59%	64%	68%	0.7
25%	60Hz	54%	59%	63%	0.5
	30Hz	55%	60%	64%	0.5
	12Hz	54%	59%	63%	0.5
	6Hz	54%	59%	64%	0.4
	3Hz	53%	57%	62%	0.6
	2Hz	52%	57%	61%	0.9
	1Hz	47%	52%	56%	1.0
	.5Hz	49%	55%	60%	0.2
	.25Hz	56%	63%	70%	0.5
	.1Hz	60%	65%	69%	0.4
50%	60Hz	53%	59%	65%	0.5
	30Hz	53%	59%	65%	0.5

continued on next page

Table B.23.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	53%	59%	64%	0.5
	6Hz	53%	59%	64%	0.5
	3Hz	52%	58%	63%	0.5
	2Hz	49%	55%	61%	0.8
	1Hz	50%	56%	62%	0.7
	.5Hz	47%	54%	61%	0.4
	.25Hz	55%	64%	72%	0.5
	.1Hz	59%	65%	70%	0.5
75%	60Hz	50%	59%	67%	0.5
	30Hz	50%	59%	67%	0.5
	12Hz	50%	59%	67%	0.5
	6Hz	50%	59%	67%	0.5
	3Hz	49%	58%	66%	0.5
	2Hz	51%	60%	68%	0.5
	1Hz	49%	57%	65%	0.5
	.5Hz	45%	55%	64%	0.4
	.25Hz	51%	64%	76%	0.6
	.1Hz	56%	65%	72%	0.5
90%	60Hz	46%	60%	72%	0.6
	30Hz	46%	60%	72%	0.6
	12Hz	46%	60%	72%	0.6
	6Hz	46%	60%	72%	0.6
	3Hz	44%	58%	71%	0.6
	2Hz	44%	58%	71%	0.6
	1Hz	44%	58%	71%	0.6

continued on next page

Table B.23.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	27%	42%	58%	0.6
	.25Hz	45%	67%	84%	0.6
	.1Hz	52%	66%	78%	0.6

General Linear Model : Keypress Data

Table B.24.: General Linear Model using Keypress Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
10%	60Hz	55%	59%	63%	0.44
	30Hz	56%	60%	65%	0.23
	12Hz	51%	56%	60%	0.92
	6Hz	53%	58%	62%	0.68
	3Hz	55%	60%	64%	0.71
	2Hz	55%	59%	64%	0.37
	1Hz	49%	54%	58%	0.93
	.5Hz	39%	44%	49%	1.0
	.25Hz	44%	50%	57%	1.0
	.1Hz	52%	56%	60%	1.0
	60Hz	53%	58%	63%	0.6
	30Hz	53%	58%	63%	0.7
	12Hz	52%	56%	61%	0.8
	6Hz	54%	59%	64%	0.4
	3Hz	53%	58%	63%	0.5
	2Hz	49%	54%	58%	1.0
	1Hz	54%	58%	63%	0.3
	.5Hz	39%	44%	50%	0.2
	.25Hz	48%	56%	63%	1.0
	.1Hz	66%	71%	75%	0.0
50%	60Hz	48%	54%	60%	1.0
	30Hz	59%	65%	70%	0.0
	12Hz	50%	56%	62%	0.8

continued on next page

Table B.24.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	6Hz	60%	66%	71%	0.0
	3Hz	58%	63%	69%	0.0
	2Hz	51%	57%	62%	0.6
	1Hz	60%	66%	72%	0.0
	.5Hz	52%	58%	65%	0.1
	.25Hz	48%	57%	66%	0.9
	.1Hz	63%	68%	74%	0.1
75%	60Hz	50%	58%	66%	0.6
	30Hz	50%	59%	67%	0.5
	12Hz	50%	59%	67%	0.5
	6Hz	49%	58%	66%	0.6
	3Hz	49%	58%	66%	0.5
	2Hz	58%	66%	74%	0.0
	1Hz	53%	61%	70%	0.2
	.5Hz	50%	59%	69%	0.1
	.25Hz	50%	63%	74%	0.7
	.1Hz	67%	75%	82%	0.0
90%	60Hz	46%	60%	72%	0.6
	30Hz	46%	60%	72%	0.6
	12Hz	46%	60%	72%	0.6
	6Hz	46%	60%	72%	0.6
	3Hz	44%	58%	71%	0.6
	2Hz	46%	60%	72%	0.4
	1Hz	44%	58%	71%	0.6
	.5Hz	33%	49%	65%	0.2

continued on next page

Table B.24.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.25Hz	35%	56%	76%	0.9
	.1Hz	54%	67%	79%	0.5

General Linear Model : Keypress and Experience Data

Table B.25.: General Linear Model using Keypress and Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
10%	60Hz	50%	54%	58%	0.99
	30Hz	46%	50%	55%	1.00
	12Hz	50%	54%	58%	0.98
	6Hz	45%	49%	54%	1.00
	3Hz	48%	53%	57%	0.99
	2Hz	48%	52%	57%	0.99
	1Hz	49%	53%	57%	0.97
	.5Hz	43%	48%	53%	0.97
	.25Hz	42%	49%	55%	1.00
	.1Hz	48%	53%	57%	1.00
25%	60Hz	54%	58%	63%	0.60
	30Hz	53%	57%	62%	0.74
	12Hz	51%	56%	61%	0.89
	6Hz	52%	57%	62%	0.74
	3Hz	50%	55%	60%	0.89
	2Hz	54%	59%	63%	0.30
	1Hz	54%	59%	64%	0.18
	.5Hz	34%	39%	44%	0.88
	.25Hz	48%	55%	62%	0.99
	.1Hz	61%	65%	70%	0.40
50%	60Hz	53%	59%	65%	0.53
	30Hz	54%	60%	66%	0.38

continued on next page

Table B.25.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	54%	60%	65%	0.39
	6Hz	52%	58%	64%	0.62
	3Hz	57%	63%	68%	0.04
	2Hz	55%	60%	66%	0.16
	1Hz	51%	57%	63%	0.57
	.5Hz	47%	54%	60%	0.42
	.25Hz	51%	60%	69%	0.80
	.1Hz	63%	69%	74%	0.06
75%	60Hz	50%	58%	66%	0.60
	30Hz	51%	60%	68%	0.47
	12Hz	53%	61%	69%	0.28
	6Hz	47%	55%	63%	0.82
	3Hz	55%	63%	71%	0.10
	2Hz	54%	62%	70%	0.14
	1Hz	54%	62%	70%	0.13
	.5Hz	36%	45%	55%	0.25
	.25Hz	42%	55%	67%	0.95
	.1Hz	61%	69%	76%	0.17
90%	60Hz	42%	56%	69%	0.75
	30Hz	47%	60%	73%	0.56
	12Hz	46%	60%	72%	0.56
	6Hz	42%	56%	69%	0.75
	3Hz	49%	63%	76%	0.25
	2Hz	58%	72%	83%	0.07
	1Hz	41%	55%	68%	0.75

continued on next page

Table B.25.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	38%	53%	69%	0.56
	.25Hz	46%	68%	85%	0.43
	.1Hz	61%	74%	85%	0.11

General Linear Model : Keypress and Interaction Time Data

Table B.26.: General Linear Model using Keypress and Interaction Time Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
10%	60Hz	46%	50%	55%	1.00
	30Hz	44%	48%	52%	1.00
	12Hz	35%	39%	44%	1.00
	6Hz	42%	47%	51%	1.00
	3Hz	48%	53%	57%	1.00
	2Hz	41%	46%	50%	1.00
	1Hz	39%	44%	48%	1.00
	.5Hz	38%	43%	48%	1.00
	.25Hz	44%	50%	57%	1.00
	.1Hz	51%	55%	60%	1.00
25%	60Hz	53%	58%	62%	0.82
	30Hz	47%	52%	57%	1.00
	12Hz	49%	54%	59%	0.98
	6Hz	53%	57%	62%	0.70
	3Hz	48%	53%	57%	0.98
	2Hz	50%	55%	60%	0.87
	1Hz	50%	55%	60%	0.97
	.5Hz	43%	49%	54%	0.93
	.25Hz	38%	45%	53%	1.00
	.1Hz	58%	63%	68%	0.74
50%	60Hz	59%	65%	70%	0.02
	30Hz	54%	60%	66%	0.34

continued on next page

Table B.26.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	57%	62%	68%	0.11
	6Hz	55%	61%	66%	0.34
	3Hz	51%	57%	62%	0.66
	2Hz	51%	57%	62%	0.62
	1Hz	51%	57%	63%	0.48
	.5Hz	51%	58%	64%	0.08
	.25Hz	45%	54%	63%	0.99
	.1Hz	64%	70%	75%	0.03
75%	60Hz	54%	62%	70%	0.23
	30Hz	56%	64%	72%	0.10
	12Hz	57%	65%	73%	0.10
	6Hz	60%	68%	75%	0.01
	3Hz	59%	67%	75%	0.01
	2Hz	59%	67%	75%	0.03
	1Hz	60%	69%	76%	0.02
	.5Hz	43%	52%	62%	0.61
	.25Hz	39%	52%	64%	0.99
	.1Hz	60%	68%	75%	0.22
90%	60Hz	51%	65%	77%	0.25
	30Hz	55%	68%	80%	0.11
	12Hz	48%	61%	74%	0.45
	6Hz	55%	68%	80%	0.11
	3Hz	55%	68%	80%	0.17
	2Hz	44%	58%	71%	0.75
	1Hz	48%	62%	75%	0.45

continued on next page

Table B.26.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	42%	58%	73%	0.32
	.25Hz	39%	60%	79%	0.74
	.1Hz	59%	72%	83%	0.17

General Linear Model : Keypress, Interaction Time and Experience Data

Table B.27.: General Linear Model using Keypress, Interaction Time and Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
10%	60Hz	40%	45%	49%	1.0
	30Hz	37%	41%	45%	1.0
	12Hz	46%	51%	55%	1.0
	6Hz	34%	38%	43%	1.0
	3Hz	42%	46%	51%	1.0
	2Hz	43%	48%	52%	1.0
	1Hz	43%	47%	52%	1.0
	.5Hz	38%	43%	48%	1.0
	.25Hz	39%	46%	52%	1.0
	.1Hz	48%	53%	57%	1.0
25%	60Hz	44%	49%	53%	1.0
	30Hz	52%	56%	61%	0.9
	12Hz	51%	56%	61%	0.9
	6Hz	50%	54%	59%	1.0
	3Hz	54%	58%	63%	0.8
	2Hz	49%	54%	59%	0.9
	1Hz	47%	52%	56%	1.0
	.5Hz	42%	47%	53%	1.0
	.25Hz	46%	54%	61%	1.0
	.1Hz	59%	63%	68%	0.7
50%	60Hz	49%	55%	61%	0.9
	30Hz	51%	56%	62%	0.8

continued on next page

Table B.27.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	56%	62%	67%	0.2
	6Hz	55%	61%	66%	0.3
	3Hz	53%	59%	65%	0.3
	2Hz	51%	57%	63%	0.6
	1Hz	55%	61%	66%	0.4
	.5Hz	40%	47%	53%	1.0
	.25Hz	48%	57%	66%	0.9
	.1Hz	63%	68%	74%	0.1
75%	60Hz	60%	68%	75%	0.0
	30Hz	54%	62%	70%	0.2
	12Hz	54%	62%	70%	0.2
	6Hz	59%	67%	75%	0.0
	3Hz	56%	64%	72%	0.1
	2Hz	56%	65%	72%	0.1
	1Hz	57%	66%	74%	0.1
	.5Hz	47%	57%	66%	0.3
	.25Hz	54%	67%	78%	0.4
	.1Hz	65%	73%	80%	0.0
90%	60Hz	51%	65%	77%	0.3
	30Hz	46%	60%	72%	0.6
	12Hz	51%	65%	77%	0.3
	6Hz	51%	65%	77%	0.3
	3Hz	48%	61%	74%	0.3
	2Hz	51%	65%	77%	0.2
	1Hz	48%	62%	75%	0.3

continued on next page

Table B.27.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	35%	51%	67%	0.1
	.25Hz	43%	64%	82%	0.6
	.1Hz	59%	72%	83%	0.2

B.2.2 Neural Networks

Neural Network : Experience Data

Table B.28.: Neural Network using Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	55%	59%	63%	0.48
	30Hz	55%	60%	64%	0.52
	12Hz	54%	59%	63%	0.52
	6Hz	54%	58%	62%	0.66
	3Hz	43%	47%	52%	1.00
	2Hz	52%	56%	61%	0.69
	1Hz	52%	57%	61%	0.52
	.5Hz	37%	42%	47%	0.48
	.25Hz	47%	53%	60%	1.00
	.1Hz	60%	64%	68%	0.52
0.25	60Hz	54%	59%	63%	0.52
	30Hz	52%	57%	62%	0.89
	12Hz	54%	59%	63%	0.52
	6Hz	54%	59%	63%	0.52
	3Hz	45%	50%	55%	1.00
	2Hz	50%	54%	59%	0.99
	1Hz	49%	53%	58%	0.94
	.5Hz	49%	55%	60%	0.27
	.25Hz	56%	63%	70%	0.53
	.1Hz	59%	64%	68%	0.64
0.50	60Hz	53%	59%	65%	0.53
	30Hz	51%	56%	62%	0.81
	12Hz	53%	59%	64%	0.52

continued on next page

Table B.28.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	6Hz	51%	57%	63%	0.74
	3Hz	52%	58%	63%	0.52
	2Hz	46%	52%	58%	0.97
	1Hz	48%	54%	60%	0.85
	.5Hz	45%	52%	58%	0.68
	.25Hz	51%	60%	69%	0.80
	.1Hz	59%	64%	70%	0.53
0.75	60Hz	50%	59%	67%	0.54
	30Hz	50%	59%	67%	0.54
	12Hz	51%	59%	67%	0.47
	6Hz	49%	57%	65%	0.66
	3Hz	47%	55%	64%	0.72
	2Hz	51%	60%	68%	0.54
	1Hz	49%	57%	65%	0.54
	.5Hz	49%	59%	68%	0.15
	.25Hz	43%	56%	69%	0.92
	.1Hz	56%	65%	72%	0.54
0.90	60Hz	46%	60%	72%	0.56
	30Hz	46%	60%	72%	0.56
	12Hz	42%	56%	69%	0.75
	6Hz	46%	60%	72%	0.56
	3Hz	44%	58%	71%	0.56
	2Hz	42%	56%	69%	0.66
	1Hz	44%	58%	71%	0.56
	.5Hz	25%	40%	56%	0.68

continued on next page

Table B.28.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.25Hz	37%	58%	78%	0.86
	.1Hz	52%	66%	78%	0.56

Neural Network : Keypress Data

Table B.29.: Neural Network using Keypress Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	56%	60%	64%	0.28
	30Hz	59%	63%	67%	0.03
	12Hz	49%	53%	58%	1.00
	6Hz	45%	49%	54%	1.00
	3Hz	60%	64%	68%	0.07
	2Hz	62%	66%	70%	$p \leq 0.001$
	1Hz	61%	65%	69%	$p \leq 0.001$
	.5Hz	45%	50%	55%	0.40
	.25Hz	57%	63%	70%	0.53
	.1Hz	65%	69%	73%	0.01
$p \leq 0.001$	60Hz	60%	64%	69%	0.01
	30Hz	54%	59%	63%	0.52
	12Hz	61%	66%	70%	$p \leq 0.001$
	6Hz	60%	65%	69%	$p \leq 0.001$
	3Hz	60%	65%	69%	$p \leq 0.001$
	2Hz	60%	65%	69%	$p \leq 0.001$
	1Hz	60%	64%	69%	$p \leq 0.001$
	.5Hz	44%	49%	55%	$p \leq 0.001$
	.25Hz	43%	50%	57%	1.00
	.1Hz	67%	71%	75%	$p \leq 0.001$
0.50	60Hz	60%	65%	71%	0.01
	30Hz	59%	65%	70%	0.02
	12Hz	61%	66%	72%	$p \leq 0.001$

continued on next page

Table B.29.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	6Hz	60%	66%	71%	0.01
	3Hz	58%	64%	70%	0.01
	2Hz	58%	64%	69%	0.01
	1Hz	59%	65%	71%	0.04
	.5Hz	49%	56%	63%	0.19
	.25Hz	55%	64%	72%	0.54
	.1Hz	65%	71%	76%	0.01
0.75	60Hz	55%	64%	71%	0.14
	30Hz	57%	66%	73%	0.05
	12Hz	56%	65%	72%	0.08
	6Hz	54%	63%	70%	0.18
	3Hz	57%	65%	73%	0.04
	2Hz	59%	67%	75%	0.03
	1Hz	55%	64%	72%	0.07
	.5Hz	51%	60%	70%	0.08
	.25Hz	51%	64%	76%	0.56
	.1Hz	62%	70%	77%	0.10
0.90	60Hz	58%	72%	83%	0.04
	30Hz	53%	67%	79%	0.17
	12Hz	58%	72%	83%	0.04
	6Hz	51%	65%	77%	0.25
	3Hz	49%	63%	76%	0.25
	2Hz	48%	61%	74%	0.35
	1Hz	44%	58%	71%	0.56
	.5Hz	42%	58%	73%	0.02

continued on next page

Table B.29.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.25Hz	46%	68%	85%	0.43
	.1Hz	48%	62%	74%	0.76

Neural Network : Keypress and Experience Data

Table B.30.: Neural Network using Keypress and Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	54%	58%	63%	0.62
	30Hz	54%	58%	62%	0.66
	12Hz	54%	59%	63%	0.52
	6Hz	42%	47%	51%	1.00
	3Hz	53%	57%	61%	0.65
	2Hz	48%	52%	57%	0.99
	1Hz	46%	50%	55%	1.00
	.5Hz	55%	60%	65%	$p \leq 0.001$
	.25Hz	44%	50%	57%	1.00
	.1Hz	66%	70%	74%	$p \leq 0.001$
0.25	60Hz	52%	56%	61%	0.87
	30Hz	62%	67%	71%	$p \leq 0.001$
	12Hz	60%	65%	69%	$p \leq 0.001$
	6Hz	60%	65%	69%	$p \leq 0.001$
	3Hz	45%	50%	55%	1.00
	2Hz	59%	64%	69%	$p \leq 0.001$
	1Hz	52%	57%	62%	0.56
	.5Hz	39%	45%	50%	0.15
	.25Hz	56%	63%	70%	0.53
	.1Hz	64%	69%	73%	0.03
0.50	60Hz	60%	66%	71%	0.01
	30Hz	57%	63%	68%	0.09

continued on next page

Table B.30.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	53%	59%	64%	0.52
	6Hz	56%	62%	68%	0.11
	3Hz	61%	67%	72%	$p \leq 0.001$
	2Hz	56%	62%	67%	0.07
	1Hz	58%	64%	69%	0.01
	.5Hz	52%	59%	66%	0.03
	.25Hz	55%	64%	72%	0.54
	.1Hz	66%	71%	76%	0.01
0.75	60Hz	57%	65%	73%	0.08
	30Hz	56%	64%	72%	0.10
	12Hz	56%	65%	72%	0.08
	6Hz	60%	68%	75%	0.01
	3Hz	57%	65%	73%	0.04
	2Hz	51%	60%	68%	0.34
	1Hz	56%	65%	73%	0.04
	.5Hz	41%	51%	61%	0.03
	.25Hz	51%	64%	76%	0.56
	.1Hz	61%	69%	77%	0.13
0.90	60Hz	51%	65%	77%	0.25
	30Hz	57%	71%	82%	0.07
	12Hz	51%	65%	77%	0.25
	6Hz	42%	56%	69%	0.75
	3Hz	60%	74%	84%	0.01
	2Hz	60%	74%	84%	0.04
	1Hz	48%	62%	75%	0.34

continued on next page

Table B.30.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	31%	47%	62%	0.86
	.25Hz	43%	64%	82%	0.59
	.1Hz	63%	76%	86%	0.06

Neural Network : Keypress and Interaction Time Data

Table B.31.: Neural Network using Keypress and Interaction Time Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	47%	51%	55%	1.00
	30Hz	51%	55%	60%	0.95
	12Hz	41%	45%	50%	1.00
	6Hz	50%	54%	58%	0.99
	3Hz	52%	56%	60%	0.99
	2Hz	53%	57%	61%	0.59
	1Hz	44%	49%	53%	1.00
	.5Hz	57%	62%	67%	$p \leq 0.001$
	.25Hz	49%	55%	62%	0.99
	.1Hz	58%	62%	66%	0.89
0.25	60Hz	60%	64%	69%	0.03
	30Hz	55%	60%	64%	0.37
	12Hz	53%	58%	63%	0.60
	6Hz	56%	61%	65%	0.21
	3Hz	53%	58%	62%	0.56
	2Hz	60%	64%	69%	$p \leq 0.001$
	1Hz	59%	64%	68%	0.05
	.5Hz	54%	60%	65%	0.01
	.25Hz	48%	56%	63%	0.99
	.1Hz	67%	71%	76%	$p \leq 0.001$
0.50	60Hz	56%	62%	67%	0.19
	30Hz	59%	65%	70%	0.02

continued on next page

Table B.31.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	59%	65%	71%	0.01
	6Hz	58%	64%	69%	0.05
	3Hz	57%	63%	69%	0.03
	2Hz	57%	63%	69%	0.03
	1Hz	55%	61%	67%	0.08
	.5Hz	54%	61%	67%	0.01
	.25Hz	50%	59%	67%	0.88
	.1Hz	64%	70%	75%	0.03
0.75	60Hz	57%	66%	73%	0.05
	30Hz	56%	64%	72%	0.10
	12Hz	59%	67%	75%	0.04
	6Hz	62%	70%	77%	$p \leq 0.001$
	3Hz	49%	58%	66%	0.54
	2Hz	53%	62%	70%	0.28
	1Hz	60%	69%	76%	0.02
	.5Hz	46%	56%	65%	0.32
	.25Hz	43%	56%	69%	0.92
	.1Hz	65%	73%	80%	0.02
0.90	60Hz	53%	67%	79%	0.17
	30Hz	53%	67%	79%	0.17
	12Hz	44%	58%	71%	0.66
	6Hz	57%	70%	82%	0.07
	3Hz	58%	72%	83%	0.07
	2Hz	53%	67%	79%	0.25
	1Hz	46%	60%	73%	0.56

continued on next page

Table B.31.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	44%	60%	75%	0.22
	.25Hz	39%	60%	79%	0.74
	.1Hz	61%	74%	85%	0.11

Neural Network : Keypress, Interaction Time and Experience Data

Table B.32.: Neural Network using Keypress, Interaction Time and Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	46%	50%	55%	1.00
	30Hz	42%	47%	51%	1.00
	12Hz	61%	65%	69%	$p \leq 0.001$
	6Hz	55%	59%	64%	0.38
	3Hz	43%	47%	51%	1.00
	2Hz	54%	58%	62%	0.38
	1Hz	54%	59%	63%	0.20
	.5Hz	51%	56%	61%	0.09
	.25Hz	43%	50%	57%	1.00
	.1Hz	59%	63%	67%	0.78
0.25	60Hz	60%	64%	69%	0.01
	30Hz	52%	57%	61%	0.84
	12Hz	51%	56%	61%	0.89
	6Hz	54%	59%	63%	0.52
	3Hz	64%	68%	73%	$p \leq 0.001$
	2Hz	56%	61%	65%	0.10
	1Hz	57%	62%	67%	0.02
	.5Hz	30%	35%	41%	1.00
	.25Hz	42%	49%	56%	1.00
	.1Hz	67%	72%	76%	$p \leq 0.001$
0.50	60Hz	57%	63%	68%	0.11
	30Hz	53%	59%	64%	0.57

continued on next page

Table B.32.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	58%	63%	69%	0.05
	6Hz	60%	65%	71%	0.01
	3Hz	58%	64%	70%	0.01
	2Hz	58%	64%	69%	0.01
	1Hz	58%	64%	70%	0.08
	.5Hz	58%	65%	71%	$p \leq 0.001$
	.25Hz	45%	53%	62%	0.99
	.1Hz	67%	73%	78%	$p \leq 0.001$
0.75	60Hz	60%	68%	75%	0.02
	30Hz	55%	64%	71%	0.14
	12Hz	55%	63%	71%	0.14
	6Hz	61%	69%	76%	0.01
	3Hz	55%	64%	71%	0.08
	2Hz	56%	64%	72%	0.18
	1Hz	56%	65%	73%	0.13
	.5Hz	48%	58%	67%	0.20
	.25Hz	43%	56%	69%	0.92
	.1Hz	66%	73%	80%	0.01
0.90	60Hz	53%	67%	79%	0.17
	30Hz	49%	63%	76%	0.35
	12Hz	51%	65%	77%	0.25
	6Hz	55%	68%	80%	0.11
	3Hz	51%	65%	77%	0.17
	2Hz	57%	70%	82%	0.04
	1Hz	51%	65%	78%	0.17

continued on next page

Table B.32.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	27%	42%	58%	0.56
	.25Hz	35%	56%	76%	0.85
	.1Hz	55%	69%	80%	0.34

B.2.3 Random Forests

Random Forest : Experience Data

Table B.33.: Random Forest using Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	51%	55%	59%	0.96
	30Hz	50%	54%	58%	1.00
	12Hz	51%	55%	60%	0.95
	6Hz	38%	43%	47%	1.00
	3Hz	45%	50%	54%	1.00
	2Hz	50%	54%	59%	0.93
	1Hz	43%	47%	52%	1.00
	.5Hz	35%	39%	44%	0.86
	.25Hz	35%	42%	48%	1.00
	.1Hz	56%	60%	65%	0.97
0.25	60Hz	52%	57%	61%	0.82
	30Hz	49%	54%	58%	1.00
	12Hz	52%	57%	61%	0.82
	6Hz	47%	52%	56%	1.00
	3Hz	45%	50%	55%	1.00
	2Hz	52%	56%	61%	0.93
	1Hz	44%	49%	54%	1.00
	.5Hz	47%	53%	58%	0.57
	.25Hz	47%	55%	62%	1.00
	.1Hz	55%	60%	64%	0.98
0.50	60Hz	52%	57%	63%	0.70
	30Hz	51%	57%	63%	0.74
	12Hz	46%	52%	58%	0.99

continued on next page

Table B.33.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	6Hz	49%	54%	60%	0.93
	3Hz	50%	56%	62%	0.70
	2Hz	46%	52%	58%	0.97
	1Hz	47%	53%	59%	0.93
	.5Hz	46%	52%	59%	0.58
	.25Hz	47%	56%	65%	0.97
	.1Hz	57%	62%	68%	0.79
0.75	60Hz	48%	57%	65%	0.72
	30Hz	50%	58%	66%	0.60
	12Hz	48%	56%	65%	0.72
	6Hz	48%	56%	65%	0.72
	3Hz	48%	57%	65%	0.60
	2Hz	51%	60%	68%	0.54
	1Hz	46%	54%	63%	0.78
	.5Hz	43%	53%	63%	0.54
	.25Hz	43%	56%	69%	0.92
	.1Hz	52%	61%	68%	0.87
0.90	60Hz	41%	54%	68%	0.83
	30Hz	46%	60%	72%	0.56
	12Hz	39%	53%	66%	0.89
	6Hz	42%	56%	69%	0.75
	3Hz	41%	54%	68%	0.75
	2Hz	42%	56%	69%	0.66
	1Hz	41%	55%	68%	0.75
	.5Hz	27%	42%	58%	0.56

continued on next page

Table B.33.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.25Hz	29%	50%	71%	0.97
	.1Hz	54%	67%	79%	0.45

Random Forest : Keypress Data

Table B.34.: Random Forest using Keypress Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	57%	61%	65%	0.13
	30Hz	57%	62%	66%	0.12
	12Hz	58%	62%	67%	0.04
	6Hz	54%	58%	62%	0.59
	3Hz	61%	66%	70%	0.01
	2Hz	55%	59%	63%	0.45
	1Hz	55%	60%	64%	0.10
	.5Hz	45%	50%	55%	0.40
	.25Hz	47%	53%	60%	1.00
	.1Hz	63%	67%	71%	0.08
$p \leq 0.001$	60Hz	55%	60%	64%	0.41
	30Hz	58%	62%	67%	0.08
	12Hz	59%	64%	68%	0.02
	6Hz	58%	62%	67%	0.05
	3Hz	55%	59%	64%	0.27
	2Hz	49%	54%	59%	0.92
	1Hz	53%	58%	62%	0.40
	.5Hz	37%	42%	47%	0.52
	.25Hz	54%	61%	68%	0.75
	.1Hz	64%	69%	73%	0.04
0.50	60Hz	58%	64%	69%	0.05
	30Hz	52%	58%	64%	0.62
	12Hz	58%	64%	70%	0.03

continued on next page

Table B.34.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	6Hz	56%	62%	67%	0.13
	3Hz	56%	62%	68%	0.07
	2Hz	57%	62%	68%	0.04
	1Hz	58%	64%	69%	0.10
	.5Hz	51%	57%	64%	0.10
	.25Hz	55%	64%	72%	0.54
	.1Hz	64%	69%	75%	0.04
0.75	60Hz	55%	63%	71%	0.18
	30Hz	60%	68%	75%	0.02
	12Hz	53%	61%	69%	0.28
	6Hz	54%	62%	70%	0.23
	3Hz	54%	62%	70%	0.14
	2Hz	53%	61%	69%	0.34
	1Hz	50%	59%	67%	0.40
	.5Hz	49%	59%	68%	0.15
	.25Hz	46%	59%	71%	0.82
	.1Hz	66%	74%	81%	0.01
0.90	60Hz	55%	68%	80%	0.11
	30Hz	53%	67%	79%	0.17
	12Hz	55%	68%	80%	0.11
	6Hz	57%	70%	82%	0.07
	3Hz	46%	60%	72%	0.45
	2Hz	44%	58%	71%	0.56
	1Hz	44%	58%	71%	0.56
	.5Hz	29%	44%	60%	0.44

continued on next page

Table B.34.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.25Hz	39%	60%	79%	0.74
	.1Hz	52%	66%	78%	0.56

Random Forest : Keypress and Experience Data

Table B.35.: Random Forest using Keypress and Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	56%	60%	65%	0.25
	30Hz	60%	64%	68%	0.01
	12Hz	59%	63%	67%	0.02
	6Hz	61%	65%	69%	0.01
	3Hz	56%	60%	64%	0.14
	2Hz	56%	60%	64%	0.10
	1Hz	52%	56%	61%	0.62
	.5Hz	53%	58%	63%	0.02
	.25Hz	39%	46%	52%	1.00
	.1Hz	66%	70%	74%	$p \leq 0.001$
0.25	60Hz	58%	62%	67%	0.07
	30Hz	56%	61%	66%	0.18
	12Hz	57%	62%	66%	0.11
	6Hz	58%	63%	67%	0.04
	3Hz	55%	60%	65%	0.16
	2Hz	58%	63%	67%	0.01
	1Hz	57%	62%	66%	0.03
	.5Hz	38%	43%	49%	0.35
	.25Hz	46%	54%	61%	1.00
	.1Hz	64%	68%	73%	0.05
0.50	60Hz	57%	63%	68%	0.11
	30Hz	57%	63%	69%	0.07

continued on next page

Table B.35.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	57%	62%	68%	0.11
	6Hz	61%	67%	72%	$p \leq 0.001$
	3Hz	57%	63%	68%	0.04
	2Hz	56%	62%	68%	0.05
	1Hz	54%	60%	66%	0.18
	.5Hz	48%	55%	62%	0.27
	.25Hz	49%	58%	67%	0.91
	.1Hz	65%	71%	76%	0.02
0.75	60Hz	56%	64%	72%	0.10
	30Hz	53%	62%	70%	0.28
	12Hz	53%	61%	69%	0.28
	6Hz	53%	61%	69%	0.28
	3Hz	56%	64%	72%	0.05
	2Hz	57%	66%	73%	0.03
	1Hz	54%	63%	71%	0.10
	.5Hz	35%	45%	54%	0.31
	.25Hz	48%	61%	73%	0.74
	.1Hz	62%	70%	77%	0.10
0.90	60Hz	41%	54%	68%	0.83
	30Hz	55%	69%	80%	0.11
	12Hz	53%	67%	79%	0.17
	6Hz	46%	60%	72%	0.56
	3Hz	62%	75%	86%	$p \leq 0.001$
	2Hz	62%	75%	86%	0.02
	1Hz	48%	62%	75%	0.34

continued on next page

Table B.35.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	42%	58%	73%	0.32
	.25Hz	39%	60%	79%	0.74
	.1Hz	61%	74%	85%	0.11

Random Forest : Keypress and Interaction Time Data

Table B.36.: Random Forest using Keypress and Interaction Time Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	56%	61%	65%	0.23
	30Hz	61%	65%	69%	$p \leq 0.001$
	12Hz	55%	59%	63%	0.45
	6Hz	58%	62%	66%	0.05
	3Hz	62%	66%	70%	0.01
	2Hz	57%	61%	65%	0.05
	1Hz	57%	61%	65%	0.03
	.5Hz	54%	59%	64%	0.01
	.25Hz	53%	59%	66%	0.90
	.1Hz	66%	70%	74%	$p \leq 0.001$
0.25	60Hz	60%	65%	70%	0.01
	30Hz	59%	64%	69%	0.01
	12Hz	60%	64%	69%	0.01
	6Hz	60%	65%	69%	$p \leq 0.001$
	3Hz	58%	62%	67%	0.03
	2Hz	52%	57%	62%	0.56
	1Hz	59%	64%	68%	0.05
	.5Hz	55%	60%	65%	$p \leq 0.001$
	.25Hz	51%	59%	66%	0.92
	.1Hz	61%	66%	70%	0.26
0.50	60Hz	63%	68%	74%	$p \leq 0.001$
	30Hz	61%	67%	72%	$p \leq 0.001$

continued on next page

Table B.36.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	60%	65%	71%	0.01
	6Hz	57%	63%	69%	0.09
	3Hz	60%	66%	71%	$p \leq 0.001$
	2Hz	56%	62%	68%	0.05
	1Hz	60%	66%	71%	$p \leq 0.001$
	.5Hz	56%	63%	69%	$p \leq 0.001$
	.25Hz	53%	62%	70%	0.68
	.1Hz	64%	69%	75%	0.04
0.75	60Hz	57%	66%	73%	0.05
	30Hz	57%	66%	73%	0.05
	12Hz	57%	66%	73%	0.07
	6Hz	64%	72%	79%	$p \leq 0.001$
	3Hz	57%	66%	73%	0.03
	2Hz	57%	65%	73%	0.08
	1Hz	63%	71%	79%	$p \leq 0.001$
	.5Hz	49%	59%	68%	0.15
	.25Hz	45%	58%	70%	0.88
	.1Hz	61%	69%	76%	0.17
0.90	60Hz	53%	67%	79%	0.17
	30Hz	51%	65%	77%	0.25
	12Hz	49%	63%	76%	0.35
	6Hz	57%	70%	82%	0.07
	3Hz	60%	74%	84%	0.04
	2Hz	58%	72%	83%	0.07
	1Hz	55%	69%	81%	0.11

continued on next page

Table B.36.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	38%	53%	69%	0.56
	.25Hz	46%	68%	85%	0.43
	.1Hz	67%	79%	89%	0.02

Random Forest : Keypress, Interaction Time Data

Table B.37.: Random Forest using Keypress, Interaction Time and Experience Data

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
0.10	60Hz	59%	63%	67%	0.02
	30Hz	60%	65%	69%	$p \leq 0.001$
	12Hz	58%	63%	67%	0.07
	6Hz	54%	58%	63%	0.55
	3Hz	55%	60%	64%	0.18
	2Hz	55%	60%	64%	0.16
	1Hz	58%	63%	67%	0.01
	.5Hz	50%	55%	60%	0.26
	.25Hz	52%	58%	65%	0.95
	.1Hz	68%	72%	76%	$p \leq 0.001$
0.25	60Hz	61%	66%	70%	$p \leq 0.001$
	30Hz	56%	60%	65%	0.27
	12Hz	56%	60%	65%	0.23
	6Hz	62%	67%	71%	$p \leq 0.001$
	3Hz	61%	66%	70%	0.01
	2Hz	59%	64%	69%	$p \leq 0.001$
	1Hz	56%	61%	66%	0.05
	.5Hz	54%	60%	65%	0.01
	.25Hz	53%	60%	67%	0.83
	.1Hz	66%	71%	75%	$p \leq 0.001$
0.50	60Hz	60%	65%	71%	0.01
	30Hz	57%	63%	69%	0.07

continued on next page

Table B.37.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	12Hz	62%	67%	73%	$p \leq 0.001$
	6Hz	61%	67%	72%	$p \leq 0.001$
	3Hz	58%	64%	70%	0.01
	2Hz	57%	62%	68%	0.04
	1Hz	62%	68%	73%	$p \leq 0.001$
	.5Hz	53%	60%	67%	0.02
	.25Hz	51%	60%	68%	0.84
	.1Hz	69%	75%	79%	$p \leq 0.001$
0.75	60Hz	60%	68%	76%	0.01
	30Hz	57%	65%	73%	0.08
	12Hz	58%	67%	74%	0.03
	6Hz	60%	68%	75%	0.01
	3Hz	56%	64%	72%	0.05
	2Hz	58%	67%	74%	0.05
	1Hz	60%	69%	76%	0.02
	.5Hz	55%	65%	74%	0.01
	.25Hz	46%	59%	71%	0.82
	.1Hz	67%	75%	82%	0.01
0.90	60Hz	57%	70%	82%	0.07
	30Hz	46%	60%	72%	0.56
	12Hz	53%	67%	79%	0.17
	6Hz	55%	68%	80%	0.11
	3Hz	48%	61%	74%	0.35
	2Hz	53%	67%	79%	0.11
	1Hz	50%	64%	76%	0.25

continued on next page

Table B.37.: *continued*

Split	Sampling Rate	Lower Bound	Accuracy	Upper Bound	P-values
	.5Hz	31%	47%	62%	0.32
	.25Hz	46%	68%	85%	0.43
	.1Hz	55%	69%	80%	0.34

C. ADDITIONAL CONTEXT

The following chapter provides additional academic context not covered in the Background Chapter.

C.1 Human Factors

The following section provides a brief overview of three interaction domains under consideration in this exam and specific human factors related to that domain. The first section, titled Aviation and HFACS, discusses the formulation of the Human Factors analysis and classification system because of aviation accident investigations. The second section, uses the interaction domain of driving to introduce the concept of mental models. The third and final section use human computer interaction and the sub-domain of data visualization to introduce the language of data.

C.1.1 Aviation and HFACS

Human Factors analysis and classification system (HFACS) is used by analysts to classify the human causes of an accident. The results of an HFACS analysis can be used to improve training and develop accident prevention procedures(Aviation, 2000). The model is based on James Reasons Swiss Cheese Model of accident causation (Reason, 1990). The HFACS model was originally developed for commercial and military aviation but has been applied to: Air traffic control (Broach & Dollar, 2002) , Aviation maintenance (Krulak, 2004; Rashid, Place, & Braithwaite, 2014), construction (Garrett & Teizer, 2009), healthcare (ElBardissi, Wiegmann, Dearani, Daly, & Sundt, 2007) and mining operations (Patterson & Shappell, 2010). Figure C.1shows the four level of analysis considered during an HFACS analysis. These lev-

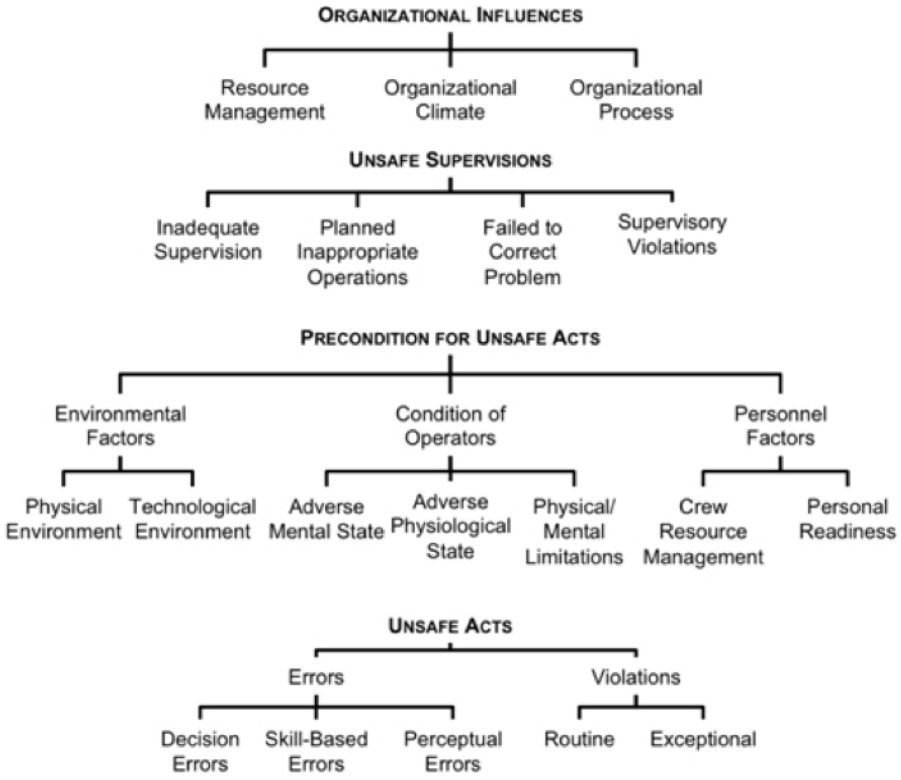


Fig. C.1. HFACS Framework courtesy skybrary.com

els, Unsafe acts, Preconditions for unsafe acts, unsafe supervision and organization influence, are directly based on Reason's Swiss Cheese model of accident causation. Analysts consider incidents and accidents using a bottom up holistic approach.

A bottom up approach allows an analyst to work up the chain of causation. Take for instance a ground vehicle which collides with a parked aircraft. This unsafe act can be simultaneously a decision error, a perceptual error and a violation. The ground crew driver may be used to taking a short cut with a small vehicle and not realize they were in a vehicle with a taller clearance which lead to this incident.

HFACS asserts that errors do not occur spontaneously but are a product of pre-existing environmental and operational conditions. The driver may have felt operational pressure to take the short cut because the ground crew was under staffed and they were needed elsewhere. Alternatively, a clearance alarm alerting the pi-

lot may not have sounded. At higher levels, unsafe supervisions and organizational influences can increase the likelihood of incidents. An organizational climate which stresses working long hours buts down plays the importance of retraining employees combined with poor supervision can increase the likelihood of errors.

Procedural Non-compliance

What are the reasons trained professionals who are aware that a procedure exists fail to comply with a procedure? According to Carthey et al. in a study of medical professionals in the National Health Service (NHS), the reasons included the length, complexity, accessibility and volume of regulations (Carthey, Walker, Deelchand, Vincent, & Griffiths, 2011). The length criteria measured was the steps in the procedure while complexity is a measure of how easy the regulation was to implement or find. Accessibility refers to how quickly the procedure could be accessed. Volume is distinct from length because it refers to the number of pages not the number of steps.

There were additional issues of redundancy and version control. Carthey et al. noted that a single task may have multiple procedures associated with it. In emergency situations, it was not always clear which procedure would take priority. Further, two departments may not be referencing the same version of a procedure. Leads to scenarios where two groups are referencing different versions of the same procedure.

This builds on previous research by Reason et al. who posited that five main factors shaped the occurrence of ten types of rule related behaviors including: availability of a procedure, the appropriateness of the procedure, whether or not it was followed, the correctness of the chosen course of action and the outcome of the action in terms of the achievement of personal goals (Reason, 1990).

The authors note that the variety of rules available will always be less diverse than the number of unsafe situations and that wholly safe situations cannot be made safe through perspective measures. Four types of organizational control can be employed: administrative controls, technical controls, social controls and self-controls.

Procedures can be inappropriate for the scenario, as in they were written for ideal conditions but emergencies, which can lead to noncompliance.

Skill Based Errors

Skill based errors occur when the individual has the right knowledge, skills and experience to execute a task but their attention is diverted from the task. This diversion can be due to external or internal factors. These errors occur once the task has become routine to the operator, which allows them to exert less cognitive effort on the task but increases the likelihood of distraction.

There are two categories of skill based errors: memory lapses and slips of action. Memory lapses occur during the task planning stage and typically involve a failure to include a step or losing one's place in the task sequence. Slips are unintended actions and occur during task execution and can take the following forms including: the right action on the wrong object, reordering steps in a procedure and performing the wrong action on the correct object.

Exceptional Violations

Exceptional violations are one of three identified types of violations within the human factors literature. The first category of errors, called situational errors, arise from pressures to keep the task moving towards completion. These errors occur when individuals believe that following the normal rule is no longer safe or will not produce the desired outcome, thus the decision is made to violate the exiting rule. Such violations are usually singular events (Lawton, 1998) .

The second category, routine violations, are the most commonplace and arise as shortcuts to the task at hand. Per researching conducted by Rebecca Lawton on Railway workers, routine violations can occur when the original rules and norms are seen as too onerous to follow (Lawton, 1998). The third and final category of

violations are Exceptional Violations which occur when an unusual situation requires an exceptional response.

Time pressure and performance

Hwang proposed an additive model of decision making under time pressure. Time pressure would positively impact task difficulty. Task difficulty negatively impacts decision strategy and positive impacts goal commitment (Hwang, 1994). According to a meta-analysis by Wofford et al., facing the challenge of a difficult task raises determination to reach a goal but task complexity can offset this determination (Wofford, Goodwin, & Premack, 1992) .

Hwang decomposed decision strategy into three categories based on research by Beach and Mitchell. These categories are: aided analytic,unaided, and unaided analytic. (Beach & Mitchell, 1978). Aided analytic methods require the decision maker to apply a prescribe procedure and solve the problem with the help of decision aids. Unaided analytic methods are applied when the decision maker attempt to explore the dimensions of the problem without aids. The last category includes non-analytic strategies such as flipping a coin or applying a heuristic.

Decision making behavior during gambling tasks is affected by time pressure. Under high time pressure conditions, individuals picked the less risky choice. Three mechanisms were proposed as a way to account for time pressure (Zur & Breznitz, 1981). Acceleration occurs when individuals process all information at a faster rate. Avoidance occurs when the individuals avoids making a choice. Finally, filtration is a compromise strategy where individuals select the subjectively important information for processing (MILLER, 1960).

Bear and Oldham studied the effect of time pressure has on individual creativity. They found that manufacturing employees creativity could be modeled as an inverted u-curve when time pressure was applied. However, this u-curve only occurred in

scenarios where the individuals were open to creativity and they were being supported. Otherwise, creative time pressure could be modeled as a line (Baer & Oldham, 2006).

Ahituv et al studied Israeli military commanders 5 key findings worth discussing. First, complete information improves task performance but that improvement is not always significant. Second, time pressure usually-but is not guaranteed- to impair performance. Experienced individuals are better at digesting more information when decision time is constrained. Less experienced individuals make more decisions within a given interval. Everyone made more decisions early in the scenario but slowed down once they realized the decisions would take time to be effective (Ahituv, Igbaria, & Sella, 1998).

Parkinson‘s law is an adage which states that work expands to fill time allowed for completion. There is some empirical evidence which suggests this phenomenon does exist. Bryan and Locke altered amount of time students had to complete a series of mathematics problems. Their findings indicated students who had been given double the amount of time worked significantly longer than student who were given a minimal amount of time to complete the tasks (Bryan & Locke, 1967). Work by Locke with wood harvesters found that wood harvesters who faced time restrictions on their production quota would exert more effort towards task completion than their less stressed counterparts. In the lumber industry, mills would restrict the number of days per week they would buy from a lumber harvester because the mill had more wood in stock than it could process. Mills would employ different strategies to restrict purchases, including only buying wood from harvesters on a first come first serve basis. Harvesters would try to minimize their lost income by increasing their output rate to accommodate this schedule (G. P. Latham & Locke, 1975).

The researchers found that when mills restricted the amount of wood they would purchase, they implicitly encouraged harvesting crews to adopt a higher production goal. This is because the crews were trying to fit five days of work into two or three days. The author‘s argued these findings validated Parkinsons law by stating that

given proper incentivizing, workers would adjust the rate they worked at to accord with the time restraints imposed by the task (G. P. Latham & Locke, 1975).

These results have been replicated in different organizations. In a five-year study that tested the association between time pressure and productivity found evidence of the impact of time pressure on individual productivity. In this study, performance was assessed along the lines of innovation, usefulness and productivity. A key finding of this study was that performance varied as a function of the absolute amount of time reported (Andrews & Farris, 1972) .

Peters et al. continued these organizational studies and found a positive correlation between performance and perceived time pressure of .19 (p¡0.05). They noted that while this relationship is not particularly strong, it is above chance and replicates the work of previous researchers. They suggested that time pressure is more associated with the rate of work rather than the amount of work done. Further rate of work is one of several factors which influences performance (Peters, O'Connor, Pooyan, & Quick, 1984).

Eckhardt et al. examined the question of how social network users experienced feelings of stress and fatigue. The proposed that users who were under higher pressure to perform had an objectively higher degree of strain and better performance than users without the performance pressure. Further, they proposed that higher levels of techno-stress would lead to lower objective performance and higher subjective and objective levels of stress (Eckhardt, Maier, & Buettner, 2012). Techno-stress is a combination of techno-uncertainty, techno-complexity, techno-invasion and techno-overload (Ayyagari, Grover, & Purvis, 2011)). Using a combination of measures of eye-tracking and electro-dermal activity, Eckhardt et al. found that a pressure to perform and high levels of system level experience contributed to a higher level of performance. This time pressure did enhance individual measures of strain (Eckhardt et al., 2012).

These studies suggest indicate that time pressure has an impact of human performance. Performance can be measured in a variety of ways- from individual creativity,

productivity, and output- depending on the context of the study. However, the presence of perceived and real time pressures has an impact on individual performance.

C.2 Human Machine Interaction Events

The following five sections briefly outline research on cognitive overload, distraction, fatigue, multi-tasking and pressure. These events are discussed because they can be induced while using the MATB-II.

C.2.1 Cognitive Overload

Cognitive overload which is related to cognitive load - occurs when an individual is presented with so much information and stimuli they cannot process everything. David Kirsh outlined four systems of causes of cognitive overload: an oversupply of information, too much information demand, inadequate tools to helps with metacognition and the need to deal with distractions (Kirsh, 2000).

Per Kirsh, these four causes of cognitive overload can be further sub-divided. For example, information oversupply can be divided into an oversupply of pushed information and an oversupply of retrievable information. These both manifest as a linear increase in quality information increases at a relatively constant rate but the volume of all information increases exponentially. As such, knowledge workers develop a series of strategies to cope with this clutter including just in time learning or just in case learning (Kirsh, 2000).

Cognitive overload has an impact on instructional domains. Mayer and Moreno identified five types of over-load scenarios in multi-media instruction. The first type occurs when essential processing in a sensory channel are greater than the cognitive capacity of that channel. The second type occurs when essential processing in the visual and auditory channels are greater than the cognitive capacity overall (Mayer & Moreno, 2003). The other three types are discussed in more detail in the literature review.

Strategies for overcoming cognitive overload including off-loading, segmenting and pre-training. Segmenting occurs when the trainer allows for time between information bites. Pre-training occurs by providing some degree of pre-familiarization with the names or characteristics of the components. Off-loading is moving information from one presentation domain to another (Mayer & Moreno, 2003).

Fatigue

Fatigue is a natural physiological and psychological reaction to prolonged periods of stress. This state corresponds with a decrease capacity to perform physical or mental work (Queensland Government, 2009). Unlike distraction or cognitive overload where individuals can assess how overloaded or distracted they are, individuals are less capable of assessing how fatigued they are (Goel, Basner, Rao, & Dinges, 2013).

There are several biological and social factors which influence fatigue. The first factor is how mentally and physically demanding the task is to perform. This factor can manifest as sub-optimal planning and decreased flexibility. Van der Linden et al. induced fatigue by having participants perform a cognitively complex task for two hours. They found that participants showed greater perseveration errors during the Wisconsin card sorting task and long planning time during a tower of London task than non-fatigued individuals (van der Linden, Frese, & Meijman, 2003).

The next set of factors influencing fatigue involve sleep. Being awake for more than sixteen hours, having inadequate amounts of sleep and being active during an adverse time of the circadian cycle have a performance detriment. One domain where sleep deprivation is noticed is the medical professional. Sleep deprivation induced fatigue contributes to an increase in the number and severity of surgical errors (Eastridge et al., 2003) and the number of misjudgments made by residents (Baldwin & Daugherty, 2004).

Distraction

A distraction is anything which draws an individual's attention away from the task which they should be employed in. Distractions are a natural part of work environment - i.e. the urgent email from your boss - but they are avoidable. Distractions can occur in the same modality a loud noise occurring while listening to someone talk-or in different modalities such as car passing by during the conversation (Lee, Young, & Regan, 2008).

One of the most immediately available examples of the negative impacts of distraction on performance at least in the public consciousness- is driving performance. Horberry and colleagues found that performing an additional in vehicle task such as using your cellphone or tuning the radio while driving can lead to degradation of driving performance under certain road conditions. The authors noted that drivers of different ages did attempt to compensate for potential in-vehicle distractions by altering their driving behavior (Horberry, Anderson, Regan, Triggs, & Brown, 2006). Further research in the domain of distracted driving demonstrated that cellphones were more distracting than passenger conversations. This is because conversations between the passenger and driver can focus on traffic conditions while cellphone conversations cannot. Thus, interpersonal conversations can increase situational awareness while cellphone conversations cannot (Drews, Pasupathi, & Strayer, 2008).

Distractions also negatively impact howl individuals perform recall tasks. Banbury et al. found that recall memory is susceptible interference by irrelevant sounds. Specifically, they found that seriation the cognition function used for maintaining order in short term memory- was very susceptible to this interference. Further, this susceptibility is not limited to tones but also to speech like sounds (Banbury, Macken, Tremblay, & Jones, 2001). Additional research on irrelevant background speech found that speech in native or non-native languages to the listener are equally distracting (Marsh & Jones, 2010).

Loss of Vigilance

Vigilance- or sustained concentration- is the ability to maintain concentrated attention over a prolonged period of time. During this period of sustained concentration an individual is attmeping to detect the appearance of a target stimulus. This stimulus can appear at an unknown time (Warm, Parasuraman, & Matthews, 2008). A practical example of a task requiring vigilance is lifegaurding (Warm, Matthews, & Finomore Jr, 2017) . In contrast, loss of vigilance occurs when the operator loses concentration on the task.

Fatigue, distraction and loss of vigilance are similar enough at first glance that one could confuse one for another. However, fatigue and loss of concentration occur over a period of sustained activity while distraction occurs over a faster time period. Further, both distraction and fatigue can contribute to a loss of vigilance.

Multi-tasking

Multi-tasking refers to the apparent human ability to perform two or more tasks or activities over a short period. Researchers have known for about 50 years that performance during multi-tasking activities often result in decreased performance in both tasks. This decrease in performance has been attributed to a phenomenon called cognitive bottle-necking.

Pashler argued that the psychological refractory period effects acts as a bottleneck which interferes with an individuals ability to process and respond to incoming stimuli (Pashler, 1994). This cognitive bottlenecking theory suggests that individuals have finite attentional resources that can only be allocated to one task at a time, therefore stimuli are filtered so only the most relevant are perceived (*Bottleneck Theory*, 0; Borst, Taatgen, & van Rijn, 2010).

Patten and colleagues studied measures of driver workload in two studies. The first study focused on driver distraction when using a cellphone as a function of task complexity. Participants drove on a series of different simulated roads (motorways,

cities and rural areas) while having a conversation on a hands-free or hand held device. Participants all drove the same route on a defined section of motorway. The researchers measured driver workload using the NASA TLX. Patten et al found that the content or complexity of the conversation had a greater influence over driver distraction than the complexity of the road (Patten et al., 2004).

C.3 Experience and Expertise

C.3.1 Experience

Experience refers to knowledge or skill in a particular activity that has been gained because and individual has performed that activity for a long while. The operational definition of experience will depend on the task. For example, Blaauw defined experienced drivers as individuals who had driven for more than 3 years and had logged over 30,000 kilometers while inexperienced drivers were those who had either just passed their driving test or were taking classes. Blaauw found that experienced drivers found lateral and longitudinal driving control tasks to be less difficult than inexperienced drivers (Blaauw, 1982). There is evidence that task experience changes how individuals will approach a task. Groves et. al showed that while expert clinicians committed more errors during data collection and interpretation but produced more accurate diagnoses than novice clinicians (Groves, O'Rourke, & Alexander, 2003) . Experts do not produce and test more hypotheses than novices do . Rather, they produce and test better hypotheses (Neufeld, Norman, Feightner, & Barrows, 1981).

There are specific conditions where experience can be related to age. In the context of driving, younger drivers are said to be less experienced than older drivers (Peck, 1993). Age is not necessarily a poor proxy for experience, but it is insensitive to contextual factors such as gender (C. F. Miller, 1993).

Littlepage et al. provide an important insight on task experience. The authors defined task experience as previous experience with similar tasks (Littlepage, Robison, & Reddington, 1997). Prior research by Goodman and Shah state the task

experience is an important contributor job knowledge (Goodman & Shah, 1992). Additionally, there is a large body of research on training effectiveness which emphasizes that individual increases with prior experience in similar tasks (Burke & Day, 1986; Guzzo, Jette, & Katzell, 1985; Hellervik, Hazucha, & Schneider, 1992) . Experience performing a particular task can be directly measured. Feigenspanet al noted that in the context of programming and software engineering, experience could be evaluated by examining programming language constructs (see (Bacher, Mac Namee, & Kelleher, 2018),(Bergersen, Sjberg, & Dyb\a a, 2014), (Feigenspan & Siegmund, 2012) in (Feigenspan, Kastner, Liebig, Apel, & Hanenberg, 2012)) . However, the authors note that experience can also be examined through the leans of education and self-estimation (Bunse, 2006; Ricca, Di Penta, Torchiano, Tonella, & Ceccato, 2007).

C.3.2 Expertise

One of the challenges in human factors and cognitive psychology is the question of how to define expertise. There are some domains where a gold standard in the form of exhaustive documentation of domain knowledge or professional licensing- have been codified and is accessible for use by others in the domain. However, the challenge remains in how to generate an operationalized definition of expertise. Hoffman proposed that expertise can be defined in terms of cognitive development, the experts knowledge structure and the experts reasoning process (Hoffman, 1998).

In this context, cognitive development is not correlated with measures of intelligence. Rather, it refers to a developmental progression from a literal interpretation of and superficial understanding of the problems to a more principled and nuanced understanding of the same problem. Skill is acquired based on accumulated practice and not age. This leads to the question of how to distinguish between someone who is at a novice level and the expert level.

Research suggests that the developmental process from novice to expert have level-like shifts. First, with practice skills require less effort and become more automatic in execution. Second, it is rare for individuals to skip levels. Expert teachers can anticipate the mistakes novice learners are about to make. Finally, it is rare for individuals to regress unless they fall out of practice .

How do you empirically measure an individual‘s level of expertise at a task? One proposed method of evaluating expertise is the Cochran-Weiss-Shanteau (CWS) index. This index a ratio of the ability to consistently discriminate various stimuli in a domain and consistent treatment of those stimuli. The CWS formula is defined as discrimination over inconsistency. According to Weiss et al, discrimination is an individuals differential evaluation of stimuli while inconsistency refers to variations in the individuals evaluations of the stimuli overtime (Weiss & Shanteau, 2003).

The CWS is important in the study of expertise for two reasons. First, it builds on the existing body of research which states that differences exist between novice and expert billiard players, American football fans and chess players as well as technical professions such as doctors and air traffic controllers (Abernethy, Neal, & Koning, 1994; Holding, 1979; Werner & Thies, 2000). Further, the measure is related to the F- ratio. This gives the CWS a degree of statistical meaning.

This metric has been validated in the context of aviation decision making. Pauley et al. examined the performance of qualified pilots, student pilots ,and geography students at weather related risk assessment tasks. The authors found that the three groups did not have significantly different CWS scores. Further, they found that when tasks decreased reliance on memory, there was a relationship between flight experience and CWS scores (Pauley, O'Hare, & Wiggins, 2009).

Applying the CWS as a measure of expertise also validates assumptions about the intermediate effect. The intermediate effect refers to a U-shaped curve which appears when measuring the performance of individuals who work in domains where extensive diagnostic knowledge is required. Witterman et al. applied the CWS to first year counseling students, master level counseling students and professional coun-

selors and found that the mean CWS scores for the first year and professional counselors were significantly different than the master levels students (C. L. M. Witteman, Weiss, & Metzmacher, 2012). Witterman et al. noted that their findings mirrored previous work which demonstrated the intermediate effect (C. L. Witteman & Van Den Bercken, 2007).

C.4 Gamification

C.4.1 Does Gamification work?

The research on the effectiveness of gamification is mixed. Similar experiences of gamification do not elicit the same results in different contexts. According to Hamari et al outcomes are either positive or negative or positive or neutral (Hamari & Tuunanen, 2014). The positive and negative effects of gamification varies among individuals (Passos, Medeiros, Neto, & Clua, 2011).

Other demographic factors can influence the outcome of a gamified experience. The individual'age along with their familiarity with games impacts interest in and use of a gamified experience (Bagley, 2012). Research also suggested that women were more engaged in gamified experiences than men (McDaniel, Lindgren, & Friskics, 2012). While this result surprised the authors, it should be noted that women are commonly assumed to not enjoy games.

C.4.2 Limits of Gamification

Gamification is not without limitations. The practice can lead to the introduction of unintended behaviors among users. One example given of this comes from BMW, who used an app to challenge drivers to be more fuel efficient. Unintentionally, drivers started engaging in unsafe driving practices in order to score more points.

According to Sebastian Detering, the other unitneded side effects of gamification include: hitting the target but missing the point, gaming the system and messing with

implicit social norms (Deterding, 2010b). Hitting the target but missing the point occurs when a game encourages behaviors that statisfy the challenge at hand, but fail to impart the intended point. A commonly cited example came from a review of British hospitals, where incoming patients were placed on trolleys- labeled as hospital beds- to meet the goal of admitting a patient to a bed in a 12 hour window (Bevan & Hood, 2006). Messing with implicit social norms occurs when players are encouraged to violate a real or implied social norm to win. Finallny, Gaming the system occurs when players try to hack or cheat the system in order to achieve the desired outcome (Deterding, 2010b).

C.4.3 Competition

Games present a unique opportunity for researchers to study competitive behavior. This is because gaming as a domain offers several design tools, mechanics and interaction modes designed foster competition. Design elements such as leader-boards allow researchers to directly interfere with a player's competitive drive. The mechanics of a game can be alter to handicap a player. This handicap in turn alters their behavior. Finally, games are explicitly designed to be competitive or cooperative. Competitive and cooperative games require a players to goal structure.

Peng and Hsieh studied goal structure in multiplayer gaming had on player motivation, goal commitment and performance. They studied this effect by having participants play a specially designed game. The game had three versions: solo play and two multiplayer modes. In the solo mode, participants would see their performance in real time. In the cooperative and competitive multiplayer, players would see their score and the other players score on screen (Peng & Hsieh, 2012).

The concepts of player handicapping and balancing are two mechanical ways of altering competition. Handicapping refers to a set mechanic which alter what the player can do. Notable handicapping methods include point handicapping where the winner of the last round starts the next stage with a score penalty or mechanical

handicapping. Mechanical handicapping alters what actions the player can take in game. Balancing occurs when the game accounts for player skill.

Balancing sounds like a special case of handicapping but it is not. Handicapping is a means of assigning a score compensation as a means of equalizing players chances of winning. Balancing refers to actions which help provide the player with the right mix of challenge and difficulty (Falstein, 2005). Balance is important because of player engagement. If a game is too difficult or too easy, players may become frustrated (Newheiser, 2009).

Cechanowicz et al identified four ways of balancing players: matchmaking, asymmetric roles, difficulty adjustment and assists. An assist is anything which adjusts a players performance in basic task by simplifying the input needed to perform the action. One example is auto-aiming in first person shooters. Difficulty adjustment refers to methods which alter the game difficulty to match the players skill (Cechanowicz, Gutwin, Bateman, Mandryk, & Stavness, 2014).

This process can be initiated by a player or it can be a dynamic process. However, this technique can lead to players discounting their success because an achievement feels meaningless (Bostan & Ogut, 2009). Asymmetric roles allow players to execute tasks or roles which better suited to their play level and ability. This technique is often used in multiplayer games to ensure that all team roles will be fulfilled. Finally, matchmaking is a computational process by which players of equivalent skill are grouped together (Cechanowicz et al., 2014).

Each of these techniques have their own benefits and drawbacks. Mishandled difficulty adjustment systems can lead experienced players feeling frustrated at the games mechanics which seem to favor weaker players (Newheiser, 2009). Asymmetric roles need to be balanced so each role feels like it fulfilling a unique purpose.

These methods of balancing play each have their own design implications and limitations for novice and expert players. Cechanowicz et al. sought to answer the following questions in the context of a racing game. First, does balancing work and

does balancing improve naive players experiences? Further, does balancing detract from the experiences of expert players (Cechanowicz et al., 2014)?

To test questions of balancing in action, the authors developed a series of assists and adaptive algorithms were developed. The assists were intended to improve the experience and performance of novices and act as a hindrance to the experts. The cars speed was modified by adding or subtracting a scalar value. Acceleration was adjusted by multiplicative increasing or decrease the cars air friction constant. Finally, steering was modified to induce error (Cechanowicz et al., 2014)?

Four adaptive algorithms were developed: realtime100, realtime40, rolling and MaxDistance. Realtime100 and realtime40 provided assistance or hindrance to players based on their relative distance from each other. The rolling algorithm averaged the assistance from the two realtime algorithms. Finally, maxdistance scaled the provided assistance based on the maximum distance between vehicles (Cechanowicz et al., 2014)?

The authors found that balancing does have a positive impact on the performance elements they tested. These metrics included the win/loss rate, lead changes and distance between cars. Balancing did improve the novice player's experience when either the rolling, MaxDistance or realtime40 algorithms were used. Balancing did impact the expert's perception of fairness when the rolling average was used.

This research leads to questions of how to balance performance based on individual ability. In a study focusing on balance between mobility and non-mobility impaired players, Gerling et. al found that explicit input balancing results in lower self-esteem in both players. They attributed this to the fact that input balancing is a highly visible means of manipulating performance. Time balancing balanced the score differential but did not result in the weaker player winning more. Finally, score balancing did result in the weaker player winning more games, but this resulted in the game being overbalanced (Gerling, Miller, Mandryk, Birk, & Smeddinck, 2014).

The results of this study suggested that balancing strategies used in collocated games did not affect the weaker players in game performance. However, players did

perceive the game balancing under certain circumstances. The authors extended the conversation outward to discussing overbalancing. They suggested that the intensity of balancing needs to be carefully adapted so a balance is struck between the weaker playing being able to wind and for the stronger player to feel competent (Gerling et al., 2014).

C.4.4 Manipulating Player Experience

A key question in player experience (pX) research is how to give the player a sense of success and progress without altering the experience. Player experience is measured using the Player Experience of Need Satisfaction (PENS) survey. PENS is a 21 item assessment instrument which uses a Likert scale to measure immersion, relatedness, competence, autonomy and intuitive controls (R. M. Ryan, Rigby, & Przybylski, 2006).

Johnson and colleagues used the PENS to assess if different genres of games had different levels of player experience. They found that action adventure, role playing and action role playing games fostered a sense of immersion and presence in the player. Further, games seem equally capable of nurturing experiences of flow and competence (Johnson, Nacke, & Wyeth, 2015).

While Johnsons work does provide insight in pX research, it does not address direct manipulation of player experience. One method of altering a players experience is to change their position on a leaderboard. Bowey et al. studied if manipulating a players position on a leaderboard would affect their performance and enjoyment. They found that leaderboard position does influence a players perception of competence, autonomy and presence (Bowey, Birk, & Mandryk, 2015).

Their findings also indicate significant interactions between leaderboard position and autonomy. Specifically, they found that when the players position is shown, that there is an interaction between their perceived success and autonomy. This type of

interaction is not present when the score is hidden. These findings were not affected by demographic factors such as age, gender or and hours played (Bowey et al., 2015).

These findings lead to the question of whether goal-setting is influenced by leader-board position. Chernbumroong and colleagues studied this question using three leader-boards: do-your-best, difficult and impossible. Do your best represented the scores of all participants while the difficult leader-board showed the players position with three participants above and below them. The impossible leader board only showed the top five scores (Chernbumroong, Sureephong, & Muangmoon, 2017).

The authors found that players using the difficult leaderboard had the highest scores on an online CPR training test. However, this group had the lowest task completion rate relative to the other groups. The authors also noted that users in the impossible and difficult scenarios showed the greatest improvement (Chernbumroong et al., 2017).

These findings were attributed to differences in the player's goals. The do-your-best leaderboard did not have a specific and clear goal. This reduced the players concern about their score or the competitive nature of the environment. In the difficult scenarios. Players could align their goals with the actions of others. This lead to players competing to see if they could best each other. Finally, in the impossible conditions, the players felt it was too difficult to reach the top five, so they instead focused their effort on task completion (Chernbumroong et al., 2017).

Individual motivations

Discussions of an individual motivations to play a game are sub dividable into the following topics. From a game content perspective, different types of players each want a unique experience from the game. While each of these player types have a unique set of goals, the steps they take to reach them can be described using self-determination theory.

Self Determination Theory investigates factors which facilitate or weaken intrinsic and extrinsic motivation. Intrinsic motivator underpinning games and sports (Frederick & Ryan, 1993, 1995). A sub-set of SDT is cognitive evaluation theory which is explicitly concerned with contextual factors enhancing or weakening intrinsic motivation (Deci & Ryan, 1980; K. A. Miller, Deci, & Ryan, 1988). CET states that intrinsic motivation is enhanced by conditions and events which enhance a player's sense of competence and autonomy. Conditions and events which decrease perceived autonomy and competence undermine a player's extrinsic motivation.

The player experience of needs satisfaction (PENS) builds on SDT and CET. PENS consists of five metrics: autonomy, competence, immersion, intuitive controls and relatedness. Autonomy relates a players sense of willingness or volition to complete a task (Deci & Ryan, 1980; K. A. Miller et al., 1988). Competence is a feeling of effectiveness. Immersion refers to a sense of 'being in'the world the game is presenting as opposed to feeling like you are outside the game. Intuitive controls refer to the degree which controls can be easily mastered and make sense in context. Finally, relatedness is the degree with which the player connects to others (La Guardia, Ryan, Couchman, & Deci, 2000).

While PENS is useful for describing motivation in microscopic terms, game designers and researchers categorize players by what they want from a game. One of the most prominent examples comes from Richard Bartles study of multi-user dungeons (MUDs). He observed that players could be divided into four categories: achievers, explorers, socializers and killers. Achievers set game-related goals and adopt their play style to accomplish them. Explores want to learn everything they can about the world while socializers want to interact with other players or characters in the game. Killers seek to cause distress in other players (Bartle, 1996).

Figure C.2 shows Bartles Taxonomy along two axes. These axes are representative of the player types motivations. Explorers like interacting with the world and uncovering as much about the world they can. These players tend not to be score focused. Achievers want to do things and feel the point of playing is the master as much as

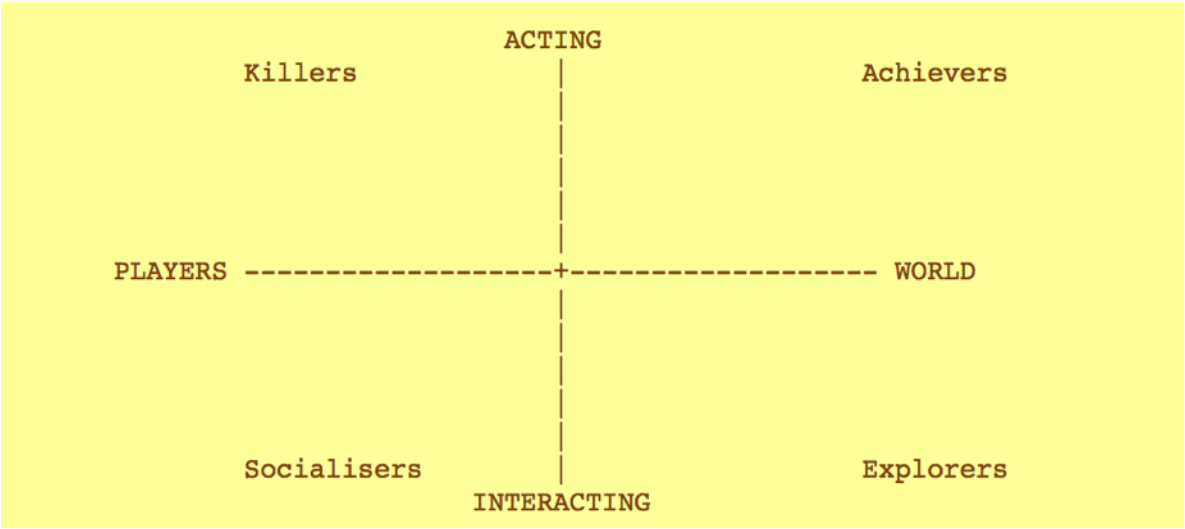


Fig. C.2. Player Types

they can. Killers want to act on people, even when these actions occur without the consent of other players. Finally, socializers want to learn about the other players through their interactions (Bartle, 1996).

In a meta-analysis of player types, Hamari and Tuunanen concluded there are seven dimensions along which to describe players motivation and behavior. These dimensions are achievement, exploration, sociability, domination, immersion, intensity and in-game demographics (Hamari et al., 2014). They noted that most player categorization systems are directly inspired by Bartles taxonomy. Some researchers, such as (Kallio, Myr, & Kaipainen, 2011), see the reliance on taxonomies as reductionist arguing that players do not strictly belong to one type.

C.4.5 Games and Executive function

According to (Schmidt, 2002), the term awareness is highly imprecise. The term is often used in association with a qualifing adjective such as 'general awareness '(Gutwin & Greenberg, 1999), 'workspace awareness '(Gutwin & Greenberg, 1996) or 'passive awareness '(Dourish & Bellotti, 1992). This comment does not account for

the topic of situational awareness, visual-spatial awareness or auditory awareness. In certain gaming contexts, awareness is best conceptualized as a way to understand the activities of others while providing context for your own actions (Dourish & Bellotti, 1992). Dourish and Bellotti's definition of awareness is the best broad definition of awareness in a gaming context.

Little research has been conducted on mode and situational awareness explicitly on games. This lack of research does not mean the two concepts are not related. The bridge between traditional situational awareness research and games research comes from research in unmanned aerial vehicles.

The reason UAVs represent a bridge case stems from control modalities. UAVs are primarily controlled through a visual medium (J. L. Drury & Scott, 2008). While there is research that focuses on providing haptic and auditory feedback to UAV operators, most of their decisions will be made based on what they can or cannot see. Videogames are controlled in a visual medium through an interaction device.

Further, many of the situations studied in the UAV literature are easily replicated in a game context. For example, Drury and Scott studied how operators would use a single UAV to coordinate unmanned ground vehicles (Riley & Endsley, 2005). This research was buttressed by previous work by Ruff et al. which showed that operator performance decreased as the number of UAVs being supervised increased (Ruff, Calhoun, Draper, Fontejon, & Guilfoos, 2004). With relatively minor tweaks to the design of the system, these same scenarios can be recreated in a game.

Expertise and executive functions

The effect of videogames on task switching and other executive control functions is often done by dividing players into expert or non-expert categories. Expert gamers are players who log more hours than non-gamers and have experience in a wider variety of games. Non-gamers are players who play less than one hour per week (Boot, Kramer, Simons, Fabiani, & Gratton, 2008).

Unsurprisingly, expert gamers outperform non-players in a variety of tasks. Experts are better able to track objects moving at higher speeds, have a more accurate short term memory in visual memory tests and are faster at switching between tasks. Further, as Boot et al. note, with few exceptions practicing videos games for more than 20 hours does not lead non-players to perform at parity with the expert players (Boot et al., 2008). This finding builds on previous research which showed it took at minimum 10 hours of practice were needed to obtain skill transfer (Green & Bavelier, 2003). Strobach et al. provided evidence showing that increased practice does improve reaction time and that these effects transfer (Strobach, Frensch, & Schubert, 2012).

Evidence exists which suggests that playing action videogames has a direct impact on the players visual attention (Green & Bavelier, 2003). Boot et al. found that a significant difference exists between expert and non-expert player's cognitive abilities. Experts were better able to detect changes in short term memory and track visual objects moving speed. What Boot was unable to resolve was if these performance differences are attributable to game experience or an untested difference in the groups.

These findings can be debated since expert players are assumed to have a deeper knowledge base to draw from. Experts are assumed to have a better and concrete grasp of the controls. This debate was addressed by Green et al. who found that expert action game players task performance is not a product of their ability to react to altered key mappings. However, expertise does not always lead to objectively better performance (Shawn Green, Sugarman, Medford, Klobusicky, & Bavelier, 2012).

There are limits to expertise. One of the few skills where experts outperform non-experts is reaction time. Having a fast reaction time is of marginal significance when cost switching. Further, performance improvement is not universal by game. Learning to play an action game will lead to faster reaction times in strategy games but the reserve is not always true (Shawn Green et al., 2012)

This research leads to the question of what distinguishes an expert from a non-expert gamer. One critique of the distinction hinges on the diversity of genres within

gaming. Action games can be subdivided into distinct genres (A. J. Latham, Patston, & Tippett, 2013). An expert in one genre of action games may not be an expert in other games.

C.5 Interaction Modalities

The following section focuses on the relvant literature for two input modalities: a) mouse and keyboard and b) trackpads. User interaction in both domains is informed by Fitts's Law. This law predicts that the time it will take for a user to move a pointer to a target is a function of the ratio of the ratio between the width of the target and the distance to the target (Fitts, 1954). Studies have shown that user performance in mouse and trackpad conditions are comparable (Rozado, 2013).

The mouse and trackpad are not the only interaction modalities that can be used in pointing tasks. Other modalities include: directional keypresses, joysticks, eye gaze and motion tracking. It has long been known that mouse movements are faster than keypresses (Card, English, & Burr, 1978). Further, there is evidence to suggest that eye gaze and motion tracking are not sufficently more complicated to implement and result in slower reaction times (General, Da Silva, Esteves, Halleran, & Liut, n.d.; Rozado, 2013). This difference in performance could be attributed to the relative rareity of devices that employ motion or eye tracking.

C.6 Data Science

C.6.1 K Nearest Neighbors

The goal of the k-Nearest Neighbors (k-NN) algorithm is to classify an unknown example into the most likely class among K nearest examples. When K-NN is used for classification, an object is classified by majority vote based on proximity to its neighbors. The object would be assigned the most common value of the neighbors.

There are several computational methods for determining the nearest neighbor. When K=1, the visually intuitive answer is to pick the only other point. At higher dimensionalities of K this intuition can fail, and the analyst must consider the distance between points. For continuous data, the analyst must calculate the Euclidean, Manhattan or Minkowski Distance. When the data is categorical, the analyst can use the Hamming Distance (Hamming, 1950).

K-NN relies on instance based learning and lazy learning. Instance based learning methods construct hypothesis directly from the training instances provided (Russell & Norvig, 2013). A consequence of this is that hypothesis can be more complex as they are given more data. The obvious downside is that more memory is required to store the training data. This can lead to slow run times.

Lazy Learning algorithms are defined by three behaviors: deferring, reply and flushing. Deferring means algorithms store all the training data and delay processing until a query is made. Replying occurs when queries are answered by combing the data using a localized learning approach. Once the query has been answered, the results are flushed (Wettschereck, Aha, & Mohri, 1997) .

One of the ways to improve the value of a large volume of data is to use pattern matching algorithms. These classes of algorithms, described in more detail below, search for and analyze underlying trends in the data. A designer or researcher can program in specific patterns to search and filter the data, x=1, or search for more complex pattern, $f(x) = n * f(n2)$.

The following paragraphs focus on three classes of data mining algorithms relevant to the project. The three classes of algorithms discussed in order are: k-means clustering, real time cluster analysis. The purpose for this ordering will be explained at the end of this section.

K-Means clustering algorithms function by partitioning n observations in k distinct clusters. K refers to the number of initial assigned means. Each observation is grouped with the closest mean. The result of a K-Means analysis is k distinct groupings. There is evidence that K-means can be applied to continuous data (Faber, 1994). The case

for using K-Means clustering in this experiment is twofold. First, one cluster of responses could be defined as 'error state'while another cluster could be defined as 'pre-error'.

The second method considered for data analysis is real-time cluster analysis. There are two method of interest: hierarchical clustering and correlation clustering. Hierarchical clustering word in either a top down or bottom up manner and seek to build the largest cluster (Blashfield, 1976). Correlation clustering works by grouping data points by how co-related they are with each-other (Bansal, Blum, & Chawla, 2004).

The third and final method under consideration are Neural Nets. In broad terms, neural nets are computing systems which progressively learn by considering examples provided to them without task specific programming. These networks can be trained using a small dataset and letting the program learn on over time (Bigus, 1996; Lu, Setiono, & Liu, 1996).

Sliding Window Clustering One of the key problems in this project is the question of which time-period needs to be evaluated to determine if an error occurred. In general terms, the data mining algorithm needs to preference the most recent elements. Streaming clustering algorithms provide a technique to solve this problem. Clustering algorithms as previously discussed-work by grouping similar data points together. The term siding window refers to an algorithm which performs an operation on a subset of data then moves on to perform the same calculation on a different 'frame'.

The simplest in terms of implementation- clustering algorithm is the K-Means clustering algorithms. This algorithm functions by partitioning n observations in k distinct clusters. Each cluster starts with a single point termed a centroid. Clusters are populated by the other data points which are assigned based on their distance from a centroid. There is evidence that K-means can be applied to continuous data (Faber, 1994).

One of the limitations when using clustering algorithms is that large volumes of streaming data can render algorithms inefficient. One-pass algorithms have been proposed as a solution to this problem, but per Aggarwal and colleagues, these algorithms

do not account for the evolution in the data. Further, they argued that the quality of the clusters can degrade when the data evolves over time and streaming algorithms need greater functionality to discover and explore clusters at different portions of the stream (Aggarwal et al., 2003).

Streaming clustering algorithms work by decomposing the stream into smaller partitions, performing an operation on that partition and then moving forward in time. The small partitions are termed micro-clusters which are stored as a snapshot. (Aggarwal et al., 2003; Dang, Lee, Ng, & Ong, 2009). Aggarwal et al. used micro-clusters as input for an off-line clustering algorithm, which used these clusters to create macro-clusters. He also introduced the concept of a pyramidal time frame, which stored snapshots at different levels of granularity. The pyramidal time frame allows for the identify and removal of repeated values while still capturing the evolving nature of the dataset (Aggarwal et al., 2003).

Dang et al. proposed a variation to Aggarwals method. Their algorithm Sliding Window with Expectation Maximization or SWEM- divides a micro-component based on the highest variance sum and size of the cluster. The algorithm then calculates the mid-point between the averages of each sub-component (Dang et al., 2009). SWEM can merge components if they are small enough and close enough. If the means of each sub-cluster are within standard deviations of each other, SWEM will form one new cluster. The algorithm will automatically remove the oldest entry

C.6.2 Training Test Split

The heuristic for splitting the data into a training and test set ranges suggests an 80/20 split is generally the most reliably accurate across conditions (Foody, McCulloch, & Yates, 1995; Liu & Cocea, 2017),Shahin2004Data. The more training data the model has to work with the more accurate it will be overall. In theory, a model given infinite training data would be able to classify with 100% accuracy given enough time. However, this heuristic is subject to limitations and most of the academic lit-

erature suggests using partition methods to sample the data. There is evidence to suggest that the size of the training set -in terms of percentage of the total data- does not have an effect on prediction accuracy (Abdelwahab, Bahgat, Lowrance, & Elmaghraby, 2015; Foody et al., 1995).

Alternate train-test splitting methods

There are alternate methods that can be employed to generate training and test data sets. One of the most common forms of cross validation is K-folds cross validation. This method divides the data into k even sized subsets. A single subset is retained for validating the model while k-1 subsets are used to build the model (McLachlan, Do, & Ambroise, 2005) . This method is repeated k times.

C.6.3 CARET Package

The CARET (classification and regression training) package focuses on simplifying model training and tuning techniques (Kuhn, 2008) . At the time of writing, this package has been referenced in 1,400 scientific articles including (Van Essen, 2012; Bischl et al., 2016) .