3-2021

Comparison of Machine Learning Techniques on Trust Detection Using EEG

James R. Elkins

COMPARISON OF MACHINE LEARNING TECHNIQUES ON TRUST DETECTION USING EEG

THESIS

James R. Elkins, 2d Lt, USAF

AFIT-ENG-MS-21-M-033

DEPARTMENT OF THE AIR FORCE
AIR UNIVERSITY

AIR FORCE INSTITUTE OF TECHNOLOGY

Wright-Patterson Air Force Base, Ohio

AFIT-ENG-MS-21-M-033

COMPARISON OF MACHINE LEARNING TECHNIQUES ON TRUST DETECTION USING EEG

THESIS

Presented to the Faculty

Department of Electrical and Computer Engineering

Graduate School of Engineering and Management

Air Force Institute of Technology

Air University

Air Education and Training Command

In Partial Fulfillment of the Requirements for the

Degree of Master of Science in Computer Science

James R. Elkins,

Second Lieutenant, USAF

March 2021

AFIT-ENG-MS-21-M-033

COMPARISON OF MACHINE LEARNING TECHNIQUES ON TRUST DETECTION USING EEG

James R. Elkins,

Second Lieutenant, USAF

Committee Membership:

Dr. Brett Borghetti, Ph. D.
Chair

Dr. Michael Miller, Ph. D.
Member

Lt. Col James Noel, Ph. D.
Member

AFIT-ENG-MS-21-M-033

Abstract

Trust is a pillar of society and is a fundamental aspect in every relationship. With the use of automated agents in today's workforce exponentially growing, being able to actively monitor an individual's trust level who is working with the automation is becoming increasingly more important. Humans often have miscalibrated trust in automation and therefore are prone to making costly mistakes. Since deciding to trust or distrust has been shown to correlate with specific brain activity, it is thought that there are EEG signals which are associated with this decision. Using both a human-human trust and a human-machine trust EEG dataset from past research, within-participant, cross-participant, and cross-task cross-participant trust detection was attempted. Six machine learning models, logistic regression, LDA, QDA, SVM, RFC, and an ANN, were used for each experiment. Multiple within-participant models had balanced accuracies greater than 70.00%, but no cross-participant or cross-participant, cross-task models achieved this. The highest balanced accuracies achieved were 64.43% and 53.37% respectively.

Acknowledgments

First and foremost, I would like to thank my loving wife for her patience and support. I'd like to thank my dog Beau for being right by my side during this entire process. I owe a large bit of gratitude to my faculty advisor Dr. Brett Borghetti who guided me through this difficult journey. I'd also like to thank my other committee members, Dr. Michael Miller and Lt. Colonel George Noel. Lastly, I'd like to thank the other members of Dr. Borghetti's advisory group, especially Major Alexander Kamrud, for being wonderful sounding boards for the past 18 months.

James R. Elkins

Table of Contents

Page

List of Figures

List of Tables

COMPARISON OF MACHINE LEARNING TECHNIQUES ON TRUST DETECTION USING EEG

I. Introduction

1.1 Background and Motivation

Trust research has been a focus area of various fields for numerous years. Psychologists, economists, information systems researchers, and academic scholars from many other fields have investigated the topic due to the enormous impact that trust has on every aspect of society (Hosking, 2002). Lately, an increasing number of scientists have explored the topic of trust from a biological perspective. This research can be broken down into three primary groups: genes, hormones, and the brain (Riedl & Javor, 2012). The rapid development of intelligent machines across various applications has also increased trust research interest from an engineering and computer science perspective. This is because the relationship between humans and these machines requires a high level of collaboration, and therefore, a high level of trust. However, one problem that humans have is that they often provide too little or too much trust in an algorithm (de Visser et al., 2018). The ability to estimate human trust levels in real-time is a critical part of the continued advancement of human-machine teams so that the individual can become aware when this problem is occurring.

The military has been a large proponent of advancing and adopting new intelligent systems. Autonomous vehicles, intelligence monitoring systems, and command and control operations are just a few notable examples (Sayler, 2020). These technologies can help protect and serve, but faulty operation could also lead to disastrous

situations. To continue to roll out these new machines into the regular workforce, humans must operate them effectively.

1.2 Problem Statement

The present understanding of human trust in machines is limited. Some studies investigating the relationships between humans and computers have found that persons have the propensity to exhibit the same behaviors toward a machine as they do in inter-personal interactions (Madhavan & Wiegmann, 2007). If people tend to treat machines like they do another person, then it seems more than possible that they demonstrate trust in machines in a similar manner as well. If this is true, then much of the knowledge on interpersonal trust should also apply to human-machine trust, increasing the current information known.

Measuring trust is non-trivial. Until recently, the most common approach used was self-reporting surveys. One primary issue with these is that they are subjective measures which can have results that greatly deviate between different individuals. The development of low-cost and effective research-grade physiological sensor technologies, though, has created the opportunity for objective methods for assessing trust. A few standard physiological signals are electroencephalogram (EEG), electrocardiography, and electromyography. Cognitive trust, or making the conscious decision to trust, originates in the brain. Therefore, researchers are investigating neural correlates, which are brain activities that correspond with and are necessary to produce a particular experience, associated with trust. Because of this, the most frequently used type of physiological

signal for assessing trust is EEG, a measure of brain activity (Ajenaghughrure et al., 2020).

As of now, there is no single agreed-upon method for objective trust detection. Additionally, no method currently exists for monitoring trust in real-time. This study intends to advance the knowledge of trust detection and monitoring by using machine learning to classify trust and distrust in data from EEGs. It will investigate trust detection on single-participant, cross-participant, and cross-task cross-domain datasets. The latter type mentioned is a dataset with observations combined from two different experiments. The experiments used different tasks as well as studied different trust domains. One investigated human-human trust, and the other researched human-machine trust.

1.3 Research Questions

The objective of this research is to identify if a machine learning model can find a functional relationship between trust and features derived from EEG. To complete this objective, the following research questions are investigated using existing data. The goal for each research question is to develop a machine learning model with an equal-class-weighted classification accuracy greater than 70% for the described task. The 70% threshold was chosen as it is representative of the current best published results in the research area (Ajenaghughrure et al., 2020).

1.3.1 Research Question 1 - Same Task Trust Detection

Can EEG be associated with specific actions which indicate the decision to trust or distrust?

1.3.1.1 Part 1 - Human - Human Trust

Hypothesis: The brain activity from a participant choosing to trust that another human will do as they expect them to will be able to be distinguished from the brain activity of a participant choosing not to trust the other human.

Research Objective: Develop a machine learning model that receives EEG data from a human-human trust experiment and is able to determine if the participant chose to trust with an equal-class-weighted classification accuracy of greater than 70%.

1.3.1.2 Part 2 - Human - Machine Trust

Hypothesis: The brain activity from a participant making a choice to trust that an automated agent is working properly will be able to be distinguished from the brain activity of a participant choosing to not trust in the machine.

Research Objective: Develop a machine learning model that receives EEG data from a human-machine trust experiment and is able to determine if the participant chose to trust with an equal-class-weighted classification accuracy of greater than 70%.

1.3.2 Research Question 2 - Cross-Task Cross-Domain Trust Detection

Can a machine learning classifier accurately detect cross-task cross-domain trust using EEG?

1.3.2.1 Part 1 - Human-Human to Human-Machine Trust

Hypothesis: A machine learning model can successfully classify trust versus distrust observations using EEG when trained on human-human trust data and tested on human-machine trust data.

Research Objective: Develop a machine learning model trained on human-human EEG trust data and tested on human-machine EEG trust data, which can determine when a participant chose to trust with an equal-class-weighted classification accuracy of greater than 70%.

1.3.2.2 Part 2 - Human-Machine to Human-Human Trust

Hypothesis: A machine learning model can successfully classify trust versus distrust using EEG when trained on human-machine trust data and tested on human-human trust data.

Research Objective: Develop a machine learning model trained on human-machine EEG trust data and tested on human-human EEG trust data, which can determine when a participant chose to trust with an equal-class-weighted classification accuracy of greater than 70%.

1.3.3 Research Question 3 - Neural Correlates of Trust

Do machine learning models achieve significantly higher classification accuracies when provided with all possible features to begin with or when a subset of features is manually selected based on past research findings investigating neural correlates of trust?

Hypothesis: The machine learning models provided with observations with subsampled feature sets based on past research will achieve higher equal class-weighted balanced accuracies as compared to the models given observations with the full set of features as input.

Research Objective: Create three new datasets based on the features shown in Table 1 from the human-human dataset and provide them as input for the machine learning models. Then compare the equal class-weighted balanced accuracies achieved to those obtained by the models created for section 1.3.1.1.

Table 1. The published feature sets and their corresponding articles for research question 3

Study Name	Year	Findings/features used
Adaptive Probabilistic Classification(Akash, Reid, et al., 2018)	2018	Mean frequency on P4, C4, and P3, peak-to-peak value of C4 and C3, root mean square of Fz, energy of Fz, variation of Fz, correlation of C4 & P4, energy of beta band on P3, CZ, C3, and variation of beta band on P3, CZ, C3
Classification for Sensing Trust (Akash, Hu, et al., 2018)	2018	Mean frequency Fz, C3, and C4, peak-to-peak value on C3, energy of theta band on P3, variance of alpha band on P4, energy of beta band on C4 and P3, mean of beta band on C3, correlation on C3 & C4 and Cz & C4, and the net phasic component from GSR
Real-Time Sensing of Trust (Hu et al., 2016)	2016	High beta band on P4, POz, and C4, Mid beta band on C3, the mean frequency on C3, C4, and P4, the net phasic component and maximum phasic component from GSR, and response time

1.4 Methodology

For the first two research questions, two EEG datasets from past experiments were used. One was from an experiment that investigated human-human trust, and the other looked at human-machine trust. Six machine learning classifier types, logistic regression, linear discriminant analysis (LDA), quadratic discriminant analysis (QDA), random forest classifier (RFC), support vector machine (SVM), and artificial neural networks (ANN) were created. Each model was trained and then tested in four ways.

- Same-task within-participant
    - Train on some of the data from participant Z and test on other data from participant Z
- Same-task cross-participant
    - Train on data from participants in group X and test on data from participants in group Y
- Cross-task cross-domain cross-participant
    - train on human-human dataset, test on human-machine dataset
    - train on human-human dataset, test on human-machine dataset

The third research question was investigated using just the human-human dataset. For this, a review of past research that used machine learning for trust detection on EEG was conducted. Three papers were chosen to be compared. Based on the feature groups which lead to the best results in the three papers, manual feature selection was made on the observations in the human-human dataset. The three datasets were then used for the six machine learning models so that the within-participant and cross-participant

performances could be compared with each other as well as the models created for section 1.3.1.1.

1.5 Assumptions

To answer the proposed research questions, the following assumptions about the experiment design were made:

- Brain activity is different when deciding to trust something versus when deciding not to trust something
- The differences in brain activity of choosing to trust and distrust are captured by EEG
- All EEG recordings were accurate, and the equipment used to record them was not faulty
- When a participant made a choice in one of the experiments, the decision demonstrated trust or distrust and this label was correctly associated with each decision

1.6 Limitations

The two experiments, in which the datasets used for this study are from, are different in many aspects. Some of the primary ones are the task being conducted by the subjects, the subjects themselves, the devices used to record the data, and the researchers conducting the experiment. Ideal conditions for the tasks being investigated would be where the only difference in the two datasets is the trust domain, being human-human trust and human-machine trust. All of the additional differing variables could present

differences in the EEG data not directly associated with the trust domain, thus making the classification more difficult.

Another difficulty within this research is the amount of data. The total observations per participant range between 100 and 150 which is a very limited amount of data. Small datasets can lead to several issues for machine learning including difficulties in separating the data into properly sized train, validation, and test sets as well as overfitting. In addition to this, both datasets had relatively large class imbalances making it even more difficult for a classifier to learn the differences between the two types of observations.

1.7 Structure of the Document

Five chapters are part of this document. Chapter II is a literature review of all relevant fields for this research to include electroencephalography, trust, and machine learning. Chapter III details the methods used to attempt to answer the research questions proposed in Section 1.3. Chapter IV then presents and discusses the results from the machine learning models created. Chapter V concludes the work by summarizing what was accomplished and discussing areas for future work.

II. Background

2.1 Chapter Overview

This chapter provides an overview of electroencephalography (EEG), human trust, and machine learning. The section provides a brief description and important definitions for all three concepts. First is a brief background on EEG and how it can be used. Then trust is discussed in-depth with a specific focus on the neural correlates of trust as well as human-machine trust. Lastly, there is a background on machine learning, different classifiers, and using these tools with EEG.

2.2 Electroencephalography

EEG is the measurement of electrical activity in different areas of the brain and the recording of the signals. It is used to detect the activity of large groups of neurons that are functioning at the same time. This physiological measurement is often used in the medical field to diagnose or treat medical disorders such as brain tumors, strokes, or schizophrenia (Dvey-Aharon et al., 2015). EEG can also be a useful tool in identifying behaviors as it has a very high degree of sensitivity. This metric allows for the distinction between different cognitive processes (Cohen, 2014). The data offers both frequency of the signals and the location of the signals at any given time.

There are two types of EEG recordings, invasive and non-invasive. Non-invasive recordings typically involve a device, like a mesh cap, which is placed on top of the head. The device has electrode sensors attached to it that monitor and collect the brain's activity. Invasive means that the recordings occur with electrodes that have been

surgically implanted on the surface or within the depth of a brain. These typically provide more accurate readings but are generally reserved for specific scenarios. Non-invasive recording procedures are cheaper, safer, quicker, and still provide accurate data, so there is rarely a need to use invasive methods. One limitation of EEG is the signals' poor spatial precision due to electromagnetic attributes of the skull and tissue (Beres, 2017). The other primary constraint is its inability to record activity from non-surface structures of the brain accurately.

To standardize EEG data collected, the electrodes on the caps use one of three standard conventions for labeling and their placements. Figure 1 is the International 10-20 system. The numbers, 10 and 20, in the name refer to the fact that the actual distances between adjacent electrodes are either 10% or 20% of the total front-back or right-left distance of the skull. Additional common electrode configurations are the 10-10 system and the 10-5 system which are shown in Figure 2. The labels correspond to the 10-10 system and the additional dots represent the added electrode locations available in the 10-5 system (Jurcak et al., 2007).

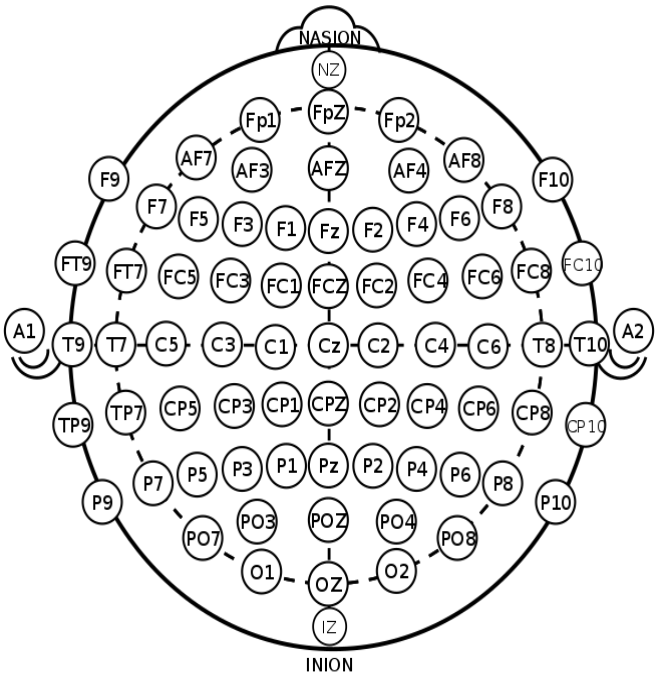

Figure 1. The International 10-20 system (Oxley, 2017)

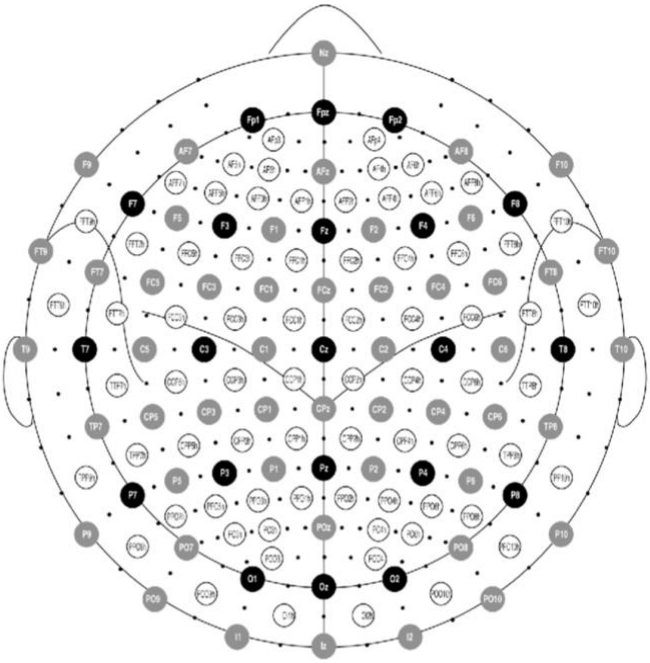

Figure 2. Electrode positions in the 10-5 system; dots indicate the additional positions as compared to the 10-10 system (Oostenveld, 2014)

2.3 Trust

Almost everyone finds trust to be crucially important, it is a vital part of any successful relationship, and it has been researched for centuries. Trust is a foundational aspect in society, exhibited by the numerous amounts of research performed in a diverse group of fields. From sociology to biology, computer science to economics, researchers from all different backgrounds have investigated this topic. Typically, each discipline has several perspectives on trust leading to numerous different definitions. However, a few comprehensive reviews on the topic have given some conceptually similar definitions. One defines trust as a trustor choosing a specific decision based on the subjective belief that a trustee will behave as anticipated by the trustor under situations of uncertainty (Cho et al., 2015). Coleman and Fishbein et al. defined trust as behavior that makes the trustor vulnerable based on the trustee's actions (Riedl & Javor, 2012). A common theme among most of the definitions is that trust is multi-dimensional. It includes many different subjective triggers such as a trustor's willingness to be vulnerable, benevolent, reliable, competent, honest, and open (Leichtenstern et al., 2011).

2.3.1 Neural Correlates of Trust

One further complexity of trust is identifying where in the brain it originates and how to identify when it is occurring. As with most neurological discoveries, experiments examine active regions in the brain when the researchers know that a participant is experiencing a specific function. With an iterative investment game as the stimuli and functional magnetic resonance imaging (fMRI) to measure brain activity, it was found that increased blood flow in the ventral striatum and medial prefrontal cortex (Figure 3)

were associated with a signal indicative of the desire to engage in collaborative interactions (Fareri et al., 2015). These areas are associated with decision making, reward-related behavior, conflict monitoring, error detection, executive control, and reward-guided learning (Euston et al., 2012). Since not everything that activates these parts of the brain deals exclusively with trust, more specific brain measures are needed to detect trust in real-time.

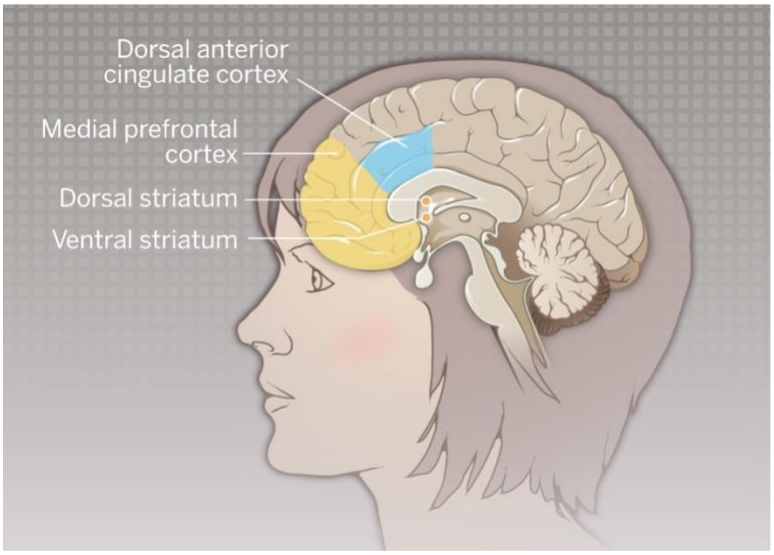

Figure 3. The location of the ventral striatum and medial prefrontal cortex (Hare, 2014)

One paper defines neural correlates as brain activity that corresponds with and is necessary to produce a particular experience (Dimoka et al., 2011). Activity in the ventral striatum and medial prefrontal cortex are considered neural correlates of trust. However, to be sure when someone is deciding to trust instead of any other decision involving these brain regions, more specific neural correlates are needed. Although several studies have

shown that brain function is associated with either trusting or distrusting others, very little is known regarding specific brain functions during the control of social attitudes, including trust and distrust. Recently, there has been an increased focus on this by studying physiological signals. At this point, though, there is no single generally accepted way to assess trust using them. Right now, EEG is the most frequently used central nervous system physiological signal sensor for assessing trust (Ajenaghughrure et al., 2020).

In pursuit of identifying the EEG-based neural correlates of trust, researchers from the University of New South Wales examined data from subjects collected while participating in an investment game (Wang et al., 2018). They computed three types of features from the signals using an autoregressive model, sample entropy, and Fourier analysis. The features that had the most significant correlation with trust were from the Delta and Gamma frequency bands with the largest effect sizes found at the Gamma band from the F5 and F3 electrode and the Delta band at the FC3 electrode. The electrode placement was based according to the international 10-5 system. The significant features found lead them to their broader findings of identifying the frontal area as the predominant brain area correlated with trust.

Purdue University also worked to find the best general set of physiological features to be used as input for a classification machine learning model (Akash, Hu, et al., 2018). This experiment placed their electrodes based on the 10-20 system. Past EEG studies have shown the importance of both time-domain features and frequency domain features to classify cognitive tasks successfully. The researchers from Purdue extracted an exhaustive set of both to start with. They then performed feature selection, which led

them to the physiological feature set detailed in Table 2. They used galvanic skin response (GSR) in addition to EEG. In a further study using the same dataset but a different feature selection algorithm, the researchers found a different optimal feature set, also shown in the table (Hu et al., 2016).

Other research has concluded that generally, the alpha and beta waves are significantly stronger in trust events, while the gamma band's power is more substantial with mistrust (Oh et al., 2017). These experiments used different stimuli, different participants, different feature selection techniques, and different EEG measurement devices, which could account for the different optimal feature sets found. A comparison of these results is important to progress toward a generally accepted feature set that can be used no matter the task. It is also important to note that these studies included those looking at both trust domains, human-human trust and human-machine trust. There are currently differing beliefs on the neurological signals shown between these two. It is expected that the overall trust process for people and automation is similar, but it is also likely that there exist specific differences between the two (Madhavan & Wiegmann, 2007). This research works under the hypothesis that they are similar enough for a machine learning model to treat them the same.

Table 2. Neural correlates of trust found from different research

Study Name	Findings/features used
Comp. Substrates of Social Value (Fareri et al., 2015)	There is increased blood flow to the ventral striatum and the medial prefrontal cortex during demonstrations of trust.
EEG-Based Neural Correlates of Trust (Wang et al., 2018)	The most statistically significant features were found for frequency band gamma from electrodes F5 and F3 as well as delta from FC3.
Adaptive Probabilistic Classification (Akash, Reid, et al., 2018)	Mean frequency on P4, C4, and P3, peak-to-peak value of C4 and C3, root mean square of Fz, energy of Fz, variation of Fz, correlation of C4 & P4, energy of beta band on P3, CZ, C3, and variation of beta band on P3, CZ, C3
Classification Model **for Sensing Trust (Akash, Hu, et al., 2018)**	Mean frequency Fz, C3, and C4, peak-to-peak value on C3, energy of theta band on P3, variance of alpha band on P4, energy of beta band on C4 and P3, mean of beta band on C3, correlation on C3 & C4 and Cz & C4, and the net phasic component from GSR
Real-Time Sensing of Trust (Hu et al., 2016)	High beta band on P4, POz, and C4, Mid beta band on C3, the mean frequency on C3, C4, and P4, the net phasic component and maximum phasic component from GSR, and response time
Study on Neurological Measure of Trust (Oh et al., 2017)	High mean frequencies of the alpha and beta band correlate with trust and a strong mean frequency of the gamma band correlates with mistrust.

2.3.2 Measuring Trust

As stated previously, there is no universally accepted way of measuring trust. Many proposed approaches from the past include a survey for a participant to self-report their trust levels. One such tool was designed by Kramer (Kramer, 1999) that used only six questions. It asked an individual to indicate their trust level with themselves as well as other members of the same group. The average score was found for each question, and together, they were used as a generalized measure of trust. Tools like this have led researchers to measure trust based on participants' survey responses before, during, and

after experiments. There are a few primary issues with using self-reporting measures to predict trust though.

- It is impossible to make predictions using these tools in real-time since it requires the participant to evaluate their behavior after it has already occurred.
- Beyond the scope of an experiment, it is not practical to continually ask humans to provide this sort of feedback.
- Any self-reporting measure is subjective, which means that the data can easily be skewed based on other factors.

Recent research in the area has focused on using physiological signals to estimate trust levels for these reasons. Even after data has been collected to measure trust, there is still a large spectrum of tools used to inspect it further. The majority of studies assessing trust using physiological signals though analyze the data with static approaches from statistics like the t-test or ANOVA. In a recent review of this field, only 17% of the studies examined used a dynamic technique like machine learning, even though these techniques are regularly used for high-level predictive analytics (Ajenaghughrure et al., 2020). This is likely because most researchers assessing trust either lack awareness of machine learning methods and their benefits or do not know how to use them. No matter the reason for the lack of dynamic techniques used in this field, it further demonstrates the gap between different disciplines researching trust.

2.3.3 Human-Machine Trust

The amount and importance of autonomous agents in society have rapidly been increasing for the past two decades. These agents need to be employed in domains not

amenable to conventional automation and which humans find difficult, dangerous, or otherwise undesirable to perform (Takayama et al., 2008). Automation is especially pressing in the time-critical and mission-critical applications of the United States military (USAF, 2010), transportation, and industry, where failure can lead to catastrophic consequences (Atkinson & Clark, 2013). Due to this, human interactions with automation have become a research area of significant interest (Sauer & Chavaillaz, 2018). For these interactions to go as planned, the operator needs to have an appropriate level of trust in the machine. The issue though, is that humans often have miscalibrated trust in automation (de Visser et al., 2018), meaning that they are very prone to either be too trusting or not trusting enough. With intelligent machines becoming a widespread norm in everyday life, it is important that operators can work with the machines effectively. Knowing the user's trust level is vital to ensure a proper relationship between a person and an autonomous machine (Jung et al., 2019).

Errors made by human-machine teams are categorized into two types, compliance and reliance errors. Compliance errors are when a human follows a recommendation made by a computer, even though they have knowledge or evidence that suggests the machine is wrong. Reliance errors, also known as misses, are when an individual does not act on something when they should because a machine did not notify them to do so (Chancey et al., 2017).

Even with a newfound focus of studies on human-machine trust, there is still little known on the subject. There is much more research investigating interpersonal trust. There is reason to believe that there is not much difference between human-human trust and human-machine trust from a neurological standpoint (de Visser et al., 2016).

Individuals exhibit a similar phenomenon where they reduce their effort when placed on a team, no matter if they are partnered with other humans or a machine. (Skitka et al., 1999). Researchers conclude from this that people will perceive automated agents as another team member or colleague, not just a computer (Parasuraman & Manzey, 2010). If an individual treats a machine similar to a human in teams, then it is likely that the neurological activity is similar in trusting between them as well.

2.4 Machine Learning

Machine learning is the study of computer algorithms that automatically improve at performing some task with increased experience. Models are created and then given data as input. Using statistics, the models find patterns in the data that can be applied to new datasets to make a prediction. These models are used for both regression and classification problems. Regression problems attempt to predict a numerical value, like estimating a home's cost based on data from other home sales. Models built to solve classification problems are supposed to correctly identify which value from a static set of categories a specific data belongs to. A classic example of this is an email filter, which decides if an incoming message should go to the regular inbox, junk mail, or a different folder. Another important distinction for machine learning models is if they are supervised or unsupervised, meaning if the dataset has the correct output associated with each piece of data or not.

2.4.1 Machine Learning with EEG

Due to EEG's complexity, it is difficult for humans to interpret all of the signal values themselves. However, machine learning models are much more suitable for modeling the functions of complex signals. Due to modern computing's processing power, they can do this at a much faster speed than any human as well. Machine learning for analysis on EEG is also superior to static techniques because they provide actual predictions based on the data instead of solely showing relationships between different variables (Ajenaghughrure et al., 2020).

EEG data to train machine learning models has successfully been used to classify different emotional states, cognitive tasks, and diagnose mental diseases. One study obtained an 88.67% classification accuracy on which cognitive task an individual was performing. The experiment used a feedforward artificial neural network (ANN) with a single hidden layer to distinguish between doing nothing, multiplication, letter composition, geometric figure rotation, and visual counting (Nuamah & Seong, 2017). Another paper describes a quadratic discriminant analysis (QDA) model that successfully classified an individual's emotional state after viewing different film clips with a better than chance accuracy (Lee & Hsieh, 2014). Support vector machines (SVM), logistic regression, and convolutional neural networks (CNN) have been used to successfully classify both patterns for seizure prediction and brain diseases (Piryatinska et al., 2017).

The wide array of successful machine learning models using EEG for different classification problems provides a reason to believe that these algorithms can be applied for many human behaviors and characteristics that can be traced back to brain activity. Given these successes and the fact that trust has been shown to correlate with increased

activity of specific regions in the brain, a machine learning model should be able to detect when a person is trusting. The achievement of a mean accuracy of 60.72%, on an unbalanced dataset, by using an ensemble of classifiers when determining if a subject was trusting of a simulated automated car's decision or not demonstrates that better than chance performance is possible (Hu et al., 2016).

2.4.1.1 Linear Discriminant Analysis

Linear Discriminant Analysis (LDA) is a machine learning classifier that uses a mathematically-linear discrimination threshold in feature space. As opposed to other linear classifiers, LDA can be used when there are more than two classes. Its parameter estimates are stable when the classes are well separated. It is stable when the number of observations is small and those observations are approximately normally distributed. LDA uses Bayes' theorem to approximate the output given an input after modeling each response class's distributions separately. This technique has long been one of the most popular methods for pattern classification. LDA had the second-best results in two recent studies that compared the performance of multiple different machine learning algorithms on classifying human behaviors using EEG signals (Binias et al., 2018; Regulinski, 1962).

2.4.1.2 Quadratic Discriminant Analysis

Quadratic Discriminant Analysis (QDA) classifiers are similar to LDAs except they are not limited to just linear boundaries between classes. Thus, QDAs are much more flexible than LDAs, which can be a benefit, especially when there is a small amount

of data and a high number of features. The flexibility does mean that QDAs can have a large variance when the data is highly dimensional. The bias to variance tradeoff is one of the primary factors when deciding to use an LDA or a QDA (Bhattacharyya et al., 2010).

2.4.1.3 Random Forest Classifier

Random forest classifiers (RFC) are tree-based learning methods that create several decision trees on bootstrapped training samples and then combine them to create a single qualitative predictor. They are a type of ensemble machine learning algorithm known as bagging. When building the decision trees, every time a split is considered, a random sample of m predictors are chosen as split candidates from the full set of p predictors. At every split, a new random set of potential predictors are chosen to be considered; this set is usually of the size $\sqrt{p}$. This offers an improvement over regular bagging because datasets with one or two very strong predictors, the first splits in the group of trees will likely be the same, leading to a group of similar trees. Random forests make it much more likely that the average number of created trees is less variable making the finished product much more reliable.

Like LDAs, RFCs have long been a popular method used for EEG classification. Notably, they have had the best results compared to other models when there is only a small amount of data. Ackermann et al. used RFCs to classify anger, sadness, and other emotions for recognizing depressive symptoms early (Ackermann et al., 2016). This method has also successfully been used to actively detect a pilot's reaction to unexpected events (Binias et al., 2018).

2.4.1.4 Artificial Neural Networks

Artificial neural networks (ANN) are another type of machine learning model. These are inspired by the way that signals are sent through synapses in the brain. ANNs are widely used when the features being looked at in the data have a complex relationship. They have been shown to do a great job at simplifying these relationships, so they are more understandable.

The original ANN is a fully connected neural network in which every neuron connects to every neuron in the next layer. Each layer in these networks receives data as input, applies a weighting to the values received, and then passes the sum of the resulting values to an activation function. ANNs improve their results by changing the weights applied to the data or modifying other parameters that are part of the system.

ANNs have been very successful in classifying EEG data. Yuen et al. used an ANN to classify five different human emotions from EEG data. The success rate of the model was over 95% (Yuen et al., 2013). In another study by Nuamah et al., five different cognitive tasks were classified using a neural network. Six people participated in this study, and on three of the sets of EEG data from the participants, the neural network had a 100% test set classification accuracy (Nuamah & Seong, 2017). Lastly, the model that performed better than the LDA in the study detecting pilots' reactions to unexpected events was an ANN (Binias et al., 2018).

2.5 Summary

Determining who and when to trust has been crucially important for as long humans have roamed the earth. The ever-increasing amount of collaboration and interaction between intelligent machines and humans significantly raises the need to be able to measure and detect trust. Beyond historically critical interpersonal situations, humans now face numerous additional situations where it would be beneficial to know the level of trust actively being exhibited. These include sitting in the driver seat of an autonomously driving car, flying as part of a swarm that is partially made up of drones, and making financial decisions for entire pension funds based on computer-generated recommendations. To achieve real-time measurement of trust, it is essential to be able to classify this behavior accurately. Using EEG along with machine learning has led to success in classifying other human cognitive behaviors making the hypothesis that successful classification of trust or distrust is possible.

III. Methodology

3.1 Chapter Overview

To ensure that human-machine teams are working together as they are designed to, it is imperative to actively monitor the trust state of the humans. Otherwise, reliance and compliance errors could outweigh the potential automation benefits. Before this can be accomplished, a system that can accurately and consistently detect trust needs to be created. Currently, trust is typically measured using post-task surveys, which are inadequate for two main reasons. The first issue is that the answers to surveys are subjective to the person answering them. The second is that one cannot simultaneously perform a task and divert their attention to respond to surveys about their current trust level. As an alternative, this research attempts to use machine learning classifiers given physiological signals as input to detect trust.

To begin this chapter, the research questions and hypothesis are presented. This is followed by a description of the two datasets used in this research and their corresponding experiments. Groups outside of AFIT collected these datasets, but the rest of the research discussed in this chapter is original research by the thesis author. Next, the steps taken to analyze the data and the machine learning approaches used are discussed. The last section is a summary of the chapter.

3.2 Research Questions

The objective of this research is to identify if a machine learning model can find a functional relationship between trust and features derived from EEG. To complete this

objective, the following research questions are investigated using existing data. The goal for each research question is to develop a machine learning model with an equal-class-weighted classification accuracy greater than 70% for the described task. The 70% threshold was chosen as it is higher than any other published results in the research area (Ajenaghughrure et al., 2020).

3.2.1 Research Question 1 - Same Task Trust Detection

Can EEG be associated with specific actions which indicate the decision to trust or distrust?

3.2.1.1 Part 1 - Human - Human Trust

Hypothesis: The brain activity from a participant choosing to trust that another human will do as they expect them to will be able to be distinguished from the brain activity of a participant choosing not to trust the other human.

Research Objective: Develop a machine learning model that receives EEG data from a human-human trust experiment and is able to determine if the participant chose to trust with an equal-class-weighted classification accuracy of greater than 70%.

3.2.1.2 Part 2 - Human - Machine Trust

Hypothesis: The brain activity from a participant making a choice to trust that an automated agent is working properly will be able to be distinguished from the brain activity of a participant choosing to not trust in the machine.

Research Objective: Develop a machine learning model that receives EEG data from a human-machine trust experiment and is able to determine if the participant chose to trust with an equal-class-weighted classification accuracy of greater than 70%.

3.2.2 Research Question 2 - Cross Task Cross Domain Trust Detection

Can a machine learning classifier accurately detect cross-task cross-domain trust using EEG?

3.2.2.1 Part 1 - Human-Human to Human-Machine Trust

Hypothesis: A machine learning model can successfully classify trust versus distrust observations using EEG when trained on human-human trust data and tested on human-machine trust data.

Research Objective: Develop a machine learning model trained on human-human EEG trust data and tested on human-machine EEG trust data, which can determine when a participant chose to trust with an equal-class-weighted classification accuracy of greater than 70%.

3.2.2.1 Part 2 - Human-Machine to Human-Human Trust

Hypothesis: A machine learning model can successfully classify trust versus distrust using EEG when trained on human-machine trust data and tested on human-human trust data.

Research Objective: Develop a machine learning model trained on human-machine EEG trust data and tested on human-human EEG trust data, which can determine when a participant chose to trust with an equal-class-weighted classification accuracy of greater than 70%.

3.2.3 Research Question 3 - Neural Correlates of Trust

Do machine learning models achieve significantly higher classification accuracies when provided with all possible features to begin with or when a subset of features is manually selected based on past research findings investigating neural correlates of trust?

Hypothesis: The machine learning models provided with observations with subsampled feature sets based on past research will achieve higher equal class-weighted balanced accuracies as compared to the models given observations with the full set of features as input.

Research Objective: Create three new datasets based on the features shown in Table 1 from the human-human dataset and provide them as input for the machine learning models. Then compare the equal class-weighted balanced accuracies achieved to those obtained by the models created for section 3.2.1.1.

3.3 Datasets

This thesis uses two datasets that were collected from past research. The following section of this chapter will describe the experiments performed to obtain the

data as well as information about the participants, the physiological recording devices used, and the data collection.

3.3.1 Experiment 1 - Human-Human Trust

Fu et al. from Fuzhou University in Fuzhou, China, collected EEG data from subjects participating in a modified version of the trust game (Fu et al., 2019). Originally named the investment game by its creators Berg et al. (Alós-Ferrer & Farolfi, 2019), this is a classic game paradigm for studying trust behaviors in laboratories. In the trust game, there are two players, the trustor and the trustee. For the experiment, the participants always acted in the trustor's role and were told that the trustee would be the same person the entire time. Ten game points were given to both sides at the beginning of each round. The trustor then had to decide to send their points to the trustee or to keep them for that round. If the latter decision was made, the round ends with both players having 10 points. If the trustor sent their initial endowment, the 10 points were tripled during the transfer process. Therefore, the trustee was then given 30 points in addition to the 10 points they were given to start. The trustee then decides whether to keep all 40 points for themselves or send 20 points back to the trustor. The game's goal is to have as many points as possible, not to have more points than the other participant. When the trustor sends their points, they open themselves up to be exploited by the trustee's decision, which is a behavioral operationalization of trust. During the initial instructions, participants were told that the trustee was an adult randomly selected from a large pool and that the experimenter had sampled and interviewed them before the experiment. In reality, the trustee was computer software that was preprogrammed so that its decision was random

across rounds but that overall, the trustor would receive 20 points if they had made the initial trusting decision 50% of the time. Twenty self-reported right-handed, to exclude potential EEG difference caused by hemispheric dominance, healthy Chinese undergraduate and graduate students were recruited for the experiment. Eleven of the participants were female and the other nine were male.

3.3.1.1 Procedure

To begin, the experimenter described the rules of the trust game in detail. To ensure the game was taken seriously, an emphasis was put on each person's performance as they were told that their compensation would be tied to the total number of points they had at the end of the game. Each participant was actually paid a flat rate that came out to about 8 U.S. dollars. Each person was then seated comfortably one meter from a computer monitor and was fitted with an electrode cap in an electromagnetically shielded room. They then participated in a practice session of 10 rounds to familiarize themselves with the procedures and make sure that the instructions were clearly understood.

For the experiment, the participants completed 150 rounds of the trust game. During this time, their brain activity was being recorded using an EEG device. Each round started with a decision tree that showed the possible outcomes based on which decision was made, as seen in Figure 4. It was shown for 1500ms. A fixation cross, shown in Figure 5, was then displayed for a variable amount of time between 800 and 1000ms. Figure 6 is the third slide, which had a picture indicating the two options that the trustor could make. This next screen was shown after either the trustor made their decision or 2,000ms following the third slide, whichever occurred first. The participants

were instructed to press “1” on the keyboard if they wanted to keep their initial endowment or hit “3” if they chose to send their points for that round. If 2000ms passed before a key had been pressed, a warning message was displayed indicating that he/she responded too slowly, and the round was skipped. The next screen was a plain black slide that appeared for 800 to 1200ms. The last two slides showed the outcome of the current trial and then a message displaying the participant’s total score. These were shown for 1,200 and 2,000ms respectively. For the current trial outcome, a 0, 10, or 20 was displayed in large yellow lettering. Figure 7 shows one example of this. The final slide’s message, written in Chinese, said the current trial number and the participants’ total points as in Figure 8.

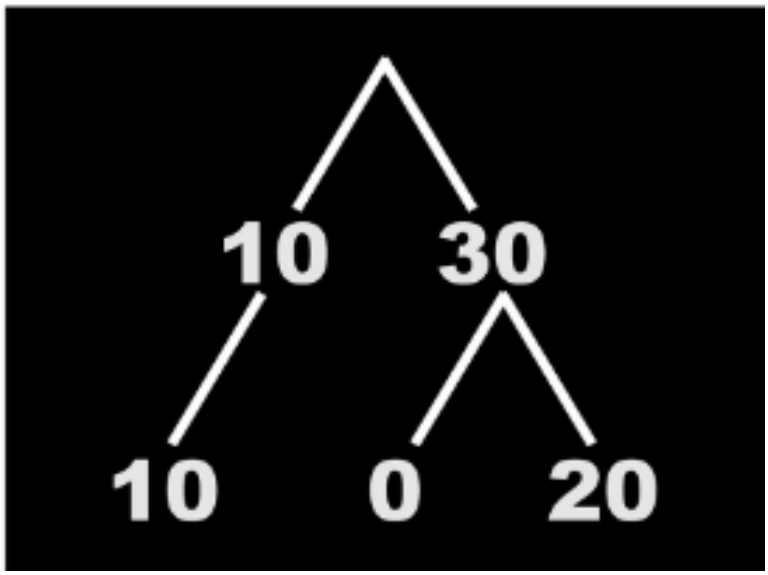

Figure 4. A simplified decision tree showing the possible outcomes for the round.

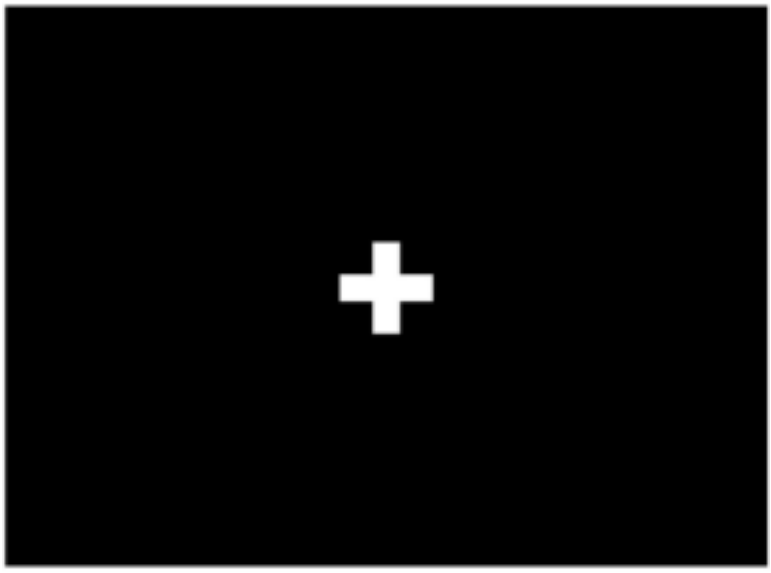

Figure 5. A fixation cross shown between the initial screen and the decision screen.

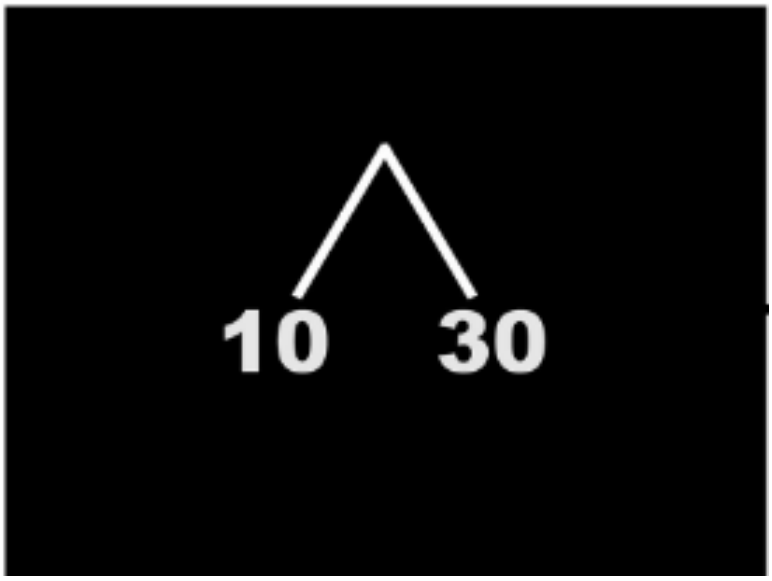

Figure 6. The decision screen showing the options that the participant may choose.

Figure 7. The outcome of the current trial is displayed. The options were either 0, 10, or 20.

Figure 8. The cumulative total slide which reads "The current trial is #3. Your total points so far are 30" in Chinese.

3.3.1.3 EEG Recording

This study used a modified 10-20 system electrode cap from Neuroscan Inc. with 64 electrodes. All of the electrode sites were first cleaned with alcohol and an impedance less than 5 kΩ between the electrodes and scalp was found and maintained throughout the experiment. The EEG was recorded using a 0.05-100Hz bandpass filter and was continuously sampled at 1000Hz with the right mastoid reference and a forehead ground. The EEG data was collected using Neuroscan software. Triggers were collected by the same software to mark the occurrence of different events from the experiment.

3.3.2 Experiment 2 - Human-Machine Trust

The human-machine trust dataset comes from research done at Purdue University by Akash et al., (Akash, Hu, et al., 2018) who set out to identify physiological features significantly correlated to human trust in intelligent systems. They recruited 48 self-reported right-handed individuals to take part in the study. The participants were told to imagine driving a car that was equipped with an image-based obstacle detection sensor. During each trial, a slide was presented saying that the sensor detected an object or that the road was clear. The subjects then had to decide whether to trust if the automated tool was correct or not.

3.3.2.1 Procedure

At first, each participant read and signed an informed consent document. Then, they were measured for and equipped with the EEG headset. After appropriately fitted with the cap, they completed a 9-minute baseline task created by Advanced Brain

Monitoring. Upon completion, they were instructed to interact with the human-machine interface (HMI). The instructions about the experimental task were then given. The subjects were told to imagine they were driving a car equipped with an image-based obstacle detection sensor. The sensor would detect obstacles on the road, and the participant would evaluate the report and choose to either trust or distrust the algorithm's findings. Each participant conducted four practice trials, read the instructions on their own, and then had a chance to ask any questions before starting the experiment.

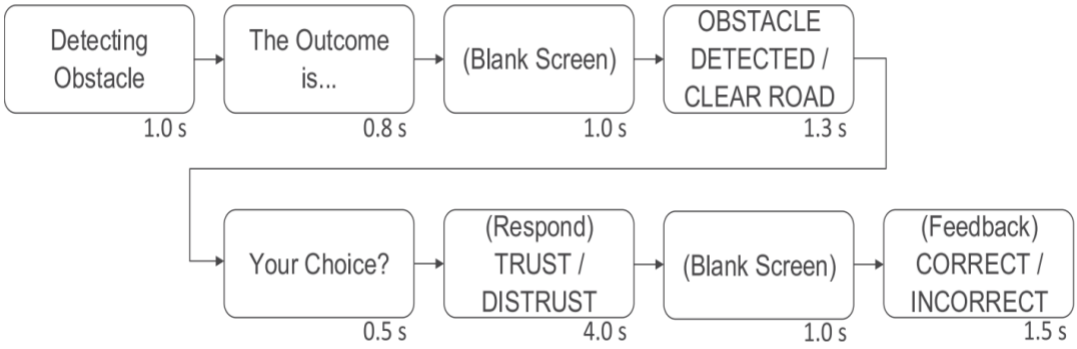

Figure 9. The sequence of events for each trial.

Each trial included eight screens that together lasted 11,100ms (Figure 9). Screen 1 displayed black text that read "Detecting Obstacle" for 1000ms and was followed by a slide that said "The Outcome is…" for 800ms. After a blank screen was shown for 1000ms, the stimuli slide was presented (Figure 10). This screen had the sensor's decision on it and was either "Obstacle Detected" or "Clear Road". After a quick 500ms slide with "Your Choice?" on it, the response slide, shown in Figure 11, appeared for

4000ms, asking the subject to make their decision. Lastly, the feedback screen was displayed for 1500ms (Figure 12) following another 1000ms blank screen. This final slide told the participant if they made the right choice and also had an image showing a clear road or an obstacle in front of the car.

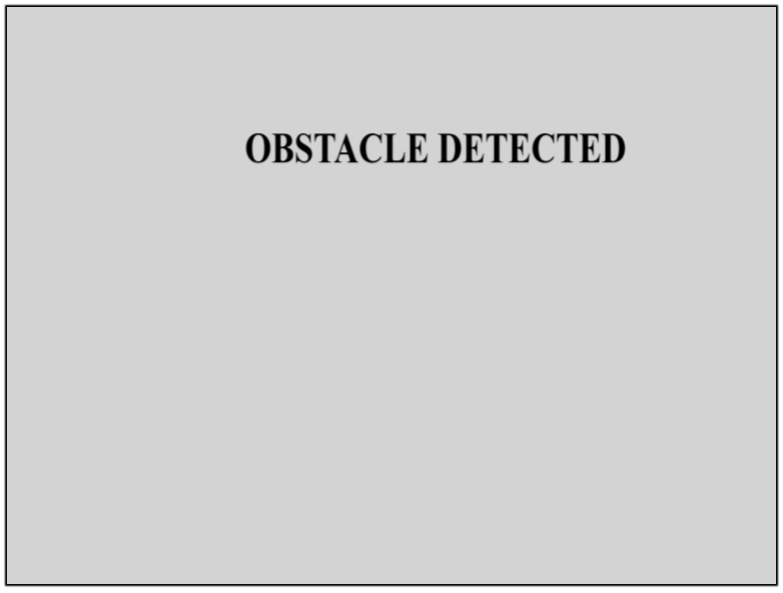

Figure 10. One of the two possible stimuli slides. The other said "clear road" on it.

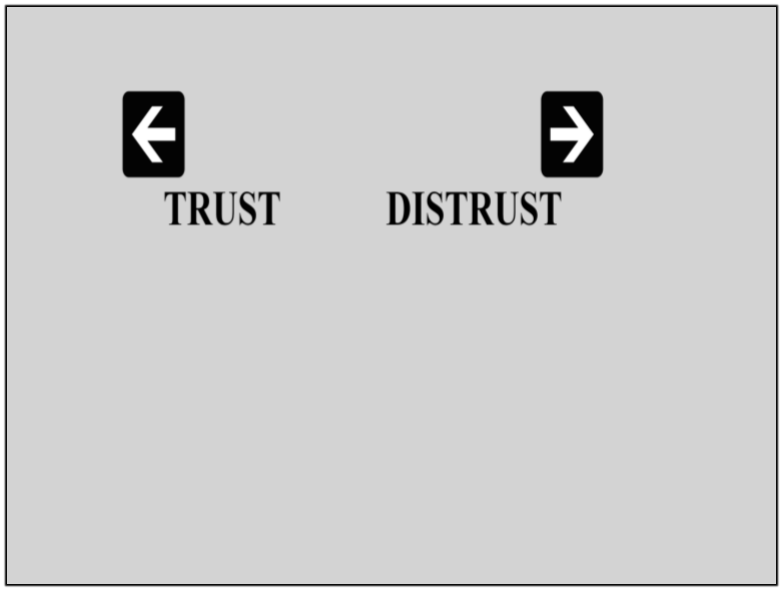

Figure 11. Screenshot of the response slide.

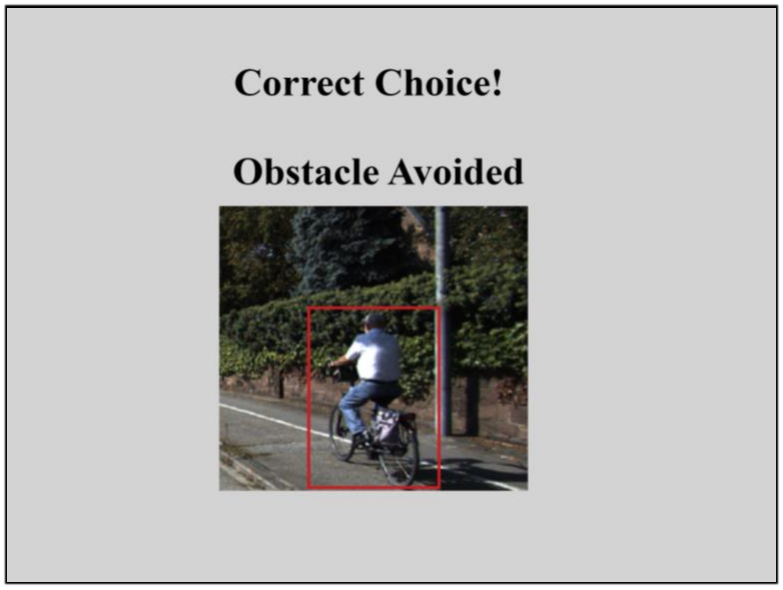

Figure 12. Example screenshot of the feedback slide.

The performance of the sensor was varied to elicit dynamic responses. Trials fell into one of two categories, reliable or faulty. In reliable trials, the algorithm was correct 100% of the time. Faulty trials had a 50% probability that the automated agent inaccurately identified the road condition. The experiment included 100 trials in total, divided into three phases, called databases, shown in Figure 13. Participants were randomly split into two groups to counterbalance any possible ordering effects. Databases 1 and 2 included either 20 reliable (A) or 20 faulty (B) trials. The third database used a pseudo-random binary sequence to switch between the two types of trials.

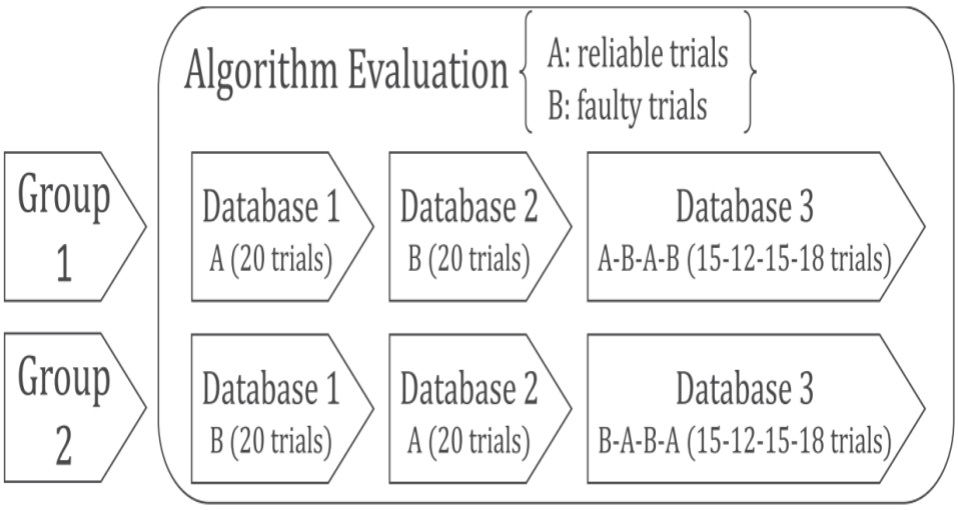

Figure 13. The organization of the databases according to the two possible groups participants could be placed in.

3.3.1.3 EEG Recording

The experiment used a B-Alert X-10 9 channel EEG device, which uses electrode placements based on the 10-20 system. The contact impedance between the skin and

electrodes was kept below 40 kΩ. The EEG was recorded at a 256Hz frequency from nine scalp sites and were referenced to the mean of the left and right mastoids.

Note: The remainder of the research discussed in the methodology is original research by the thesis author.

3.4 Machine Learning Procedures

3.4.1 Data Pre-Processing

Both datasets were processed using the 2019 version of EEGLAB, an interactive MATLAB toolbox used for processing continuous and event-related EEG. The software used in the human-machine experiment automatically decontaminated the signals for artifact removal. The decontamination process minimizes the effects of electromyography, electrooculography, spikes, saturation, and excursions. The data from the human-human experiment is the EEG in its completely raw form. At this point, both datasets were processed following the steps in the PREP pipeline (Bigdely-Shamlo et al., 2015). A summary of these steps is provided below.

1. The human-human data was down-sampled from 1000Hz to 256Hz. The human-machine data was initially sampled at 256Hz so it was not down-sampled further. Down-sampling is done to compress the data size and cut off unnecessary high-frequency information.
2. The data is put through a high-pass filter at 1 Hz using a basic finite impulse response (FIR) filter. This removes the baseline drift.

3. Line noise is removed using the CleanLine EEGLAB plugin. This process uses a sliding window to adaptively estimate sine wave amplitude to subtract.
4. The information about the channel locations is imported to allow for channel re-referencing. Both studies used electrode placement based on the International 10-20 system.
5. Channels were then inspected for rejection. They were marked as bad and then removed if they had a Z-score over five after using kurtosis.
6. The channels which were removed were then interpolated using spherical interpolation to prevent bias when re-referencing.

After the pre-preprocessing steps, the datasets were epoched into segments of 1500ms ending at the moment that a subject made their decision. In the human-human experiment this was when they pressed the "1" or the "3" on the keyboard and in the human-machine experiment this was when they pressed the left or the right arrow on the keyboard. During this step, the data was also separated into two files, one with epoched data from observations marked as trust and the other from distrust observations. Following this, a visual inspection of the data occurred to reject any segments that had noticeable large amounts of noise in comparison to the others.

3.4.2 Signal Feature Extraction

Features represent a specific property, a recognizable measurement, and a functional component obtained from a section of a pattern (Al-Fahoum & Al-Fraihat, 2014). Turning the continuous EEG data into a set of features is necessary to complete

many machine learning tasks. Since EEG signals are non-stationary, it is necessary to transform the data to a time-frequency domain. That way, knowledge about both domains is associated with the final values (Dhiman et al., 2013). In this study, both datasets were transformed to the time-frequency domain using a hybrid method that is a cross between pure wavelet analysis and short-term Fourier Transform. This method is proposed by Mike X Cohen and is regularly used for EEG signal analysis (Cohen, 2014). This mathematical operation uses a tool known as the Fast Fourier Transform along with a family of complex Morlet wavelets to generate values for the five traditional EEG bands. These bands and the corresponding frequencies they cover are delta (1-6 Hz), theta (7-11 Hz), alpha (12-15 Hz), beta (16-22 Hz), and gamma (23-30 Hz). Afterward, the mean of the power spectral density was found for each frequency band leading to five features per electrode for each epoched observation.

3.4.3 Classifier Descriptions

Six different machine learning models, along with a naïve model, were created to test the accuracy of classifying trust vs. distrust. The models were built in Jupyter Notebook version 6.0.3 using Python 3.8.3. The machine learning packages used were `keras-gpu` version 2.3.1 and `scikit-learn` 0.23.2. The Matthews correlation coefficient (MCC), found with sci-kit learn's `matthews_corrcoef` method, was used for parameter tuning instead of simply basing it off the highest accuracy. It was chosen as the validation metric because it is a more reliable statistical rate when accuracy on both classes matter. The MCC produces a high score only if the prediction obtained good results for true positives (TP), false negatives (FN), true negatives (TN), and false

positives (FP) (Chicco & Jurman, 2020). These four terms are depicted in the confusion matrix displayed in Figure 14.

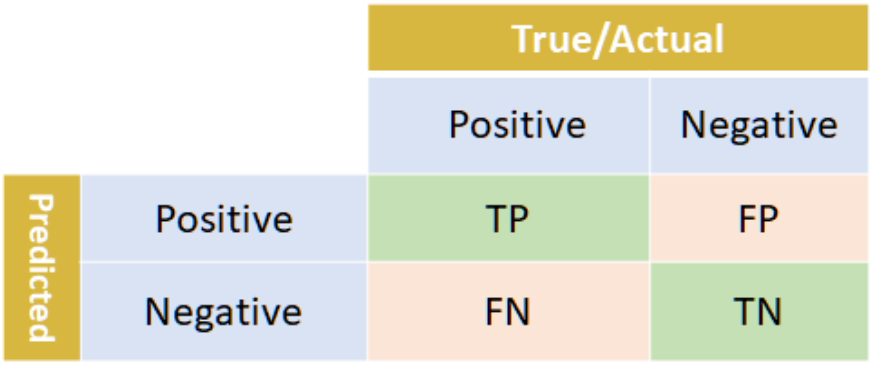

Figure 14. Confusion matrix depicting TP, FN, TN, and FP are (Shmueli, 2019)

3.4.3.1 Logistic Regression

A logistic regression model was built due to its ease of implementation as well as the fact that the trained weights of the features can often provide inference about the importance of them. Both the '`C`' and '`penalty`' parameter values were tuned during training. The options for the penalty were '`l2`' and '`none`'. The C value, which is the inverse of regularization strength, had 20 evenly spaced values on a log scale from -4 to 4. The model used '`lbfgs`' as the '`solver`' parameter and the maximum number of iterations, '`max_iter`', taken for the solver to converge was set to '`1500`'. Additionally, the class weights of each training set were provided as one of the model parameters.

3.4.3.2 Linear Discriminant Analysis

An LDA was used as it typically helps reduce high-dimensional data onto a lower-dimensional space, which is especially beneficial when there are many more features than observations. The LDA was composed with '`lsqr`', or least squares solution, as the '`solver`' parameter, and the '`shrinkage`' parameter value tuned with 100 evenly spaced values from 0 to 1 as well as the '`auto`' setting which uses the Ledoit-Wolf lemma. Shrinkage adds a penalty to coefficient estimates to help regularize the models. This technique discourages learning a more complex or flexible model to help avoid overfitting.

3.4.3.3 Quadratic Discriminant Analysis

In one study trying to classify trust using EEG, a QDA outperformed all of the other basic models tested. Datasets with a large number of features like those used in this thesis, the QDA's flexibility should be very beneficial. Scikit-learn's '`QuadraticDiscriminantAnalysis`' class only has one parameter to tune. This is the regularization parameter, which regularizes the per-class covariance estimates by transforming the scaling attribute for the given class. The model was tuned with values for '`reg-param`' of '`0.0`', '`0.00001`', '`0.0001`', '`0.001`', '`0.01`', and '`0.1`'.

3.4.3.4 Support Vector Machine

The support vector machine (SVM) scales relatively well for high dimensional data and is one of the most frequently used models for classification with EEG. It typically does not perform well when target classes are overlapping though so extremely

accurate results were not expected from this approach. ‘C’, ‘kernel’, and ‘gamma’ were the three parameters tuned. C, the regularization parameter, had ‘0.001’, ‘0.01’, ‘0.1’, ‘1’, and ‘10’ as possible values. Gamma, the kernel coefficient examined ‘0.001’, ‘0.01’, ‘0.1’, ‘1’, ‘scale’, and ‘auto’. Then the ‘linear’, ‘poly’, ‘rbf’, and ‘sigmoid’ settings were the different kernel types tuned. The class weights for the training sets were also provided for each model.

3.4.3.5 Random Forest Classifier

RFCs are built so that the important features are naturally found and often lead to the highest classification accuracy over many other models. ‘Max_samples’, ‘n_estimators’, ‘max_depth’, and ‘max_features’ were the hyperparameters chosen to be tuned. Ten evenly separated values between 0 and 1 were the percentage of observations used to train the base estimators, ‘Max_samples’. The tuned number of trees, or estimators, were values from 10 to 500 in increments of 10. The RFCs were trained and evaluated with maximum depths of 1 to 25. The recommended number of features to consider for an RFC is the square root of the feature count (James et al., 2013). For the human-human dataset this is $\sqrt{320} = 17.8$, so 18 features and for the human-machine dataset this would be $\sqrt{45} = 6.7$, so 7 features. To ensure the optimal number of features were used, a sweep of 20 and 10 features, respectively were used. Each model was also trained with the class weights as a parameter.

3.4.3.6 Artificial Neural Network

A fully connected neural network was chosen as a relatively simple deep-learning model to implement that has been shown to be able to classify EEG in other experiments with high accuracies. The '`Sequential`' method from Keras was used to build out the models. Each ANN was tuned on the number of hidden layers, the dropout rate of the first layer, and the number of hidden nodes in each hidden layer. The different dropout rates tested were '`0.1`', '`0.2`', and '`0.3`'. There were 1 to 4 hidden layers added with either '`128`', '`256`', or '`512`' nodes.

The first layer of the ANN was created as a fully connected Dense layer with one of the different amounts of node options and the '`relu`', or rectified linear unit (ReLU), activation function. One of the different dropout rates was then applied to this layer. Afterwards, the hidden layers were added each with the same number of nodes as the first layer as well as the same activation function. Each of these layers had a dropout rate of '`0.2`'. Batch normalization was then applied for every hidden layer using Keras's '`BatchNormalization`' method. The last layer is another fully connected Dense layer with a single output value and a '`sigmoid`' activation function. The sigmoid function squashes the output to a value between 0 and 1 which represents the probability that the input belongs to the trust class. Figure 15 is a diagram of the final network.

After built, each model was compiled with the Adam optimizer algorithm and '`binary_crossentropy`' as the loss function. This computes the cross-entropy loss between true labels and predicted labels for output values with two classes. Each ANN was then trained given the class weights with batch sizes of 32 and went through 30 epochs. Model training used a validation loss based early stopping method with a

‘`delta`’ of ‘`0.001`’ and a ‘`patience`’ of ‘`10`’. As an example of a resulting model, the ANN from participant AS08 in the human-machine dataset was optimized with 256 nodes per layer, 3 hidden layers, and a dropout rate of 0.1. This model had 6,144 parameters in the first layer, 65,792 parameters in each hidden layer, 1024 parameters for every batch normalization layer, and 257 parameters in the final layer, totaling a model with 206,849 parameters.

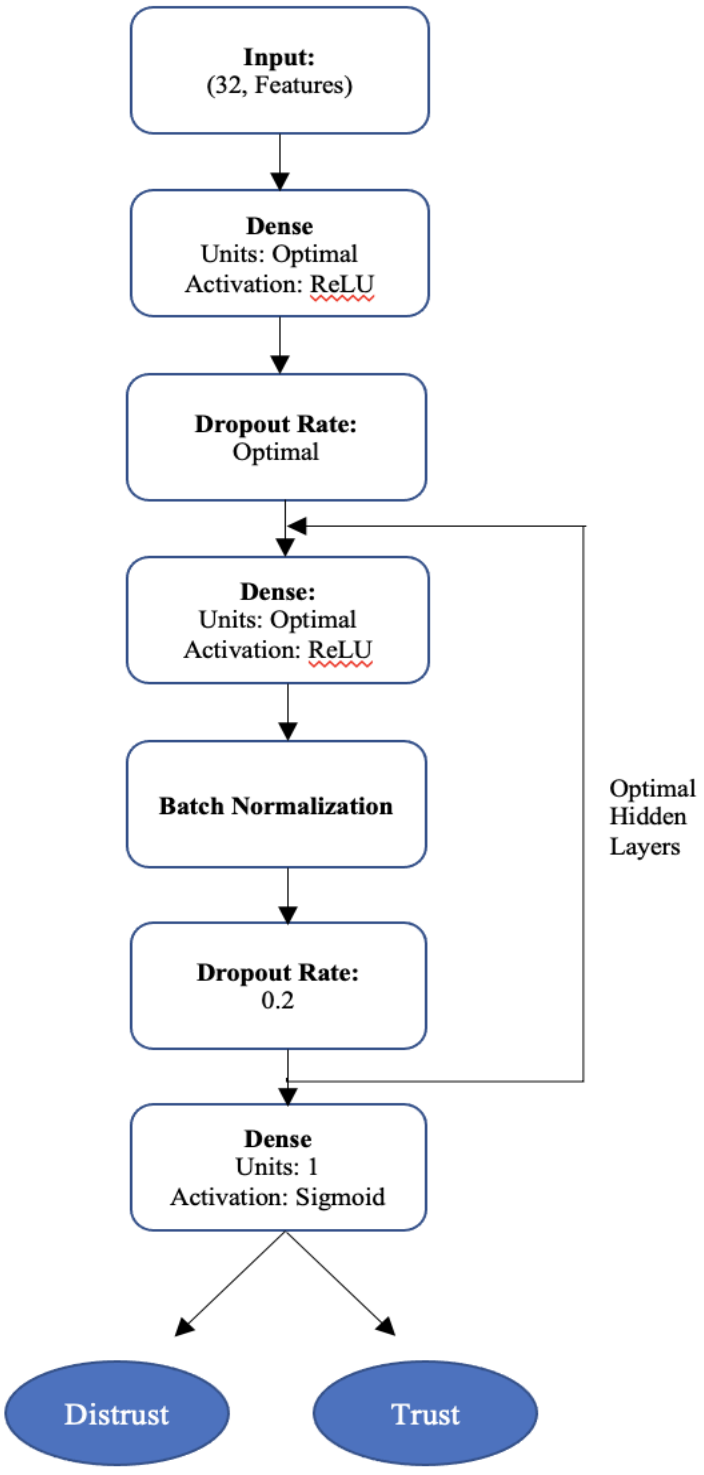

Figure 15**. Diagram of the fully connected network architecture. The number of input features was determined by the outcome of the PCA method on each dataset.**

3.4.4 Experiment Procedures

This section describes the experiments used in the thesis. Table 3 breaks each part down based on what dataset the training set and test set were from.

Table 3. The simplified look at how the datasets were used for each experiment.

Part	Training Dataset	Testing Data
A	Human – Machine	Human - Machine
B	Human - Human	Human - Human
C	Human – Human (subset)	Human - Human (subset)
D	Human – Machine	Human - Human (subset)
E	Human - Human (subset)	Human - Machine

3.4.4.1 Handling of Class Imbalance

Inspection of both datasets led to a discovery that was not ideal for achieving high balanced test accuracies with machine learning. The human-human and human-machine datasets originally had 69.81% and 75.25% observations in the 'trust' class, respectively, which are relatively large class imbalances. To mitigate this imbalance, a few different techniques were used. The first of which was removing any single participant in either dataset with a trust class of 80.00% or more of the total observations. One subject in the human-human, and seventeen participants in the human-machine dataset fit this criterion. After removing these subjects from the datasets, the trust classes were 68.41% in the human-human dataset and 70.45% in the human-machine dataset. Additionally, data on the class weights were provided to each of the models, which would accept it as a parameter. These were the logistic regression classifiers, the SVMs, RFCs, and the ANNs.

3.4.4.1 Single Dataset Tests

The first three tests were comparisons of classification accuracy across the different models using single-participant data from all of the datasets. Each participant's data was split into five evenly sized groups, four of these were used for 4-fold cross-validation training, and the last was kept separate to be the test set. During the hyperparameter tuning of each model, observations from three groups were combined and given to a model as input to learn from. The model then predicted the fourth set's values, and the MCC was found for that prediction. This process occurred four times, where the validation set was different each time. The average of the four MCCs was then computed and stored. The maximum of the set of average MCCs corresponded with the optimal parameters for each model. These were then used to train a new model provided with all four training folds as input. This model then predicted output values for the test set, and the results were evaluated.

After the within-participant models, cross-participant classifiers were made for every dataset. For these models, the train and test sets were created differently than with the single participant models. One of the primary purposes of cross-participant models is to see if an algorithm can learn enough from a group of participants' data to accurately classify observations from a completely unseen individual. If the sets had been randomly made from all of the observations, then it is very likely that every observation in the test set was from a participant whose data was also in the train and validation sets. Instead, the participants were randomly split into five groups. Four of these groups were used for cross-validation training as in the single participant tests. For testing, though, the participants' data in the final set were tested individually. This way, it could be seen if

any single subject's data could be predicted with high accuracy as well as what the overall average accuracy was upon all of the subjects in the test set.

Due to the large number of features and the relatively low number of observations, principal component analysis (PCA) was applied to all of the data to reduce the dimensionality. PCA transforms a large set of variables into a smaller one, containing a set amount of information of the full set (George, 2012). The new features are selected based on the variance that they cause to the output. The original features are transformed into principal components, which are linear combinations of the existing features. A few limitations of PCA are that the single features become less interpretable, and that information is lost in the process. To do PCA, the data was first standardized by scaling it with sklearn.preprocessing.StandardScaler and then transformed with sklearn.decomposition.PCA. The variance explained with the new components was set to 95% of the data's original variance. For the experiments, PCA was fit based on the 80% of the data used in the cross-validation training. The StandardScaler and PCA objects were then stored so that the same objects were used to transform the data for training as well as testing.

One thing to note is that the human-human trust experiment data used an EEG cap with 64 electrodes, but the human-machine trust experiment had a cap with only 9 electrodes. To combine the data for the further experiments, a subset of the human-human dataset was taken that included only the signal values from the same 9 electrodes that were on the cap in the human-machine experiment. The locations of these 9 nodes are highlighted in Figure 16. This new dataset (human-human 9-channel) was taken from

the raw data and then was pre-processed in MATLAB using the PREP pipeline. This is the dataset that was used as input for the classifiers which are shown in part C of Table 3.

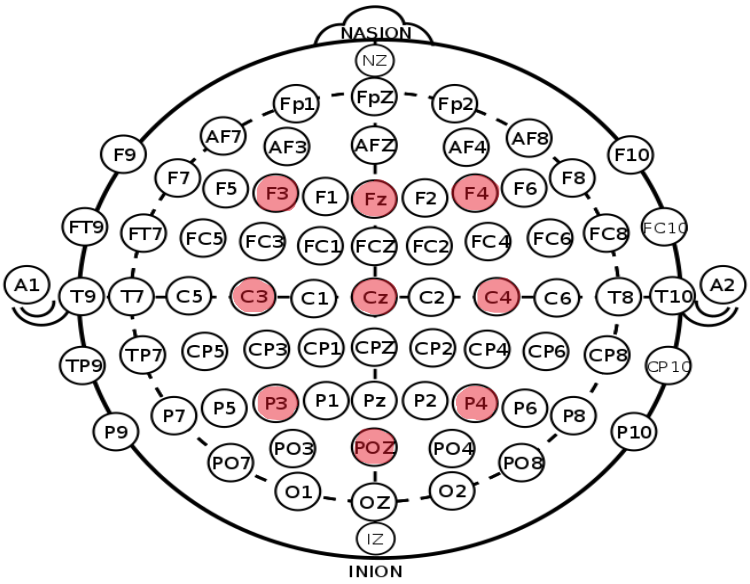

Figure 16. The highlighted nodes are those which are part of the human-machine dataset. The entire picture is the scalp locations for the International 10-20 system (Oxley, 2017)

3.4.4.2 Cross Task Cross Domain Tests

The first tests done were to establish if machine learning classifiers could successfully classify trust versus distrust using EEG as input. This section describes the experiments conducted to see if any of these same models can learn from data in one trust domain and accurately predict observations from another trust domain. How the human-machine and human-human 9-channel datasets were used for each test are shown by rows D and E in Table 3.

Each of these tests consisted of two sections that differed by what was used for the validation sets. The first method of the tests split the participants in each dataset into four groups. For the human-machine dataset, there were eight subjects' observations in three groups and seven in the final group. For the human-human 9-channel dataset, there were five participants' data in three groups and four in the final group. Four-fold cross-validation was used for hyperparameter tuning for each model, just as in the earlier experiments. After determining the optimal parameters for each model, predictions were made for each participant's observations in the other dataset. The second method evaluated used the entire dataset from one domain as the training set for each model. The participants in the second dataset were then randomly split into two evenly sized groups. One group's observations were used as the validation set to find each model's optimal parameter values. The participants in the other group were used as the different test sets. A diagram of the two processes for the models that were trained on the human 9-channel dataset and tested on the human-machine dataset is shown in Figure 17. The two different methods were used to see if a difference in balanced accuracies occurred when a model was validated using data from the same domain as the training set versus the test set if the training set and test set are from different domains. The use of the datasets for each of the four tests are shown in Table 4.

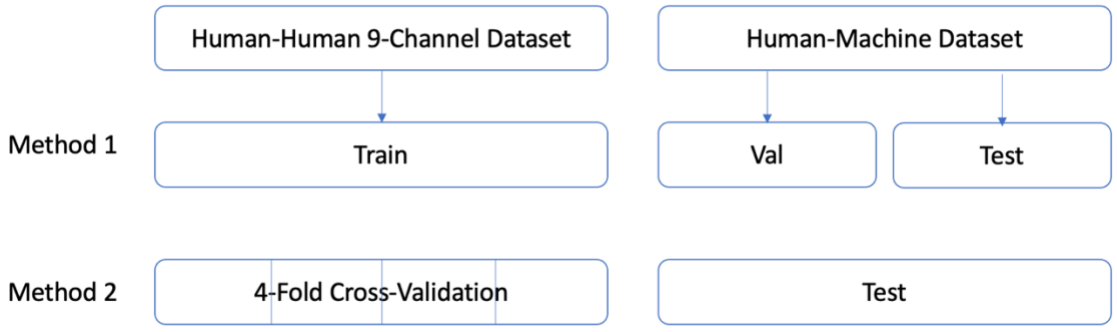

Figure 17. The two train, validate, and split processes for the cross-task cross-domain experiments

Table 4. The train, validation, and test sets for each of the four cross-participant cross-domain tests.

Test	Train Set	Validation Set	Test Set
1	Human - Machine	Human - Machine	Human - Human
2	Human - Human	Human - Human	Human - Machine
3	Human - Machine	Human - Human	Human - Human
4	Human - Human	Human - Machine	Human - Machine

3.4.4.3 Neural Correlate Feature Comparison

Many studies have used physiological signals to find the primary neural correlates of trust, but each one of them have their own findings. There is an intersection between many of these results, but still there is no generally accepted feature set that has been determined to be the best for classifying trust. Table 5 restates a few of the rows from Table 2 in Section 2.3.1. For each of the findings presented in Table 5, a new dataset was created from the human-human dataset by using the values from the corresponding features. These datasets then went through the same process described in Section 3.4.4.1.

Table 5. The neural correlate sets and their corresponding articles for the datasets used in these tests

Study Name	Year	Findings/features used
Adaptive Probabilistic Classification(Akash, Reid, et al., 2018)	2018	Mean frequency on P4, C4, and P3, peak-to-peak value of C4 and C3, root mean square of Fz, energy of Fz, variation of Fz, correlation of C4 & P4, energy of beta band on P3, CZ, C3, and variation of beta band on P3, CZ, C3
Classification for Sensing Trust (Akash, Hu, et al., 2018)	2018	Mean frequency Fz, C3, and C4, peak-to-peak value on C3, energy of theta band on P3, variance of alpha band on P4, energy of beta band on C4 and P3, mean of beta band on C3, correlation on C3 & C4 and Cz & C4, and the net phasic component from GSR
Real-Time Sensing of Trust (Hu et al., 2016)	2016	High beta band on P4, POz, and C4, Mid beta band on C3, the mean frequency on C3, C4, and P4, the net phasic component and maximum phasic component from GSR, and response time

Some of the features shown in Table 5 were not used in the feature sets for the other experiments in this research. These features include mean frequency, variance, peak-to-peak value, root mean square, energy, the correlation between two nodes, and the different frequency bands' energy and variance. All but the last two features are time-domain rather than frequency-domain features. Letting $k \in (1, n)$, where n is the total number of 2s epochs, each of length N = 512, and x_k represents the kth epoch of channel ch_x. These features were defined as

1. $mean\ frequency\ \bar{f}_k(ch_x), defined\ as\ the\ estimate\ of\ the\ mean\ frequency\ from\ the\ power\ spectrum\ of\ x_k$

2. $variance = \sigma_k^2(ch_x) = \frac{1}{N-1}\sum_{i=1}^{N} | x_{ki} - \mu_k |^2$

3. peak-to-peak = $pp_k(ch_x) = \max_{1 \le i \le N} x_{ki} - \min_{1 \le i \le N} x_{ki}$

4. root mean square = $rms_k(ch_x) = \sqrt{\frac{1}{N}\sum_{i=1}^{N} |x_{ki}|^2}$

5. energy = $E_k(ch_x) = \sum_{i=1}^{N} |x_{ki}|^2$

6. correlation between two channels = $p_k(ch_x, ch_y) = \frac{cov(x_k, y_k)}{\sqrt{var(x_k)var(y_k)}}$.

The expressions *cov(.)* and *var(.)* are the covariance and variance functions respectively. Two of the studies also included a feature found using GSR, but as that data was not available for this research those features were left out.

3.4.5 Performance Analysis

The primary metric used to judge the final performance of the models is the balanced classification accuracy score. Due to the imbalance between the trust and distrust classes, this value gives a more complete representation of each model's performance as compared to the regular classification accuracy. In addition to this value, confusion matrices were created to illustrate the model's performance on the individual classes.

Since the balanced accuracy score and the confusion matrix depend on tuning the thresholds that are used to classify a model's probability output, an Area Under the Receiver Operating Characteristic Curve (AUROC) will also be used to help determine the overall best model for each test. This is found by plotting the Receiver Operating Characteristic (ROC) curve, which is the true positive rate versus the false positive rate at various classification thresholds. An AUROC value of 1.0 means that the model had a perfect classification score while 0.0 means that it got every prediction wrong. And a score of 0.5 means the classifier is operating with the same correctness as a random guess.

Lastly, an analysis of the specific original feature values was performed. A logistic regression model was created that was fit on the entirety of each dataset. The coefficients from the model were taken and plotted using a histogram. The ten features with the greatest absolute value of their coefficients were then found. The values of these ten features from every observation in the starting dataset were graphed on a grouped boxplot, separated by the observations' truth class. This plot allows for better insight into any noticeable traits in the features deemed most influential. Additionally, the observations from each test set that all six models mispredicted were found. These observations' values of each dataset's ten influential features were then plotted on the same boxplot as different colored carrots. These added scatter points allowed for a visual inspection of the values of the most influential features for the most missed observations to see if there was an obvious explanation as to why they may have been inaccurately predicted.

3.5 Summary

This chapter began with a list of the research questions being focused on for this study. It then provided a detailed description of the two sets of data, as well as the experiments done to gather them. The machine learning pipeline, which included the pre-processing steps for the data and the different classifiers built, was then discussed. Finally, an explanation of the different experiments done on the data and the analysis techniques used was given. In the next chapter the results from these experiments are presented and analyzed.

IV. Analysis and Results

4.1 Chapter Overview

This chapter provides an in-depth analysis of the results obtained from the tests outlined in Chapter 3. Section 4.2 discusses the machine learning models' results, which were trained and tested on the same dataset. The following part provides a detailed analysis of the cross-task models' performance. Then, section 4.4 analyzes the results from the models using past research on neural correlates of trust. These sections are regarding machine learning models built to classify EEG as that from a trusting or non-trusting decision. The chapter ends with a summarization of the findings.

4.2 Single Dataset Machine Learning Models

4.2.1 Human - Human Dataset

The number of observations per participant ranged from 145-150, with an average of 148.6. Participants with fewer than 150 observations did not answer in the given amount of time at least once, so those trials were not marked with a decision and therefore not included. A breakdown of the specific class distributions for each participant is located in Appendix 1.

For the single participant classifiers, 20 percent of the observations were set aside for the test set. The remaining approximately 120 observations were split so that four-fold cross-validation could be performed during each model's training process for hyperparameter tuning. Therefore, 80% of each participant's data was split into four different sets. Then there were four repetitions where one of the sets was held out as the

validation set while the model was trained on the other three sets. Before testing, new models were created, provided the best parameters found during the cross-validation process, and then fit with all 80% of the data used during the training process.

4.2.1.1 Full Human Dataset

This section's results are from models that were provided with data derived from all 64 channels used to collect brain activity during the experiment. Each observation began with 320 features since the dataset included the mean power values from each of the five frequency bands. The number of features for each dataset after PCA was applied are shown in Appendix 1. This process successfully reduced the dimensionality for every participants' dataset. The balanced test accuracies for each of the six models on each individual participant's data are displayed in Table 6. The results for the cross-participant models are then shown in Table 7.

Averaged across the 19 participants, the top performing model was the logistic regression classifier with a mean balanced accuracy of 61.50% and a standard deviation of 8.80%. The highest single balanced accuracy of 82.50% was also achieved by this classifier type on Subject 16’s data. All six of the classifiers achieved better than random chance on average across the 19 participants. In total, there were nine models with a balanced accuracy of 70.00% or greater. These are bolded in Table 6. For the cross-participant models, only the logistic regression classifier and the LDA model achieved an average balanced accuracy better than chance. The large number of exactly 50.00% results are telling that many of the models failed to learn a difference in the classes since

this is the same score of a model which predicted the same class for every observation**.**

No test set balanced accuracy over 70.00% was achieved.

Table 6. Test set balanced accuracies achieved on the full human - human dataset. Bolded values are those 70.00% or greater.

Participant	Logistic Regression Balanced Accuracy	LDA Balanced Accuracy	QDA Balanced Accuracy	SVM Balanced Accuracy	RFC Balanced Accuracy	ANN Balanced Accuracy
Subject 01	61.93%	57.95%	64.20%	55.11%	53.41%	53.41%
Subject 02	**70.00%**	67.50%	62.50%	50.00%	50.00%	55.00%
Subject 03	67.05%	**76.14%**	60.80%	**81.82%**	47.73%	60.23%
Subject 04	53.97%	55.56%	52.38%	54.76%	59.52%	54.76%
Subject 06	**73.02%**	57.14%	42.86%	63.49%	35.71%	65.08%
Subject 07	52.84%	38.64%	50.00%	44.89%	60.23%	56.25%
Subject 08	**70.57%**	60.29%	50.00%	66.03%	50.00%	51.20%
Subject 09	56.35%	50.79%	59.52%	58.73%	59.52%	50.00%
Subject 10	60.27%	66.52%	58.48%	52.23%	63.39%	52.23%
Subject 11	63.33%	**70.00%**	66.67%	66.67%	56.67%	**70.00%**
Subject 12	53.97%	53.97%	44.44%	53.17%	47.62%	56.35%
Subject 13	57.39%	61.93%	53.41%	50.00%	50.00%	46.02%
Subject 14	50.00%	41.67%	43.75%	50.00%	50.00%	60.42%
Subject 15	55.00%	60.00%	32.50%	50.00%	62.50%	52.50%
Subject 16	**82.50%**	50.00%	55.00%	60.00%	62.50%	67.50%
Subject 17	62.78%	65.28%	63.33%	53.06%	56.39%	52.22%
Subject 18	57.50%	40.00%	42.50%	32.50%	50.00%	57.50%
Subject 19	**70.00%**	52.50%	45.00%	60.00%	55.00%	50.00%
Subject 20	50.00%	45.83%	43.75%	58.33%	50.00%	52.08%
Average	61.50%	56.41%	52.16%	55.83%	53.69%	55.93%
Standard Deviation	8.80%	10.40%	9.41%	10.09%	6.89%	6.28%

Table 7. Test set cross participant balanced accuracies on the human - human dataset.

Participant	Logistic Regression Balanced Accuracy	LDA Balanced Accuracy	QDA Balanced Accuracy	SVM Balanced Accuracy	RFC Balanced Accuracy	ANN Balanced Accuracy
Cross - Subject 03	50.00%	50.00%	50.00%	50.00%	50.00%	54.20%
Cross - Subject 06	52.59%	54.42%	40.93%	50.00%	50.00%	47.93%
Cross - Subject 09	50.63%	50.32%	51.59%	50.00%	50.00%	54.29%
Cross - Subject 20	50.00%	50.00%	50.00%	50.00%	50.00%	50.00%
Average	51.25%	51.09%	47.53%	50.00%	50.00%	49.76%
Standard Deviation	1.23%	2.16%	4.86%	0.00%	0.00%	3.16%

Another metric used to evaluate the classifiers are ROC curves and their corresponding AUROCs. The values for each model provided with a single participant's data are shown in Table 8. Based on this criterion, the logistic regression models once again performed best with an average AUROC of 0.63 and standard deviation of 0.11. All scores greater than or equal to 0.80 are bolded in the table. In Table 9 are the AUROC results from the cross-participant test sets. The only standout result is the 0.62 achieved by the RFC when testing Subject 20's data.

Table 8. Single participant AUROCs on the human-human dataset. AUROCs greater than or equal to 0.80 are bolded.

Participant	Logistic Regression AUROC	LDA Balanced AUROC	QDA Balanced AUROC	SVM Balanced AUROC	RFC Balanced AUROC	ANN Balanced AUROC
Subject 01	0.51	0.52	0.61	0.57	0.45	0.53
Subject 02	**0.80**	**0.83**	0.66	0.51	0.51	0.55
Subject 03	0.70	0.68	0.63	**0.84**	0.41	0.60
Subject 04	0.58	0.61	0.51	0.30	0.63	0.55
Subject 06	0.75	0.73	0.45	0.69	0.30	0.65
Subject 07	0.46	0.31	0.49	0.31	0.59	0.56
Subject 08	0.73	0.68	0.50	0.72	0.49	0.51
Subject 09	0.50	0.60	0.60	0.43	0.56	0.50
Subject 10	0.58	0.67	0.62	0.34	0.58	0.52
Subject 11	0.66	0.70	0.74	0.72	0.59	0.70
Subject 12	0.56	0.54	0.48	0.50	0.42	0.56
Subject 13	0.61	0.66	0.43	0.39	0.40	0.46
Subject 14	0.51	0.43	0.53	0.74	0.39	0.60
Subject 15	0.57	0.56	0.29	0.50	0.54	0.52
Subject 16	**0.80**	0.73	0.60	0.70	0.68	0.68
Subject 17	0.61	0.69	0.66	0.35	0.41	0.52
Subject 18	0.55	0.51	0.43	0.40	0.53	0.58
Subject 19	**0.84**	0.66	0.61	0.67	0.57	0.50
Subject 20	0.69	0.63	0.38	0.34	0.34	0.52
Average	0.63	0.62	0.54	0.53	0.49	0.56
Standard Deviation	0.11	0.12	0.11	0.17	0.10	0.06

Table 9. Cross-participant AUROCs on the human-human dataset

Participant	Logistic Regression AUROC	LDA AUROC	QDA AUROC	SVM AUROC	RFC AUROC	ANN AUROC
Cross - Subject 03	0.40	0.39	0.40	0.51	0.45	0.54
Cross - Subject 06	0.56	0.57	0.37	0.48	0.50	0.48
Cross - Subject 09	0.53	0.53	0.54	0.52	0.54	0.54
Cross - Subject 20	0.54	0.55	0.47	0.55	0.62	0.50
Average	0.52	0.52	0.47	0.51	0.49	0.50
Standard Deviation	0.07	0.08	0.08	0.03	0.07	0.03

To try to understand where the models did not perform well, confusion matrices were created for each model. These figures provide detail about what type of errors occurred. For each of the six classifiers, the confusion matrices for all of the single-participant data predictions are shown in Figure 18. In the pictures, "1" represents the trust observations, the positive class, and 0 represents the distrust observations, the negative class. The prediction count and prediction percentage per class is shown for each cell in the matrix. The false-positive rate (FPR) and false-negative rate (FNR) show the likelihood of an error given an observation's truth value. For both of these metrics, 0% is perfect, and 100% is wholly misclassified. These values were calculated for all of the within-participant predictions on each of the six models and are displayed in Table 7.

Table 10. The FPRs and FNRs from the combined predictions of all the participants

Within Participant	Logistic Regression	LDA	QDA	SVM	RFC	ANN
FPR	46.93%	67.60%	63.13%	55.87%	78.21%	39.66%
FNR	31.28%	17.95%	33.33%	30.00%	16.67%	47.18%

Based on the values in Table 10 and the within-participant confusion matrices, logistic regression and the ANN are the best two models. Both the LDA and RFC performed very well in terms of the FNR, but it was at the cost of having the two highest FPRs. The confusion matrices from each model type for the cross-participant tests are shown in Figure 19. The FPR and FNRs derived from these confusion matrices are shown in Table 11. The confusion matrices show that the SVM and RFC models predicted every observation to be in the trust class and both the LDA and QDA did nearly the same. Given this, it is apparent that these models were unable to learn anything to differentiate between the two classes on the cross-participant data. Logistic regression and the ANN both had approximately a 70% FPR and a 30% FNR which are not great but are still vastly better than the other models.

Table 11. FPRs and FNRs from the predictions on the cross-participant models

Cross-participant	Logistic Regression	LDA	QDA	SVM	RFC	ANN
FPR	70.25%	98.10%	94.94%	100.00%	100.00%	70.25%
FNR	31.82%	1.14%	10.00%	0.00%	0.00%	30.23%

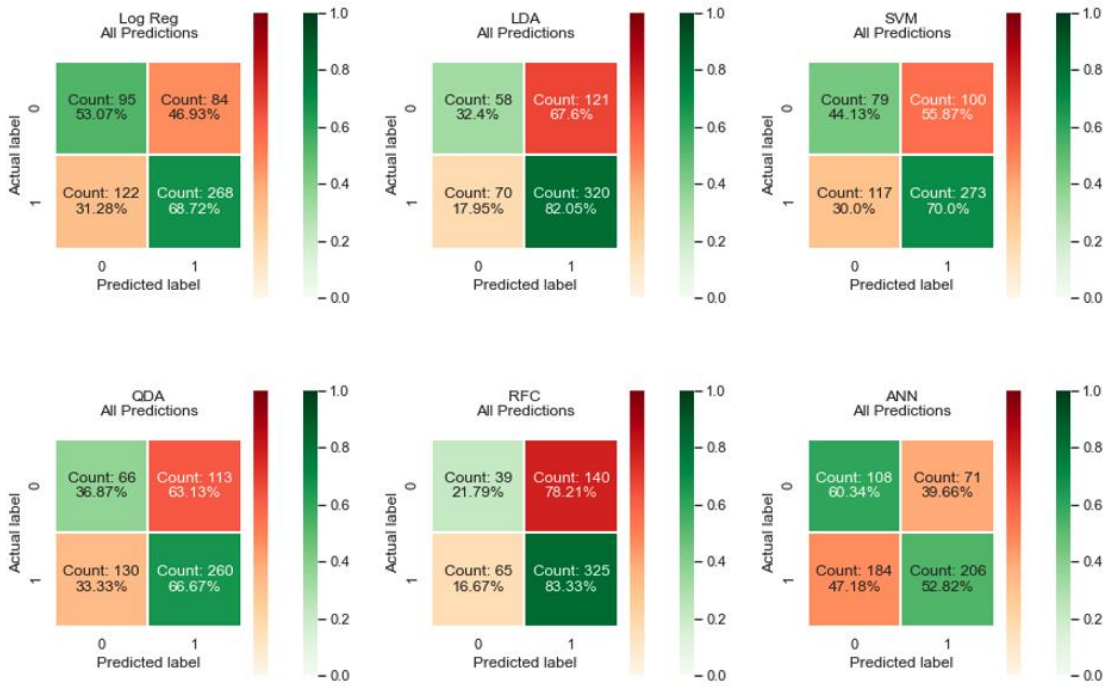

Figure 18. Confusion matrices for the combined predictions on the single participant human-human models

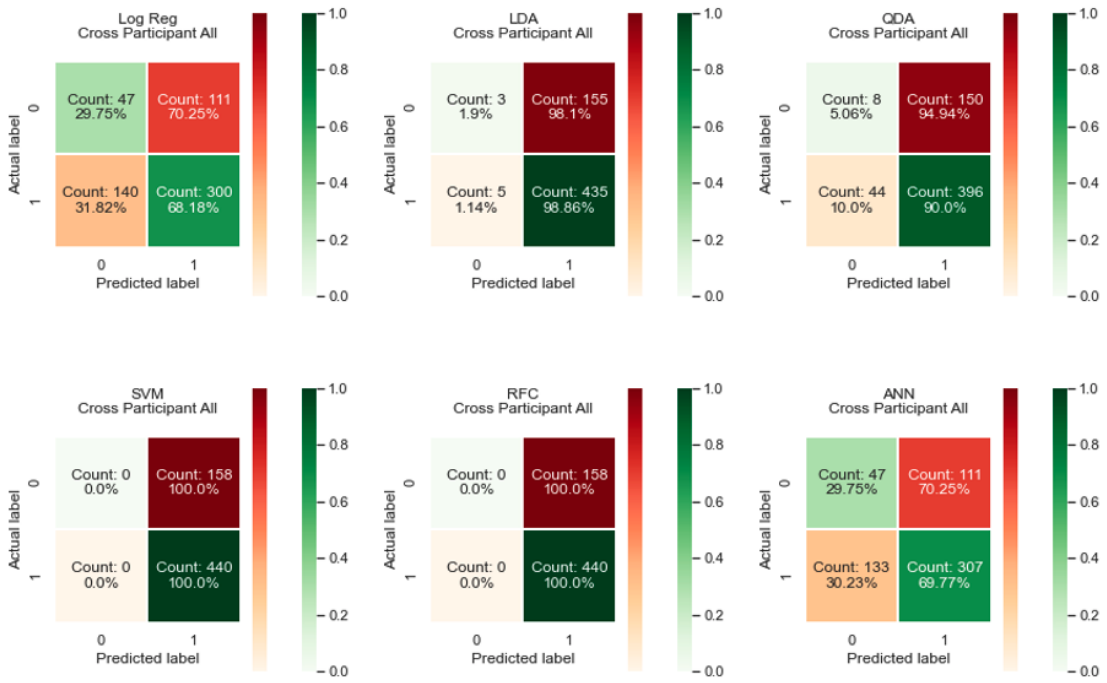

Figure 19. Confusion matrices for the cross-participant human-human models

To see if any noticeable traits could be found to explain the errors, observations from a participant which were incorrectly predicted by all six models were looked at further. The number of each of these per subject can be found in Appendix 2. The detailed process described in Section 3.4.5 was used for this part. Figure 20 shows a histogram with the features in the dataset and their coefficients for the logistic regression model fitted with the full dataset. It is clear that very few features were much more impactful toward the predictions than the others with only a few coefficients being greater than 0.05 and the majority of features in the first bar in the graph.

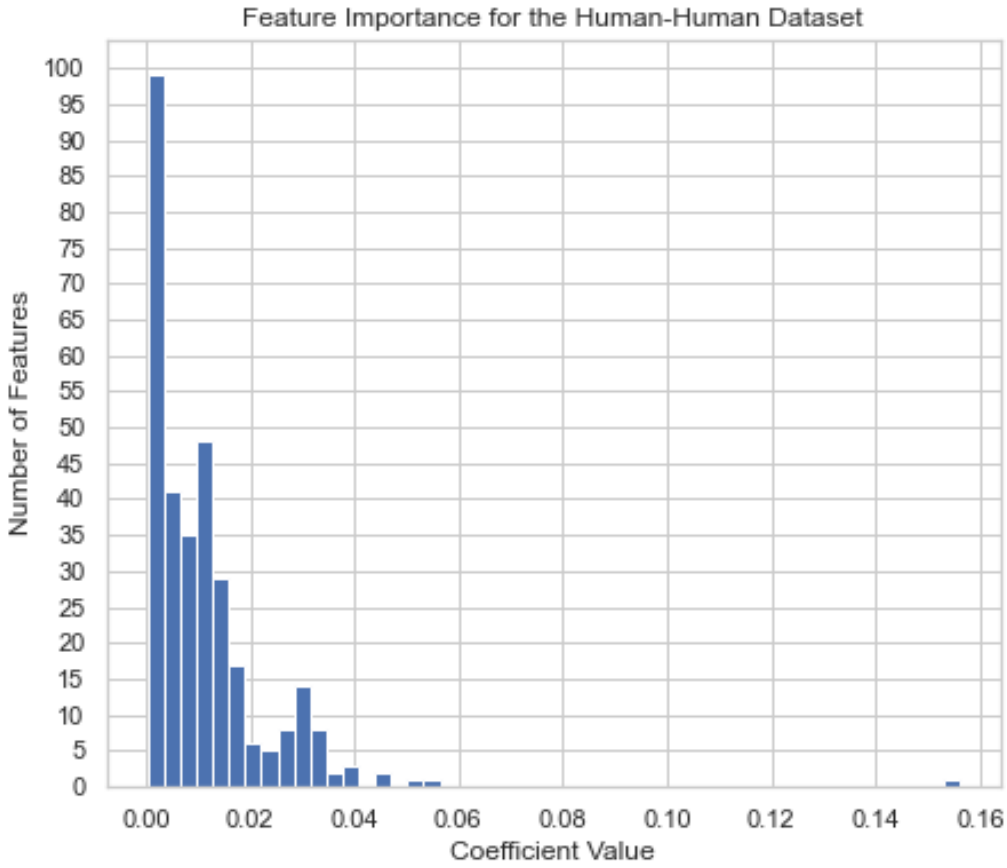

Figure 20. Histogram of the human-human dataset features and their logistic regression coefficients

To look closer at the most impactful features for the model's predictions, the 10 with the greatest absolute value were separated. The values for each feature were graphed on grouped boxplots by those from trust observations vs. distrust observations. The features are graphed in descending order of influence on predictions of the models. The feature values from the observations which were incorrectly predicted by all six models

were then added to the graphs. The FPs are the red carrots, and the FNs are the green carrots. This was done to see if there is anything obvious to distinguish the missed observations from the rest of the values. As an example, the graph created for Subject 11's observations are shown in Figure 21. Subject 11 was chosen as it is one of the only subjects with at least one positive and negative observation missed on every model. The most telling information from the group of boxplots are how similar the average and quartile values are between the two classes for every influential feature. For no feature on any participant is the median value from one class outside of the first and third quartile in the opposite class. Without many apparent differences between features, even the ones deemed most impactful, it makes since that the classifiers did not perform well.

In respect to the FP and FN observations, there were some examples, like the downward pointing carrot over F8_delta in Figure 21, where the feature value is an outlier of its true class and may explain why it was guessed incorrectly but the majority of these values were not out of the ordinary for the subject. This info shows that the values for the most influential features of each participants data were very similar between the two trust classes making it seemingly very difficult to separate the classes in order to make predictions from unseen observations.

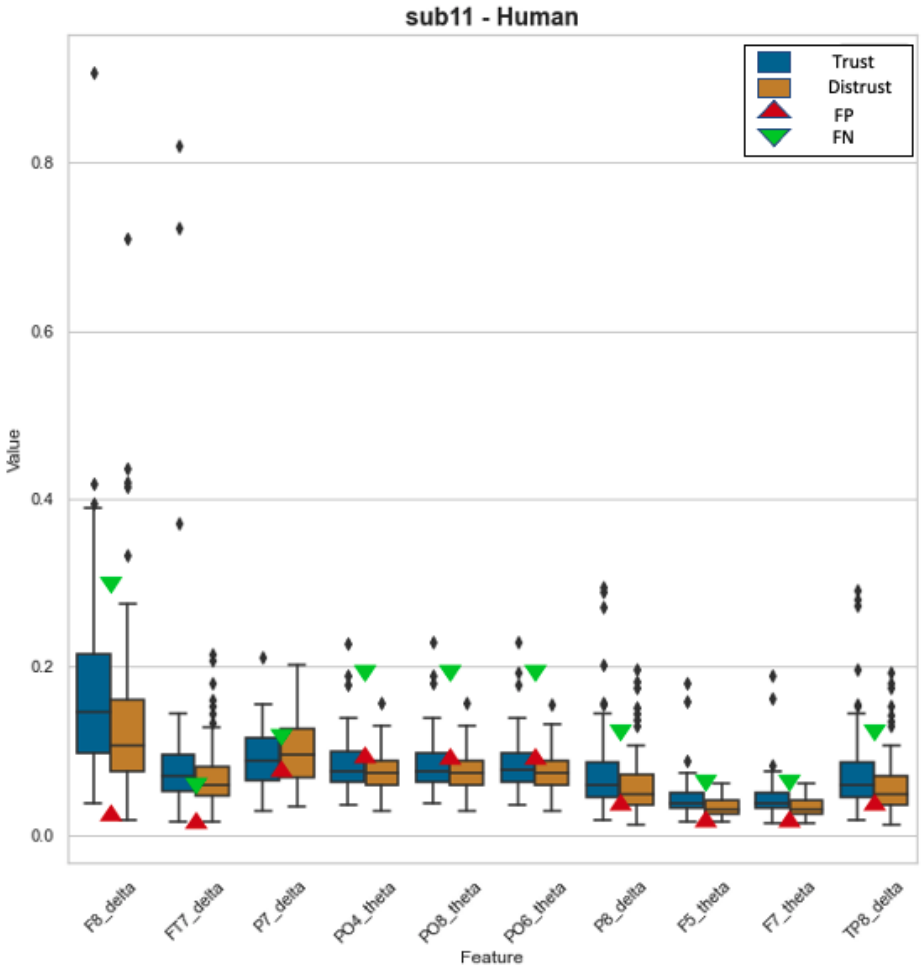

Figure 21. Observation values from the top 10 most influential features for prediction for Subject 11

4.2.1.2 Human-Human 9-Channel Dataset

This section includes results from the models which were trained on the datasets derived from data collected from 9 channels in the human - human experiment. The five values for each electrode resulted in 45 features per observation for this dataset. The

number of features for each dataset after PCA was applied is shown in Appendix 1. The balanced accuracies found from the six models are reported in Table 12.

Table 12. Test set balanced accuracies achieved on the cross-participant 9 channel models. Scores greater than 70.00 are bolded.

Participant	Logistic Regression Balanced Accuracy	LDA Balanced Accuracy	QDA Balanced Accuracy	SVM Balanced Accuracy	RFC Balanced Accuracy	ANN Balanced Accuracy
Subject 01	60.80%	55.11%	63.64%	50.00%	50.00%	50.00%
Subject 02	65.00%	60.00%	52.50%	62.50%	52.50%	62.50%
Subject 03	**82.39%**	50.00%	53.41%	72.16%	**72.16%**	60.23%
Subject 04	**71.43%**	**73.02%**	**70.63%**	**81.75%**	**76.19%**	**71.43%**
Subject 06	**71.43%**	65.08%	57.14%	55.56%	55.56%	57.14%
Subject 07	48.86%	45.45%	51.70%	50.00%	50.00%	51.14%
Subject 08	69.86%	**75.84%**	**75.12%**	67.22%	67.46%	**70.57%**
Subject 09	65.87%	53.17%	56.35%	61.90%	38.10%	65.87%
Subject 10	52.68%	52.68%	62.95%	52.68%	63.84%	59.38%
Subject 11	60.00%	60.00%	56.67%	63.33%	63.33%	63.33%
Subject 12	61.90%	58.73%	65.08%	55.56%	50.00%	55.56%
Subject 13	64.77%	67.61%	53.41%	52.84%	52.84%	56.25%
Subject 14	41.67%	47.92%	45.83%	58.33%	58.33%	43.75%
Subject 15	55.00%	40.00%	50.00%	65.00%	52.50%	47.50%
Subject 16	**87.50%**	**85.00%**	50.00%	**82.50%**	**80.00%**	**75.00%**
Subject 17	57.22%	64.72%	53.61%	59.72%	59.72%	57.22%
Subject 18	65.00%	62.50%	55.00%	50.00%	50.00%	52.50%
Subject 19	65.00%	52.50%	45.00%	50.00%	52.50%	57.50%
Subject 20	39.58%	41.67%	50.00%	47.92%	50.00%	60.42%
Average	62.42%	58.47%	56.21%	59.95%	57.63%	58.80%
Standard Deviation	11.71%	11.45%	7.79%	10.05%	10.28%	7.96%

Table 13. Test set cross-participant balanced accuracies on the human-human 9-channel dataset.

Participant	Logistic Regression Balanced Accuracy	LDA Balanced Accuracy	QDA Balanced Accuracy	SVM Balanced Accuracy	RFC Balanced Accuracy	ANN Balanced Accuracy
Cross - Subject 03	50.19%	50.00%	51.42%	50.00%	52.85%	48.89%
Cross - Subject 06	50.00%	51.35%	47.83%	50.00%	50.00%	46.77%
Cross - Subject 09	46.98%	51.27%	49.68%	50.63%	45.87%	53.49%
Cross - Subject 20	50.45%	53.18%	54.43%	50.00%	58.52%	54.20%
Average	49.19%	51.28%	52.88%	50.16%	51.46%	51.99%
Standard Deviation	1.63%	1.31%	2.81%	0.32%	5.31%	3.59%

As with the full human-human dataset, the logistic regression model had the highest average balanced accuracy across all participants as well as the single highest score of 87.50%. This was achieved on participant 16's test set which is the same participant and model type from the highest score on the full human-human dataset. As a whole, there were 16 models with balanced accuracies greater than 70.00% which are bolded in Table 12. Based on these metrics, the logistic regression model is the best classifier for single participants on this dataset.

Surprisingly given this, the logistic regression model is the only classifier to have an average balanced accuracy under 50.00% for the cross-participant models. There were some notably high single-participant performances though, especially on Subject 20's data. The RFC given this data scored 58.52% and was followed by the QDA then the ANN with 54.43% and 54.20% respectively. The ANN tested on Subject 09's data also performed noticeably above random chance with a score of 53.49%. Even though higher

accuracies were achieved on the cross-participant models with this dataset as compared to the full human-human dataset, there were still no balanced scores greater than 70.00%.

The AUROC values found for these single participants are displayed in Table 14. The LDA had the highest average score of 0.65 and a standard deviation of 0.12 with logistic regression right behind it with a mean of 0.64 and standard deviation of 0.12. The three highest scores were all from models on Subject 16's data. Logistic regression, LDA, and SVM achieved values of 0.88, 0.90, and 0.91 respectively. Table 15 shows the results found on the cross-participant models. The logistic regression and LDA classifiers had average AUROCs below 0.50, but the average AUROC from the RFC was 0.57. The majority of the AUROC values being near 0.50 means that there is a large overlap in the probabilities being predicted by the models on the two different truth classes. This helps explain the poor performance seen by the majority of the models on this dataset and indicates that there is likely too much noise in the data.

Table 14. Single participant AUROCs on the human-human 9-channel dataset.

AUROCs greater than or equal to 0.80 are bolded.

Participant	Logistic Regression AUROC	LDA Balanced AUROC	QDA Balanced AUROC	SVM Balanced AUROC	RFC Balanced AUROC	ANN Balanced AUROC
Subject 01	0.64	0.62	0.69	0.46	0.74	0.50
Subject 02	0.70	0.69	0.55	0.73	0.60	0.45
Subject 03	0.77	0.77	0.60	0.73	0.48	0.60
Subject 04	0.78	0.76	0.72	0.78	0.75	0.70
Subject 06	0.72	0.76	0.68	0.69	0.66	0.52
Subject 07	0.46	0.44	0.41	0.55	0.42	0.48
Subject 08	0.74	0.78	0.71	0.66	0.78	0.52
Subject 09	0.63	0.70	0.69	0.24	0.58	0.50
Subject 10	0.55	0.54	0.63	0.57	0.60	0.58
Subject 11	0.67	0.63	0.66	0.68	0.70	0.57
Subject 12	0.60	0.62	0.61	0.54	0.40	0.56
Subject 13	0.66	0.74	0.51	0.57	0.52	0.56
Subject 14	0.47	0.44	0.40	0.57	0.50	0.38
Subject 15	0.51	0.48	0.54	0.61	0.48	0.48
Subject 16	**0.88**	**0.90**	0.63	**0.91**	0.65	0.75
Subject 17	0.68	0.67	0.62	0.64	0.60	0.57
Subject 18	0.57	0.59	0.54	0.48	0.62	0.50
Subject 19	0.71	0.71	0.50	0.45	0.44	0.58
Subject 20	0.41	0.59	0.46	0.72	0.55	0.38
Average	0.64	0.65	0.59	0.61	0.58	0.53
Standard Deviation	0.12	0.12	0.09	0.14	0.11	0.09

Table 15. Cross-participant AUROCs on the human-human 9-channel dataset

Participant	Logistic Regression AUROC	LDA AUROC	QDA AUROC	SVM AUROC	RFC AUROC	ANN AUROC
Cross - Subject 03	0.45	0.48	0.53	0.54	0.46	0.49
Cross - Subject 06	0.56	0.55	0.41	**0.62**	0.46	0.47
Cross - Subject 09	0.49	0.49	0.47	0.48	0.46	0.53
Cross - Subject 20	0.44	0.45	0.54	**0.61**	0.55	0.54
Average	0.46	0.47	0.51	0.57	0.50	0.52
Standard Deviation	0.05	0.04	0.06	0.07	0.04	0.04

The confusion matrices for the single participant models on this dataset are displayed in Figure 22. The FPRs and FNRs which can be derived from the confusion matrixes are listed in Table 16. The logistic regression model was the only one with a FPR under 50.00% but it also did have the highest FNR at 31.02%. This shows that it sacrificed predicting the positive class accurately to have a better accuracy on the negative class. Even though the ANN and SVM's FPRs are almost 60.00%, their confusion matrices show that they likely predict similarly to logistic regression. The LDAs, QDAs, and RFCs, however, predicted almost every observation to be in the trust class. This is surprising, especially for the RFC, since it is provided with the class weight breakdown for each dataset.

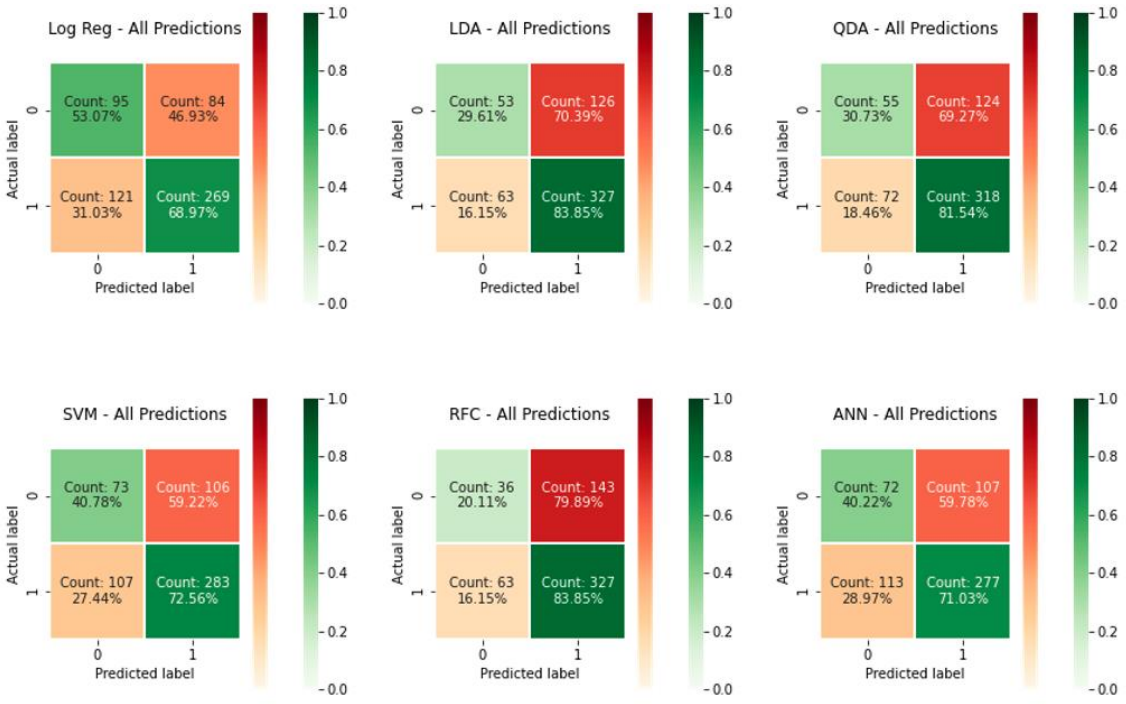

Figure 22. Confusion matrices from the single participant human-human 9-channel models

Table 16. FPRs and FNRs from the predictions on the within-participant human-human 9-channel models

Within-participant	Logistic Regression	LDA	QDA	SVM	RFC	ANN
FPR	46.93%	70.39%	69.27%	59.22%	79.89%	59.78%
FNR	31.03%	16.15%	18.46%	27.44%	16.15%	28.97%

Figure 23 displays the confusion matrices for the cross-participant models and Table 17 has the corresponding FPRs and FNRs for each model. Logistic regression and QDA were the only two models which performed differently between the single-

participant and cross-participant models. The logistic regression model favoring the negative class in the cross-participant model could be due to the difference in class weights in the training sets as opposed to the test sets. The other four models all had similar, just more exaggerated, FPRs and FNRs in both cases. Between the balanced accuracies, AUROCs, and confusion matrices there is not a definitive best performing model based on every metric for the human-human 9-channel dataset.

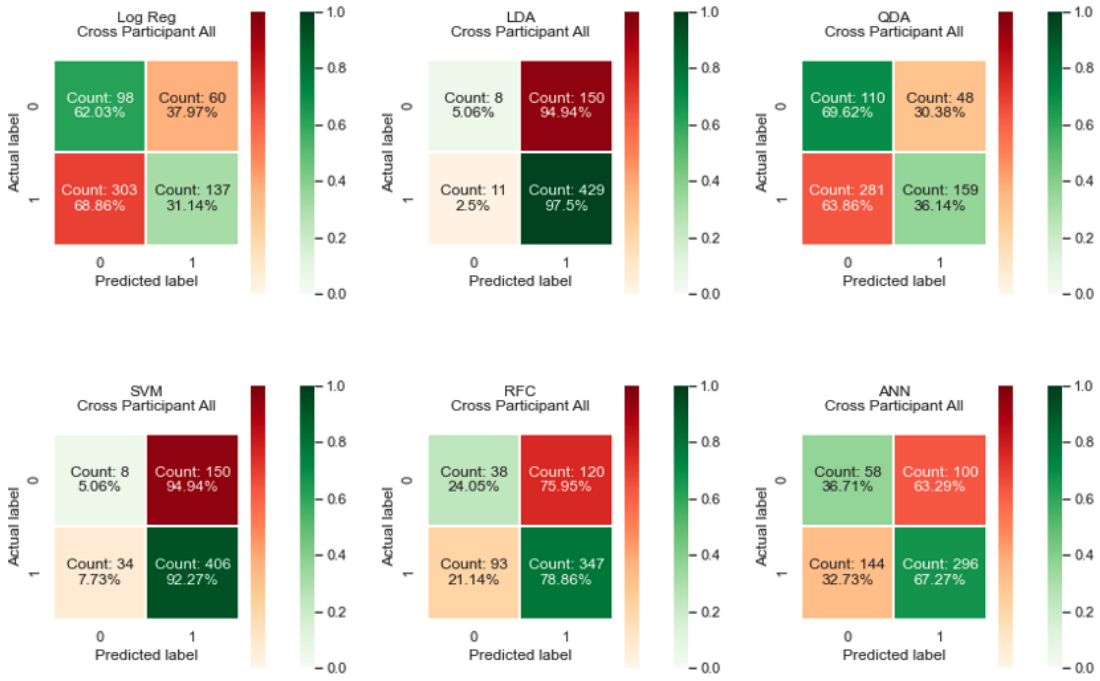

Figure 23. Confusion matrices from the cross-participant human-human 9-channel models

Table 17. FPRs and FNRs from the prediction on the cross-participant human-human 9-channel models

Cross-participant	Logistic Regression	LDA	QDA	SVM	RFC	ANN
FPR	37.97%	94.94%	30.38%	94.94%	75.95%	63.29%
FNR	68.86%	2.50%	63.86%	7.73%	21.14%	32.73%

As with the full human-human dataset, the individual features in each dataset were analyzed. Since multiple models given Subject 16's data performed extremely well on all the metrics analyzed, the feature set histogram and grouped boxplot from this dataset are displayed in Figures 24 and 25. The histogram shows that nearly 20% of the absolute values of the feature coefficients are greater than 0.2. Provided that almost half of the features are in the first bar of the histogram, this means that if a model could learn the difference between the classes based on just a few of these features than it would likely have a high classification accuracy.

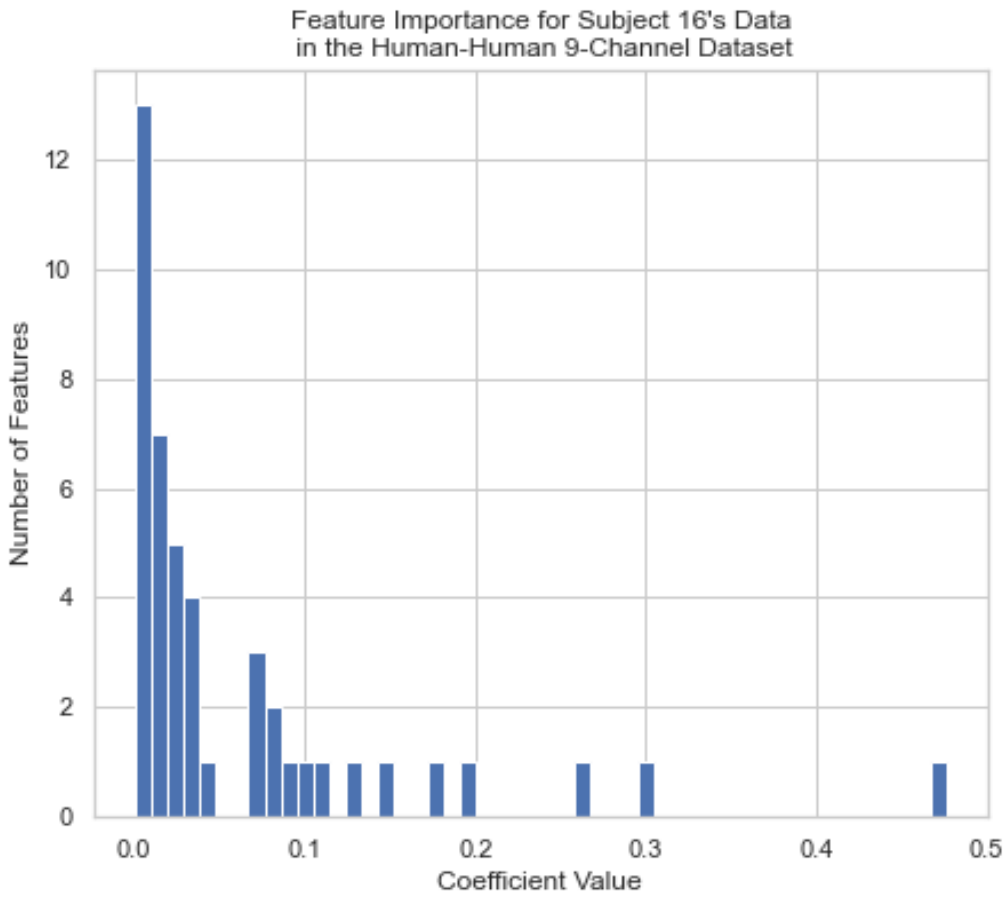

Figure 24. Histogram of the coefficients for the features in Subject 16's data in the human-human 9-channel dataset

There are a few noticeable things in the grouped boxplot which may further explain the high performance of the models on Subject 16's data. The median of the positive class on F4_delta and P4_delta is equal to the third quartile boundary of their respective negative class. These are two of the very few times where a difference of this size between the two classes occurs across all of the datasets and their ten most

influential features. The number of FPs and FNs missed on every model per subject can be found in Appendix 3. With respect to the one of these noted in Figure 25, its F4_delta value was an outlier for the negative class, which it belonged, but was within one of the whiskers of the positive class box. For all other nine features, the value of this observation does not standout with respect to either of the classes. This shows that due to how similar the majority of the feature values are between the two classes the machine learning models were likely very sensitive to any outlier values.

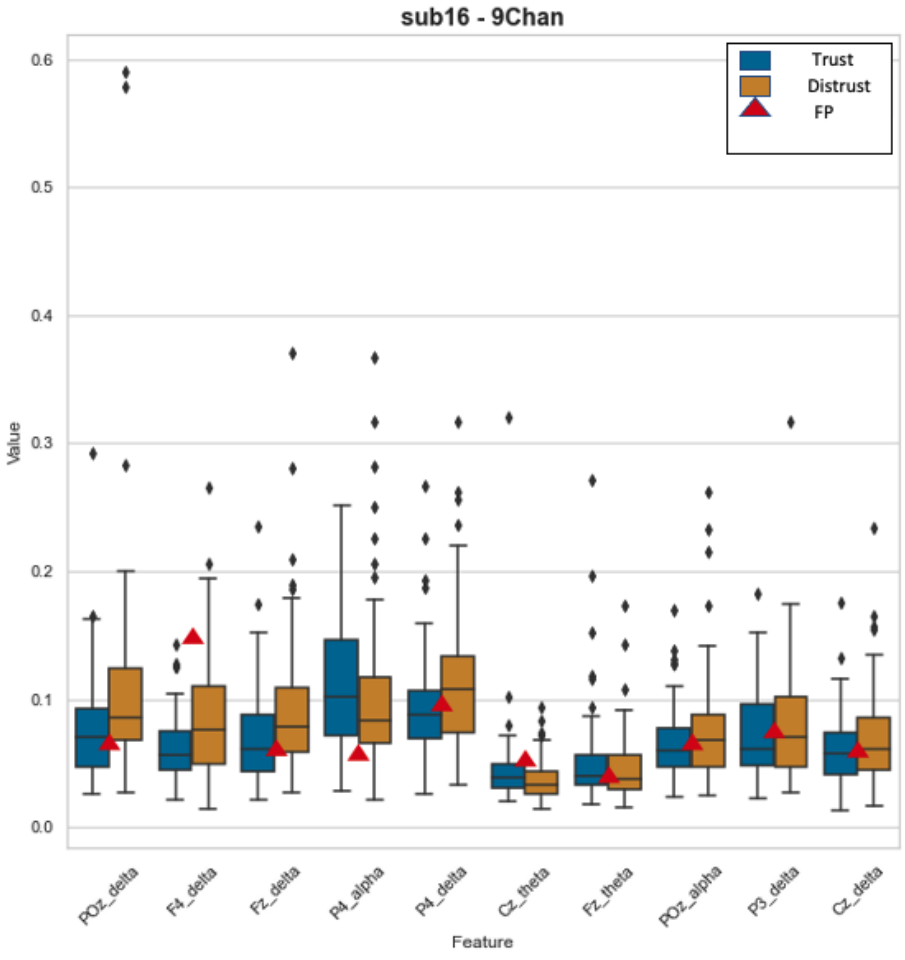

Figure 25. Observation values from the top 10 most influential features for prediction for Subject 16's data from the human-human 9-channel dataset

4.2.2 Human - Machine Dataset

There were 100 observations for every participant in this dataset, with an average percent of trust observations of 70.45%. The individual class breakdowns can be found in Appendix 4. To start, each observation had 45 features as with the human-human 9-channel dataset. The number of features in each dataset post application of PCA is shown in Appendix 4.

The train, validate, and test split procedure for the human-machine data was the same as what was done with the human-human dataset. Four-fold cross-validation was also used, meaning that only 20 observations were part of each set. The balanced test set accuracies of the six models are displayed in Table 18. The highest average balanced score, of 57.61% with a standard deviation of 10.39%, was again achieved by logistic regression. On this dataset, though, the next best averages were much higher with the QDA, SVM, and RFC all over 57% as well. Also different from the results on the two human-human datasets, the highest single score of 86.67% was from an LDA model. This accuracy occurred with AS23's data as input. All scores 70.00% or greater are bolded in the table.

Table 18. Test set balanced accuracies of the models on the single participant human-machine dataset. Accuracies 70.00% or greater are bolded.

Participant	Logistic Regression Balanced Accuracy	LDA Balanced Accuracy	QDA Balanced Accuracy	SVM Balanced Accuracy	RFC Balanced Accuracy	ANN Balanced Accuracy
AS05	**72.62%**	55.95%	**72.62%**	**70.24%**	66.67%	59.52%
AS07	36.90%	57.14%	58.33%	54.76%	47.62%	50.00%
AS08	33.33%	29.76%	34.52%	46.43%	63.10%	48.81%
AS10	54.17%	50.00%	58.33%	54.17%	66.67%	**70.83%**
AS12	61.90%	55.95%	57.14%	55.95%	67.86%	63.10%
AS13	61.90%	42.86%	40.48%	50.00%	41.67%	48.81%
AS14	56.25%	60.42%	64.58%	**75.00%**	58.33%	43.75%
AS15	**75.00%**	46.88%	46.88%	53.13%	65.63%	43.75%
AS16	56.67%	66.67%	43.33%	50.00%	63.33%	53.33%
AS17	51.19%	54.76%	39.29%	58.33%	54.76%	54.76%
AS20	53.33%	56.67%	**70.00%**	63.33%	63.33%	50.00%
AS21	**70.33%**	67.03%	**74.73%**	66.48%	57.14%	59.89%
AS23	**76.67%**	**86.67%**	**80.00%**	50.00%	66.67%	43.33%
AS24	50.00%	45.24%	48.81%	46.43%	50.00%	54.76%
AS27	60.00%	56.67%	66.67%	**76.67%**	**70.00%**	50.00%
AS28	56.67%	53.33%	53.33%	60.00%	60.00%	53.33%
AS30	50.00%	40.00%	46.67%	50.00%	53.33%	46.67%
AS33	60.00%	66.67%	**76.67%**	**70.00%**	50.00%	50.00%
AS35	54.76%	66.67%	55.95%	34.52%	54.76%	35.71%
AS36	53.03%	49.49%	53.03%	48.48%	60.61%	53.03%
BS01	59.38%	46.88%	50.00%	43.75%	50.00%	28.13%
BS02	53.33%	50.00%	60.00%	60.00%	50.00%	60.00%
BS04	51.52%	48.48%	42.93%	50.00%	41.92%	41.92%
BS05	60.00%	56.67%	56.67%	**73.33%**	60.00%	66.67%
BS06	62.50%	62.50%	40.63%	46.88%	62.50%	43.75%
BS07	50.00%	40.48%	48.81%	53.57%	54.76%	46.43%
BS09	52.75%	49.45%	60.44%	50.00%	52.75%	67.58%
BS10	67.86%	59.52%	58.33%	67.86%	63.10%	54.76%
CS01	41.67%	45.83%	66.67%	50.00%	62.50%	45.83%
CS03	**73.33%**	**80.00%**	**73.33%**	**80.00%**	50.00%	66.67%
CS04	68.75%	**71.88%**	**71.88%**	68.75%	50.00%	59.38%
Average	57.61%	55.50%	57.13%	57.36%	57.39%	52.08%
Standard Deviation	10.39%	11.93%	12.34%	11.13%	7.69%	9.60%

The balanced test scores for the cross-participant models are displayed in Table 19. Every one of the six models scored better than random chance, which did not occur with either of the human-human datasets. The greater number of participants used in these models' training likely helped generalize the data, leading to the better results. The highest average balanced accuracy of 56.02% was achieved with logistic regression. Also, for the first time, a cross-participant model eclipsed a balanced accuracy of 60.00%. This achievement occurred when AS14's data and CS03's data were the test input for the ANN. These performances are notable, but they still did not accomplish the research objective of having a model achieve a balanced accuracy greater than 70.00%. One other important thing to note from these results is that the RFC scored exactly 50.00% on all but one of the participants' data. Once again, the RFCs demonstrate that they seemingly are unable to learn any notable differences in the two trust classes with cross-participant data.

Table 19. Test set cross-participant balanced accuracies on the human-machine dataset.

Participant	Logistic Regression Balanced Accuracy	LDA Balanced Accuracy	QDA Balanced Accuracy	SVM Balanced Accuracy	RFC Balanced Accuracy	ANN Balanced Accuracy
AS05	54.55%	50.27%	52.87%	50.00%	50.00%	55.89%
AS14	50.00%	58.07%	52.96%	60.05%	50.00%	63.33%
AS16	51.27%	52.86%	59.28%	51.27%	50.00%	48.28%
AS20	57.85%	49.22%	46.10%	50.00%	50.00%	46.67%
AS23	57.57%	49.34%	55.15%	54.39%	50.00%	54.28%
BS06	51.32%	50.00%	49.56%	56.47%	50.00%	46.82%
CS03	50.00%	50.83%	54.87%	50.00%	53.13%	64.43%
Average	56.02%	51.51%	52.97%	53.17%	50.45%	54.24%
Standard Deviation	4.54%	3.14%	4.22%	3.95%	1.18%	7.49%

Table 20 has the AUROCs from the single participant models. The highest average score was 0.59, with a standard deviation of 0.14. This value was from the RFC models, which also had a high average balanced accuracy on this dataset. Even though the RFC was consistently one of the worst-performing classifiers on the two human-human datasets, it, along with logistic regression, outperformed the other four models on both metrics. All the values of 0.80 or greater are bolded. Five out of the ten occurrences of this were from models using data from CS03.

Table 20. Single participant AUROCs on the human-machine dataset. Scores 0.80 or greater are bolded

Participant	Logistic Regression AUROC	LDA Balanced AUROC	QDA Balanced AUROC	SVM Balanced AUROC	RFC Balanced AUROC	ANN Balanced AUROC
AS05	0.73	0.62	0.69	0.75	0.67	0.60
AS07	0.36	0.52	0.56	0.58	0.52	0.50
AS08	0.21	0.24	0.29	0.42	0.42	0.49
AS10	0.63	0.74	0.58	0.63	0.70	0.71
AS12	0.57	0.64	0.57	0.70	0.63	0.63
AS13	0.64	0.45	0.40	0.45	0.25	0.49
AS14	0.58	0.61	0.63	0.75	0.59	0.44
AS15	0.78	0.56	0.50	0.39	0.75	0.44
AS16	0.59	0.60	0.39	0.52	0.62	0.53
AS17	0.42	0.40	0.46	0.46	0.40	0.55
AS20	0.52	0.45	0.69	0.49	0.74	0.50
AS21	0.59	0.54	0.68	0.73	0.68	0.60
AS23	**0.87**	**0.91**	**0.83**	0.31	0.75	0.43
AS24	0.55	0.56	0.44	0.44	0.48	0.55
AS27	0.49	0.69	0.64	0.69	**0.87**	0.50
AS28	0.64	0.55	0.40	0.32	0.57	0.53
AS30	0.57	0.55	0.52	0.52	0.64	0.47
AS33	0.60	0.72	0.75	0.65	0.60	0.50
AS35	0.36	0.37	0.33	0.33	0.44	0.36
AS36	0.54	0.49	0.49	0.47	0.63	0.53
BS01	0.50	0.56	0.63	0.47	0.66	0.28
BS02	0.59	0.60	0.41	0.60	0.37	0.60
BS04	0.44	0.48	0.40	0.42	0.45	0.42
BS05	0.69	0.61	0.67	0.89	0.60	0.67
BS06	0.69	0.73	0.36	0.61	0.48	0.44
BS07	0.43	0.42	0.56	0.50	0.70	0.46
BS09	0.54	0.49	0.56	0.47	0.40	0.68
BS10	0.60	0.60	0.64	0.62	0.52	0.55
CS01	0.41	0.36	0.61	0.38	0.61	0.46
CS03	**0.80**	**0.88**	**0.83**	**0.84**	**0.81**	0.67
CS04	0.75	0.69	**0.84**	0.56	0.69	0.59
Average	0.57	0.57	0.56	0.55	0.59	0.52
Standard Deviation	0.14	0.14	0.15	0.15	0.14	0.10

The AUROC values for the cross-participant models are displayed in Table 21. The logistic regression and LDA models with AS14's data achieved the highest AUROC scores of 0.74. This was the highest cross-participant AUROC on this dataset as well as on either of the previous cross-participant models. Once again, logistic regression outperforms the other models with the greatest average AUROC of 0.57. The better cross-participant results on this dataset compared to the previous two datasets are probably due to the increased number of participants' data included in the training.

Table 21. Cross-participant AUROCs on the human-machine dataset

Participant	Logistic Regression AUROC	LDA Balanced AUROC	QDA Balanced AUROC	SVM Balanced AUROC	RFC Balanced AUROC	ANN Balanced AUROC
AS05	0.49	0.49	0.53	0.53	0.52	0.56
AS14	0.74	0.74	0.55	0.61	0.54	**0**.63
AS16	0.68	0.68	0.60	0.52	0.58	0.48
AS20	0.50	0.50	0.49	0.49	0.41	0.47
AS23	0.54	0.54	0.45	0.41	0.50	0.54
BS06	0.55	0.56	0.49	0.55	0.57	0.47
CS03	0.46	0.45	0.49	0.46	0.53	0.64
Average	0.57	0.56	0.51	0.51	0.52	0.54
Standard Deviation	0.10	0.10	0.05	0.07	0.06	0.07

Given that this dataset was more imbalanced than the human-human dataset, it would make sense that the models tended to overpredict the trust class as most of the previous models did. To see if this is true, the confusion matrices for each model type were created and are displayed in Figure 27. The FPRs and FNRs for the models are then shown in Table 22. As hypothesized, these values are in line with those obtained on the

previous two datasets. Logistic regression and the ANN once again have the most balanced predictions while the other four models predict the positive class most of the time. The confusion matrices for the cross-participant models are displayed in Figure 28, and their FPRs and FNRs are in Table 23. The human-machine dataset's more significant class imbalance looks to have affected the cross-participant models more than it did for the single-participant classifiers. The LDA and RFC models predicted ten and two negative class observations, respectively, when there were actually 191 distrust observations in the combined test sets. The LDA's prediction breakdown is surprising to see as it had the second-highest average AUROC out of the six cross-participant models.

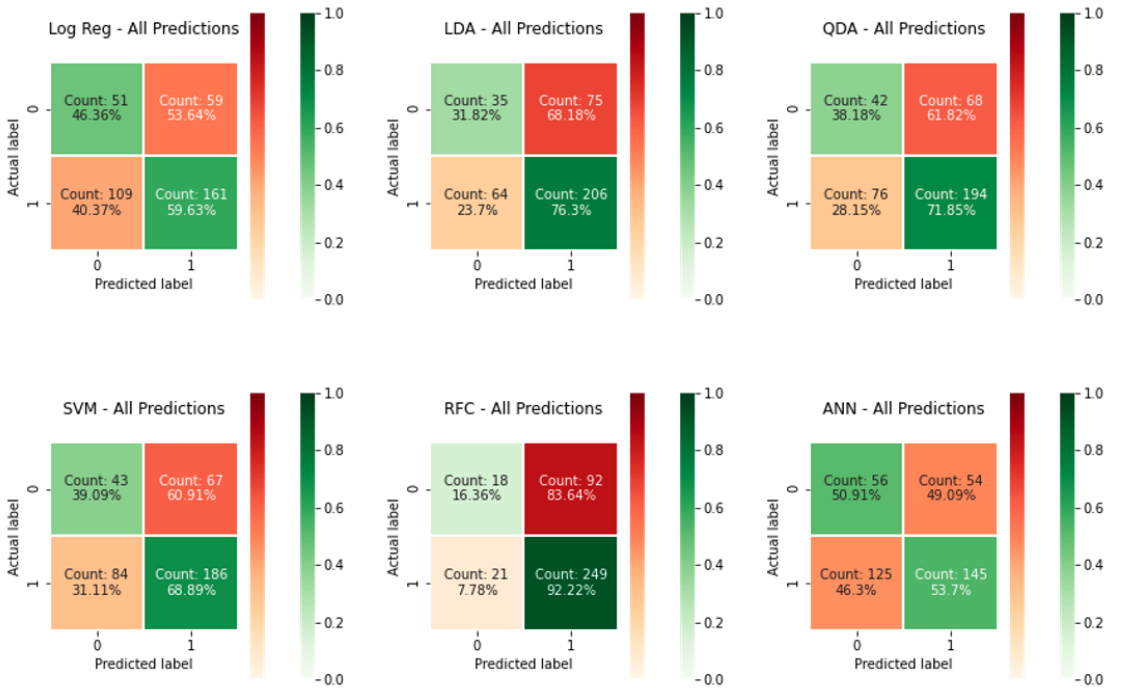

Figure 26. Confusion matrices from the single participant human-machine models

Table 22. FPRs and FNRs for the within-participant human-machine models

Within-participant	Logistic Regression	LDA	QDA	SVM	RFC	ANN
FPR	37.97%	94.94%	30.38%	94.94%	75.95%	63.29%
FNR	68.86%	2.50%	63.86%	7.73%	21.14%	32.73%

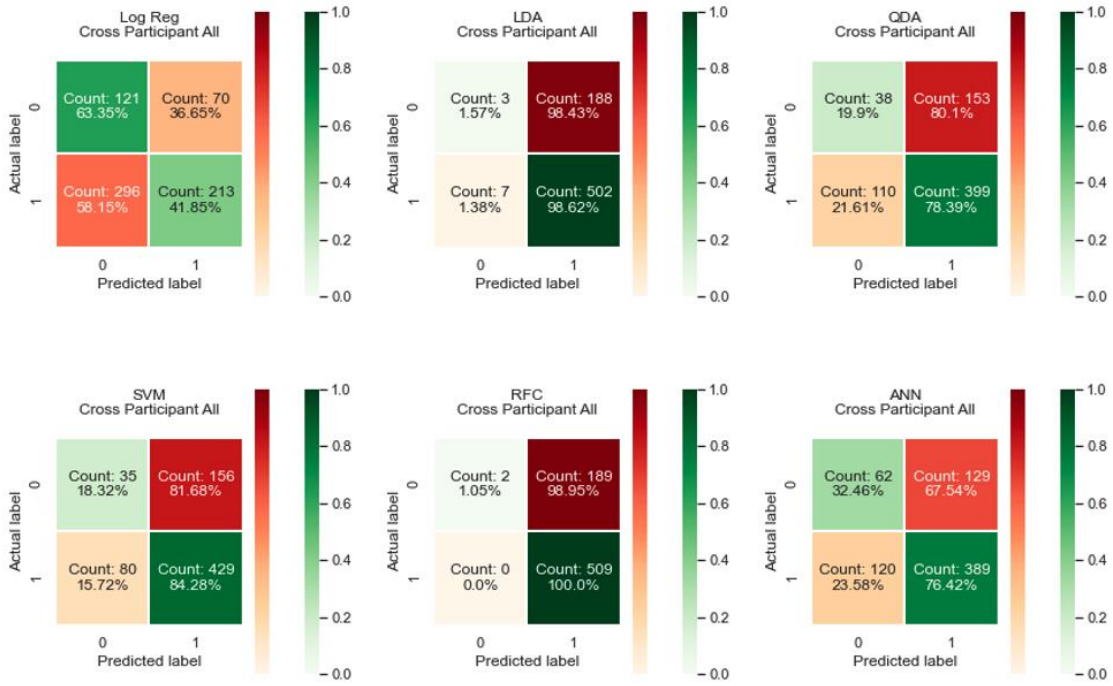

Figure 27. Confusion matrices from the cross-participant human-machine models

Table 23. FPRs and FNRs for the cross-participant human-machine models

Cross-participant	Logistic Regression	LDA	QDA	SVM	RFC	ANN
FPR	37.97%	94.94%	30.38%	94.94%	75.95%	63.29%
FNR	68.86%	2.50%	63.86%	7.73%	21.14%	32.73%

Logistic regression was once again used to look at the influence of the individual features in the dataset and see if there are any apparent reasons to explain the models' performances. The histogram of the absolute value of the coefficient values for each feature is shown in Figure 29. The distribution of coefficient values is still skewed, but with a lower percentage of features in the first few bins. There are even two features with absolute valued coefficients greater than 1.0.

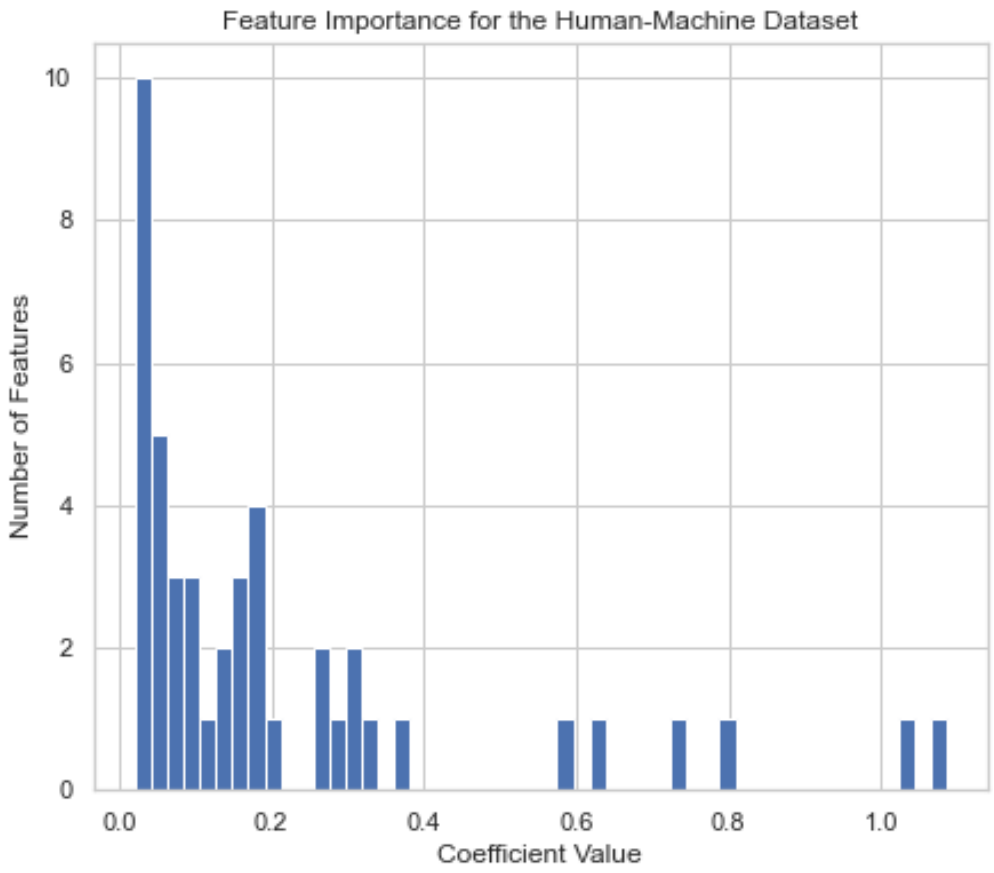

Figure 28. Histogram of the human-machine dataset features and their logistic regression coefficients

The two grouped boxplots in Figures 30 and 31 are of AS08 and CS03's data as they had, respectively, the lowest and highest average balanced accuracies across the six classifier types. After examining the two figures, it is relatively straightforward as to why the models performed as they did. The medians and quartile ranges between the two classes for each impactful feature on AS08's data look to be remarkably similar. On the other hand, the median of the trust class of the six most impactful features in CS03's data is greater than three quarters of the values from the negative class on the same feature.

Patterns like this are easily learned and are likely the reason for the excellent performance by many of the models given this subjects' data.

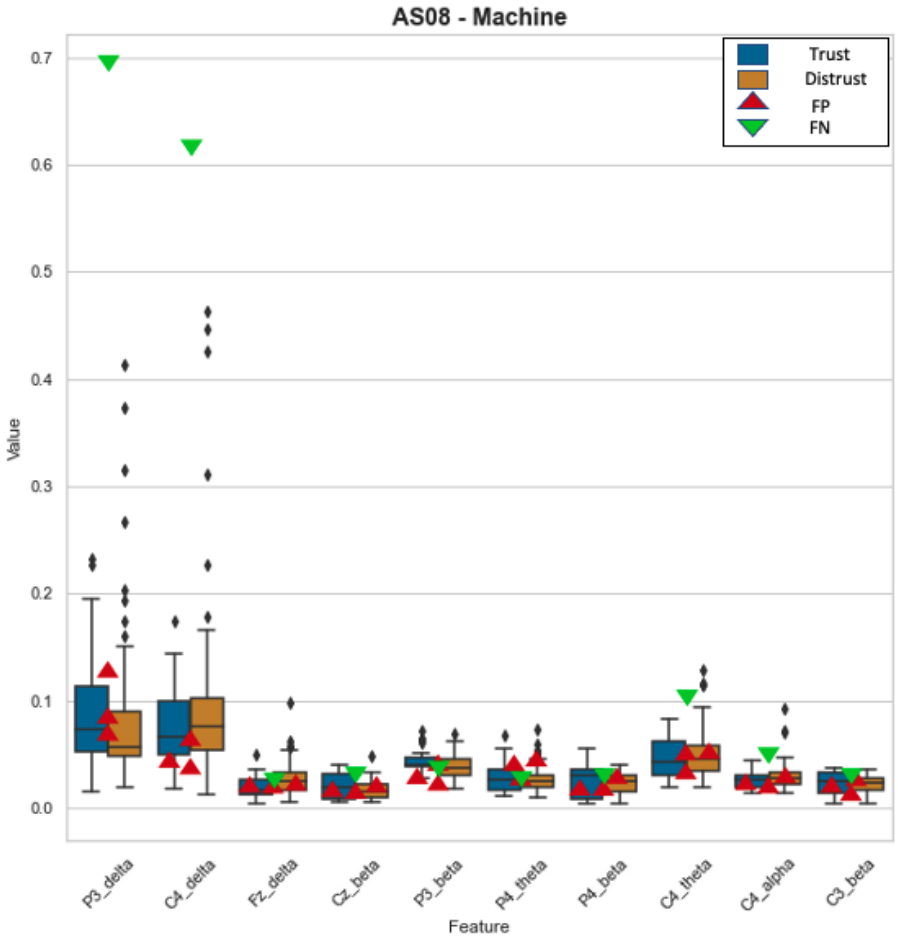

Figure 29. Observation values from the top 10 most influential features for prediction for AS08's data

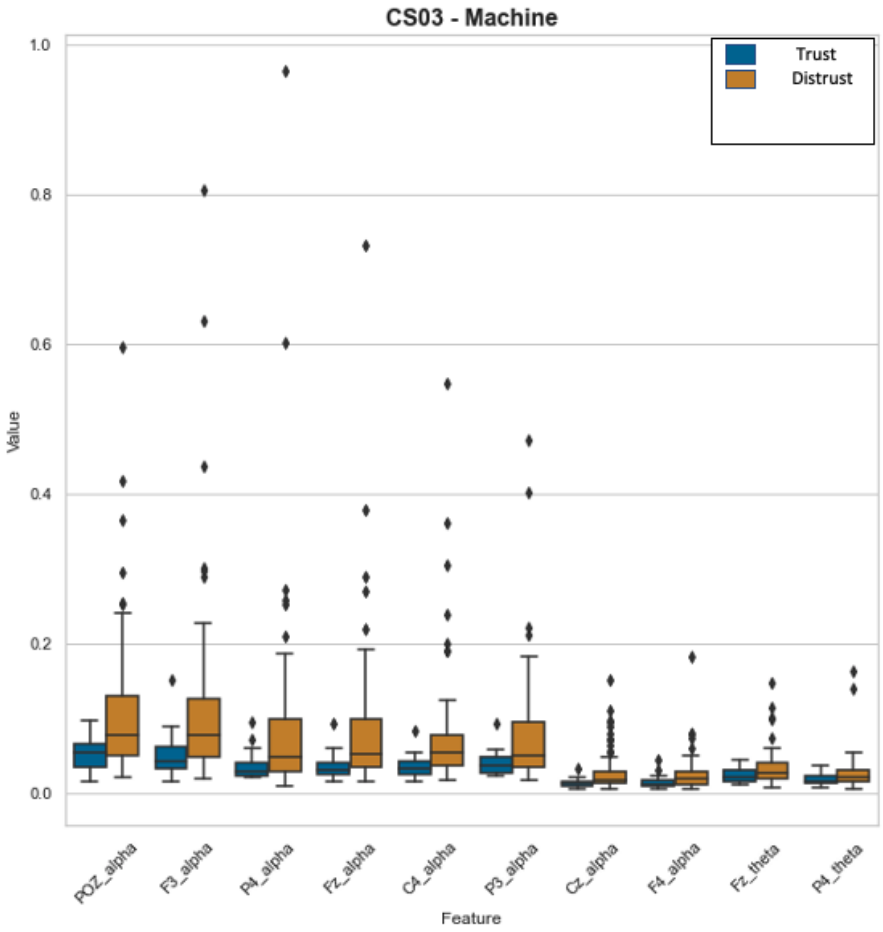

Figure 30. Observation values from the top 10 most influential features for prediction for CS03's data

4.3 Cross Task Machine Learning

This section will examine the performance of the machine learning classifiers that were trained on one of the datasets and then tested on the other. The datasets used in this section are the human-human 9-channel dataset and the human-machine dataset, both of which have the same 45 features per observation. The nine electrode locations are the

same between the two of these datasets, but it should be noted that different branded EEG caps were used in the two experiments. There were two different train, validate, and test splits that were used. The first used one dataset for training and validation and the other dataset was just the test data. In these scenarios, the participants from one dataset were split into four different evenly sized groups, and four-fold cross-validation was used during the training of each model. The other scenario trained on the entirety of one dataset, used half of the other dataset as a validation set, and then the second half was the test set. The balanced accuracies from the results on the test sets from all of these models are displayed in Table 24.

Table 24. Test set cross-task cross-participant balanced accuracies

Train Set	Validate Set	Test Set	Logistic Regression Balanced Accuracy	LDA Balanced Accuracy	QDA Balanced Accuracy	SVM Balanced Accuracy	RFC Balanced Accuracy	ANN Balanced Accuracy
Human	Human	Machine	46.58%	48.77%	50.16%	47.46%	49.51%	49.58%
Human	Machine	Machine	49.42%	42.74%	50.20%	52.88%	49.37%	50.35%
Machine	Machine	Human	48.59%	34.23%	49.05%	47.23%	48.92%	50.26%
Machine	Human	Human	46.74%	50.57%	49.29%	50.00%	49.04%	53.37%

Almost every one of the models performed approximately as well as random chance, if not worst. The highest balanced accuracy of the models trained on the human-human 9-channel dataset and tested on the human-machine dataset was an SVM, which achieved 52.88%. The model with the highest balanced accuracy that trained on the human-machine dataset then predicted the observations in the human-human 9-channel

dataset was an ANN with a score of 53.37%. It can be seen that, on average, the models validated and tested on the same data did perform slightly better than the others.

The confusion matrices for all 24 models were created to see how each classifier made its predictions. The six models trained on the human-machine data and then validated and tested on the human-human 9-channel data are in Figure 32. Only these matrices are shown because the four confusion matrices from the same model on the different tests all have a similar breakdown for five of the classifier types. The logistic regression and QDA models predicted close to half of the observations to be in each class. The LDAs, SVMs, and RFCs, however, predicted nearly every observation to be in the positive class. The only model which is different is the ANN. The ANN that used the human-machine data for the validation and test set made predictions like the ANN whose confusion matrix is depicted. The other two ANNs predicted almost every observation to be trusting.

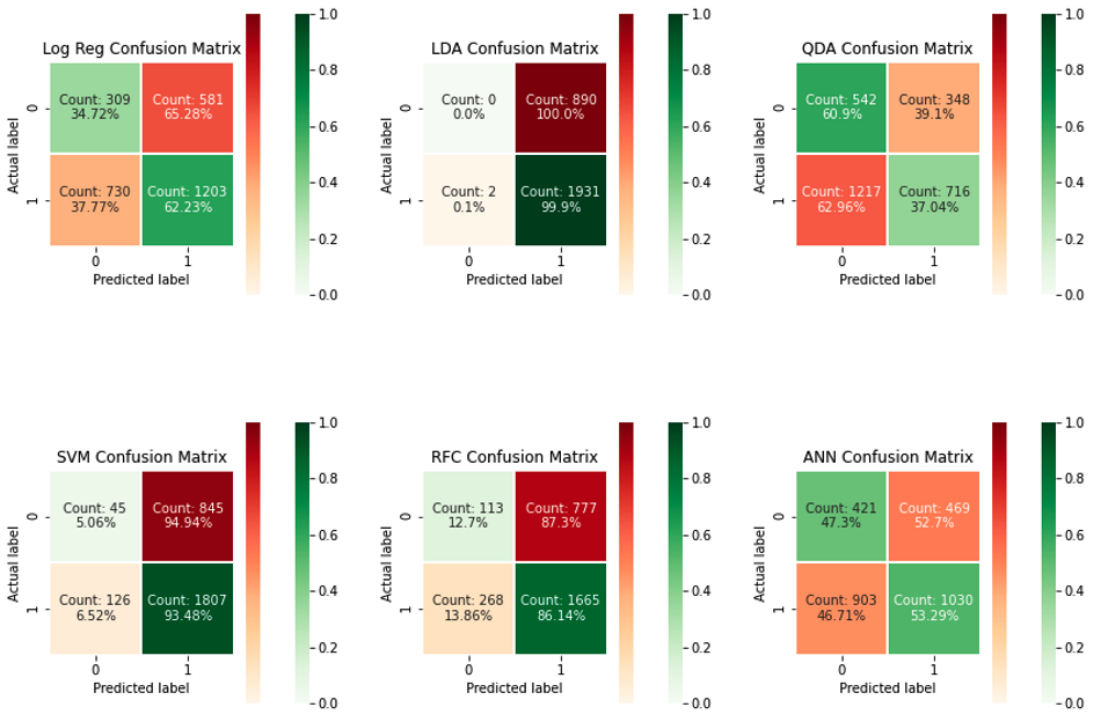

Figure 31. Confusion matrices from the models trained on the human-machine dataset then validated and tested on the human-human 9-channel dataset.

The five most influential features from each dataset were found based on the coefficient values from logistic regression and then were graphed using grouped boxplots. These two graphs showed the feature values for the observations from both datasets though rather than just the one in which data was used to fit the model. The graph with the human-human 9-channel dataset's influential features is in Figure 33, and then Figure 34 is the plot for the s influential features of the human-machine dataset. Only one feature, denoted as Fz_alpha, was one of the top five most influential features in both datasets. The two graphs show no apparent similarities between the same trust class's feature values in different domains. Instead, it looks as if observations would be

much easier to distinguish based on their trust domain than if they were trusting or distrusting. This finding leads to the belief that accurate cross-task cross-participant classification is not possible with this data and that there may be significant differences between the brain activity for interpersonal trust and human-machine trust.

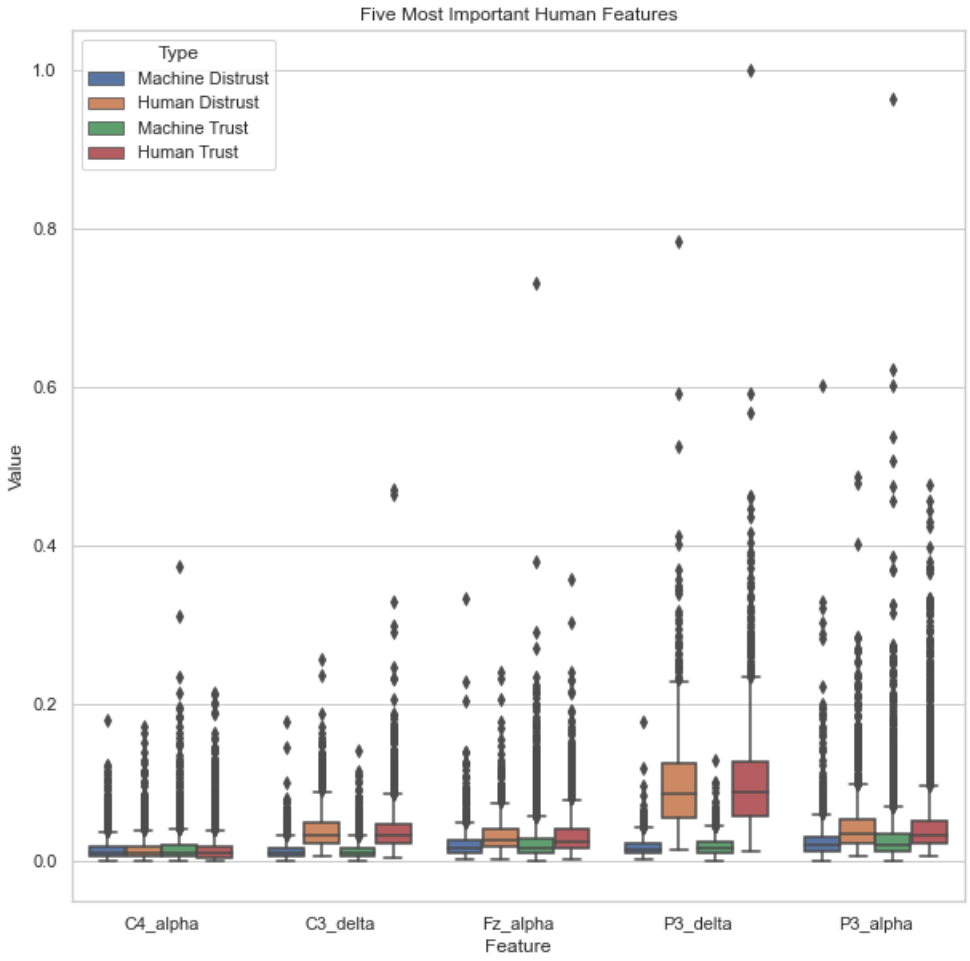

Figure 32. Grouped boxplot of the observation values from both datasets for the five most influential features in the human-human 9-channel dataset.

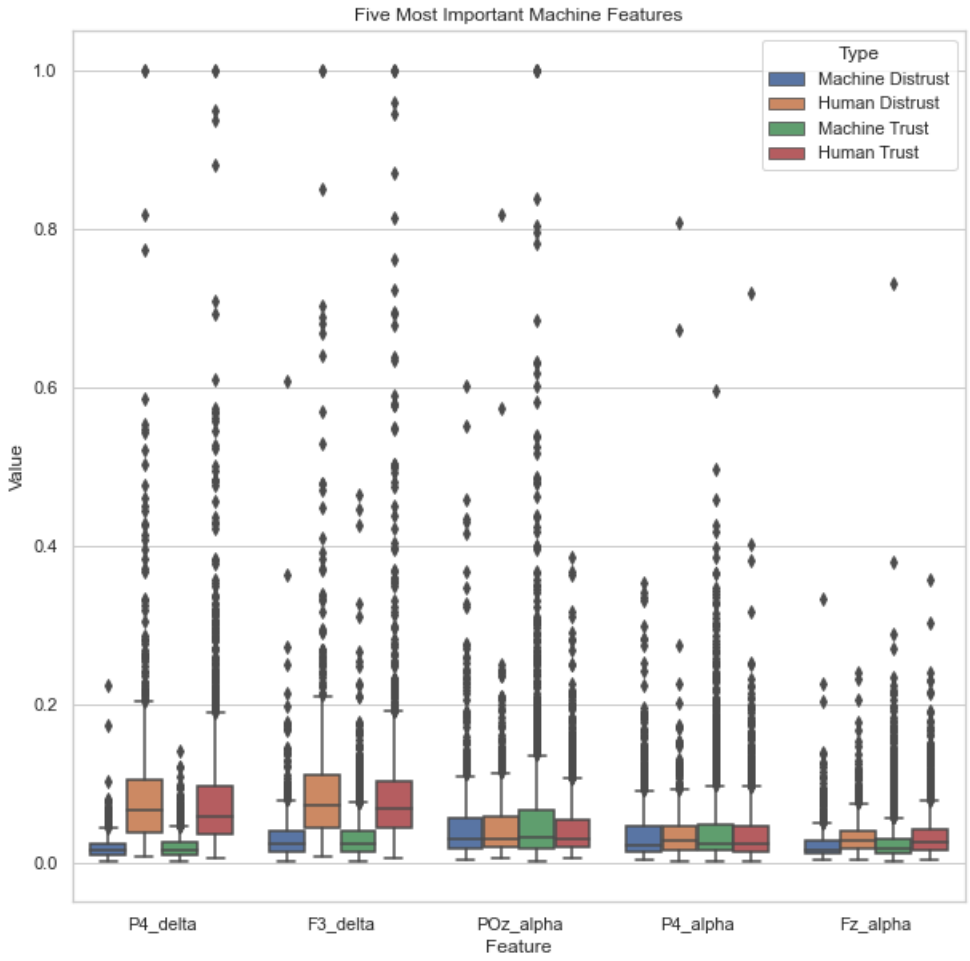

Figure 33. Grouped boxplot of the observation values from both datasets for the five most influential features in the human-machine dataset.

4.4 Manually Picked Feature Datasets

This section describes the results from the experiments investigating if any one set of EEG features claimed to be neural correlates of trust from past research is better than any other one. Since this is a comparison of the sets of features themselves, PCA was not applied to the datasets. The balanced accuracies for the logistic regression models are

displayed in 25. Values in the table are bolded if they are greater than the corresponding human-human dataset balanced accuracy. Logistic regression is shown since these models consistently performed the best in the experiments from Section 4.2 The balanced accuracies from the other five classifier types are in Appendices 5-9.

Table 25. Balanced accuracies from the logistic regression models trained on the neural correlate datasets. Bolded scores are where the neural correlate dataset score is higher than the human-human dataset score

Logistic Regression	Human-Human Balanced Accuracy	Real-Time Sensing Balanced Accuracy	Classification Model Balanced Accuracy	Adaptive Probabilistic Balanced Accuracy
Subject 01	61.93%	39.77%	50.57%	42.05%
Subject 02	70.00%	65.00%	50.00%	42.50%
Subject 03	67.05%	58.52%	53.98%	59.09%
Subject 04	53.97%	36.51%	**63.49%**	**61.90%**
Subject 06	73.02%	61.11%	48.41%	63.49%
Subject 07	52.84%	42.05%	48.30%	26.70%
Subject 08	70.57%	70.57%	61.96%	54.31%
Subject 09	56.35%	**64.29%**	53.17%	**58.73%**
Subject 10	60.27%	53.57%	50.45%	54.46%
Subject 11	63.33%	50.00%	56.67%	36.67%
Subject 12	53.97%	**61.11%**	**69.84%**	42.86%
Subject 13	57.39%	51.70%	57.39%	43.75%
Subject 14	50.00%	**79.17%**	**58.33%**	**58.33%**
Subject 15	55.00%	35.00%	47.50%	50.00%
Subject 16	82.50%	50.00%	62.50%	42.50%
Subject 17	62.78%	49.72%	**73.89%**	62.78%
Subject 18	57.50%	50.00%	37.50%	45.00%
Subject 19	70.00%	55.00%	40.00%	62.50%
Subject 20	50.00%	41.67%	**62.50%**	47.92%
Cross-Participant	51.25%	48.42%	50.66%	49.15%

Based on these results, no group of neural correlates from the past research is clearly better than any other. The most notable thing is that in all three cases, the performance was overall, worse than on the full human-human dataset. Each neural

correlate group had balanced accuracies greater than the corresponding score from the full human-human dataset model on three within-participant datasets. This meant that the models with the full dataset scored better 85% of the time.

4.5 Summary

The experiments investigated whether machine learning could be used to determine a difference between the decision to trust and distrust in EEG. Two datasets from past research were used to investigate this research question. The average frequency values of the alpha, beta, delta, gamma, and theta frequency bands were extracted from the raw EEG data to be used as features. These datasets were used to answer the following research questions.

Research Question 1

Can EEG be associated with the decision to trust or distrust?

Result: Nine single-participant human-human dataset models, 17 single-participant human-human 9-channel dataset models, and 23 single-participant human-machine dataset models achieved a balanced accuracy of 70.00% or greater. No model achieved a mean balanced accuracy across all the participants in a dataset greater than 70.00%. No cross-participant model achieved a balanced accuracy greater than 70.00%.

Research Question 2

Can a machine learning classifier accurately detect cross-task cross-domain trust using EEG?

Result: No cross-task cross-domain model achieved a balanced accuracy greater than 70.00%. The highest accuracy was 53.37% which was achieved by an ANN trained on the human-machine dataset, then validated and tested on the human-human 9-channel dataset.

Research Question 3

Do machine learning models achieve significantly higher classification accuracies when provided with all possible features to begin with or when a subset of features is manually selected based on past research findings investigating neural correlates of trust?

Result: The machine learning models provided with all of the features performed better than the models given the observations with subsampled features 85% of the time.

V. Conclusions and Recommendations

5.1 Conclusions of Research

The results of this research are inconclusive. Based on the machine learning models' performance, there is no way to say for sure that there are EEG signals associated with the decision to trust or not trust. Additionally, there is no evidence that a machine learning model can be used for cross-participant trust and distrust detection of any kind. These conclusions were arrived at after doing thorough testing and comparisons of multiple machine learning models on the two datasets described in Chapter 3.3.

The first research questions (Section 3.2.1) investigated if EEG data can be associated with the decision to trust or distrust. This was researched using machine learning models to classify single task trust data in both within-participant and cross-participant tests. There were four different scenarios examined for this research question, 1) within-participant on human-human trust, 2) within-participant on human-machine trust, 3) cross-participant on human-human trust, and 4) cross-participant on human-machine trust. There were individual models in scenarios 1 and 2 that achieved balanced accuracies greater than 70.00%, but this was not the case on average for any model on any dataset or for any models part of scenarios 3 and 4. These results did not prove any of the hypothesis to be correct.

The second research question investigated whether cross-task cross-domain trust detection was possible. The hypothesis was that a machine learning model could achieve a balanced accuracy greater than 70.00% on this task. The results found do not support the hypothesis but also do not prove that this is not possible. The top performing model

out of both scenarios investigated only achieved an equal-class-weighted score of 53.37%. After further investigation into the feature values across the two datasets, it is unclear as if the brain activity for trusting another human and trusting a machine is similar enough to complete this task.

5.2 Significance of Research

Trust is often considered one of our society's fundamental pillars. It is the foundation of many relationships and allows people to go through their day with some peace of mind. There is a balance needed with trust, however. Going through life blindly, trusting every word read, or voice heard will likely lead one into some danger. Trust must be adequately calibrated for specific scenarios, especially with the exponential growth of computers and algorithms in everyone's daily life. The methods in this research do not allow for entirely accurate or real-time trust classification. As of now, trust is typically evaluated using post-task surveys, these have no way of being done while a task is occurring, and they are prone to bias as they are subjective measures. The results presented here add to the present body of work within trust detection and provide evidence to support the claim that it is a task that can most likely be accomplished using machine learning. However, the results go further than just trust detection, showing a few examples of reasonably accurate cross-participant trust detection. The majority of the models performed just slightly better than chance and still need to be vastly improved, but they will act as an important steppingstone toward what can be done.

5.3 Recommendations for Future Research

The main issue with this experiment is it used two datasets with numerous different variables between them. Only one of which, being the type of entity the trust is being placed in, is desired. The experiments used to collect the data had different participants, asked them to complete different tasks, used a different EEG cap and EEG collection software brand, and were completed by different researchers. The precise effect of these differences on the results is unknown. However, the models' performance would be much less murky if the only difference in the two datasets is that one is with human-human trust and the other is with human-machine trust. It is suggested that further research on this topic includes new experiments and data collection to be conducted.

There are many other avenues as well which can be explored for possible future work. For this paper, only the frequency feature dataset using the five clinical bands' mean spectral values was used. However, many other features could be explored, including those using the time domain rather than just the frequency domain. Some that could be further investigated are entropy, the variation of band frequencies, peak-to-peak values, the mean and variation of frequencies, energy values, and correlated values found between two or more different channels. Additionally, there are many other physiological signals which could be used on their own or along with EEG to help with trust detection.

The last suggestion for future work is to look into different and potentially more complex machine learning classifiers. This paper did not look at some popular deep learning classifiers like recurrent neural networks or convolutional neural networks, nor did it test any ensemble methods. Following any of these potential paths provided, or a

combination of them, may very well lead to an improvement in the machine learning results presented.

5.4 Summary

This thesis explored the possibility of trust detection using machine learning classifiers that were given EEG data. The results found some within-participant models achieving balanced accuracies greater than 70.00%, but not any cross-participant or cross-task cross-participant models. The work used two datasets, one from a human-human trust experiment and one from a human-machine trust experiment. These models indicate that there are EEG signals associated with the decision to trust or distrust, and machine learning classifiers can learn that relationship. It is unclear whether the described relationship is similar across different people and if cross-participant classification of this behavior is possible. From the two datasets used in this experiment, it does not seem likely that cross-task trust detection is possible. Multiple improvements to the current experiment have been suggested that could lead to possible advancements.

Appendix

Appendix 1. Observation count, class distributions, and post-PCA feature count for the human-human dataset and human-human 9-channel dataset

Participant	Observations	Percent Trusting	Percent Distrusting	Post-PCA Human-Human Feature Count	Post-PCA Human-Human 9-Channel Feature Count
Subject 01	148	72.97%	27.03%	59	25
Subject 02	146	66.43%	33.57%	63	23
Subject 03	150	73.33%	26.67%	41	23
Subject 04	150	70.66%	29.34%	11	17
Subject 06	150	70.66%	29.34%	69	30
Subject 07	150	74.66%	25.34%	38	26
Subject 08	149	63.75%	36.25%	67	28
Subject 09	148	70.94%	29.06%	35	19
Subject 10	149	53.02%	46.98%	52	27
Subject 11	149	49.66%	50.34%	41	23
Subject 12	149	69.12%	30.88%	57	26
Subject 13	147	74.14%	25.86%	55	27
Subject 14	148	78.37%	21.63%	57	27
Subject 15	150	67.33%	32.67%	55	25
Subject 16	149	65.10%	34.90%	37	24
Subject 17	145	68.96%	31.04%	55	28
Subject 18	150	66.66%	33.34%	45	21
Subject 19	146	65.75%	34.25%	24	10
Subject 20	150	79.33%	20.67%	43	24
Total	**2972**	**68.41%**	**31.59%**		

Appendix 2. Number of FPs and FNs missed by all six models for each subjects' data in the human-human dataset

Participant	All False Positive Count	All False Negative Count
Subject 01	1	0
Subject 02	1	1
Subject 03	0	1
Subject 04	0	0
Subject 06	0	2
Subject 07	2	0
Subject 08	2	0
Subject 09	2	0
Subject 10	0	1
Subject 11	1	1
Subject 12	0	0
Subject 13	0	0
Subject 14	0	0
Subject 15	3	0
Subject 16	1	0
Subject 17	1	0
Subject 18	3	0
Subject 19	0	0
Subject 20	3	0
Cross - Subject 03	18	0
Cross - Subject 06	7	0
Cross - Subject 09	16	0
Cross - Subject 20	40	0

Appendix 3. Number of FPs and FNs missed by all six models for each subjects' data in the human-human 9-channel dataset

Participant	All False Positive Count	All False Negative Count
Subject 01	1	0
Subject 02	0	0
Subject 03	1	0
Subject 04	0	2
Subject 06	1	0
Subject 07	5	0
Subject 08	1	0
Subject 09	4	0
Subject 10	2	1
Subject 11	0	1
Subject 12	3	0
Subject 13	1	0
Subject 14	3	0
Subject 15	4	0
Subject 16	1	0
Subject 17	4	0
Subject 18	0	1
Subject 19	1	0
Subject 20	4	0
Cross - Subject 03	6	0
Cross - Subject 06	0	0
Cross - Subject 09	3	0
Cross - Subject 20	3	0

Appendix 4. Observation count, class distributions, and post-PCA feature count for the human-machine dataset

Participant	Observations	Percent Trusting	Percent Distrusting	Post-PCA Human-Human Feature Count
AS05	100	68.00%	32.00%	22
AS07	100	70.00%	30.00%	23
AS08	100	71.00%	29.00%	25
AS10	100	62.00%	38.00%	24
AS12	100	71.00%	29.00%	24
AS13	100	69.00%	31.00%	21
AS14	100	59.00%	41.00%	23
AS15	100	79.00%	21.00%	23
AS16	100	76.00%	24.00%	23
AS17	100	68.00%	32.00%	22
AS20	100	77.00%	23.00%	21
AS21	100	67.00%	33.00%	24
AS23	100	74.00%	26.00%	21
AS24	100	70.00%	30.00%	25
AS27	100	75.00%	25.00%	21
AS28	100	73.00%	27.00%	19
AS30	100	73.00%	27.00%	25
AS33	100	73.00%	27.00%	25
AS35	100	69.00%	31.00%	26
AS36	100	53.00%	47.00%	24
BS02	100	77.00%	23.00%	22
BS04	100	57.00%	43.00%	23
BS05	100	74.00%	26.00%	19
BS06	100	79.00%	21.00%	26
BS07	100	71.00%	29.00%	25
BS09	100	63.00%	37.00%	26
BS10	100	71.00%	29.00%	25
CS01	100	61.00%	39.00%	25
CS03	100	76.00%	24.00%	27
CS04	100	78.00%	22.00%	19
Total	**3100**	70.45%	29.55%	

Appendix 5. Balanced accuracies from the LDA models trained on the neural correlate datasets

LDA	Human-Human Balanced Accuracy	Real-Time Sensing Balanced Accuracy	Classification Model Balanced Accuracy	Adaptive Probabilistic Balanced Accuracy
Subject 01	57.95%	50.00%	45.45%	50.00%
Subject 02	67.50%	55.00%	45.00%	52.50%
Subject 03	76.14%	51.70%	50.00%	50.00%
Subject 04	55.56%	50.00%	55.56%	55.56%
Subject 06	57.14%	56.35%	51.59%	47.62%
Subject 07	38.64%	50.00%	43.18%	45.45%
Subject 08	60.29%	51.20%	55.74%	55.74%
Subject 09	50.79%	53.17%	47.62%	47.62%
Subject 10	66.52%	52.68%	59.82%	49.55%
Subject 11	70.00%	53.33%	56.67%	36.67%
Subject 12	53.97%	56.35%	56.35%	48.41%
Subject 13	61.93%	62.50%	50.00%	62.50%
Subject 14	41.67%	50.00%	47.92%	58.33%
Subject 15	60.00%	40.00%	40.00%	40.00%
Subject 16	50.00%	50.00%	50.00%	37.50%
Subject 17	65.28%	53.61%	66.67%	53.06%
Subject 18	40.00%	50.00%	45.00%	47.50%
Subject 19	52.50%	50.00%	47.50%	45.00%
Subject 20	45.83%	50.00%	47.92%	50.00%
Cross-Participant	51.25%	50.00%	50.32%	50.63%

Appendix 6. Balanced accuracies from the QDA models trained on the neural correlate datasets

QDA	Human-Human Balanced Accuracy	Real-Time Sensing Balanced Accuracy	Classification Model Balanced Accuracy	Adaptive Probabilistic Balanced Accuracy
Subject 01	64.20%	41.48%	61.36%	48.86%
Subject 02	62.50%	35.00%	50.00%	45.00%
Subject 03	60.80%	44.89%	62.50%	58.52%
Subject 04	52.38%	50.79%	70.63%	52.38%
Subject 06	42.86%	62.70%	50.79%	49.21%
Subject 07	50.00%	50.00%	42.05%	42.61%
Subject 08	50.00%	64.83%	53.11%	58.85%
Subject 09	59.52%	42.86%	47.62%	47.62%
Subject 10	58.48%	40.18%	40.18%	54.91%
Subject 11	66.67%	50.00%	60.00%	56.67%
Subject 12	44.44%	44.44%	48.41%	39.68%
Subject 13	53.41%	61.36%	57.95%	49.43%
Subject 14	43.75%	54.17%	39.58%	56.25%
Subject 15	32.50%	40.00%	42.50%	37.50%
Subject 16	55.00%	45.00%	50.00%	42.50%
Subject 17	63.33%	51.11%	65.28%	59.72%
Subject 18	42.50%	60.00%	52.50%	52.50%
Subject 19	45.00%	55.00%	45.00%	45.00%
Subject 20	43.75%	45.83%	50.00%	47.92%
Cross-Participant	51.25%	49.09%	47.51%	48.32%

Appendix 7. Balanced accuracies from the SVMs trained on the neural correlate datasets

SVM	Human-Human Balanced Accuracy	Real-Time Sensing Balanced Accuracy	Classification Model Balanced Accuracy	Adaptive Probabilistic Balanced Accuracy
Subject 01	55.11%	52.27%	52.27%	52.27%
Subject 02	50.00%	50.00%	50.00%	47.50%
Subject 03	81.82%	50.00%	50.00%	50.00%
Subject 04	54.76%	50.00%	50.00%	50.00%
Subject 06	63.49%	50.00%	50.00%	50.00%
Subject 07	44.89%	37.50%	55.68%	32.95%
Subject 08	66.03%	51.20%	69.14%	60.77%
Subject 09	58.73%	53.17%	47.62%	53.17%
Subject 10	52.23%	50.00%	50.00%	49.11%
Subject 11	66.67%	50.00%	56.67%	53.33%
Subject 12	53.17%	53.97%	62.70%	45.24%
Subject 13	50.00%	48.86%	45.45%	41.48%
Subject 14	50.00%	60.42%	70.83%	58.33%
Subject 15	50.00%	35.00%	37.50%	47.50%
Subject 16	60.00%	50.00%	50.00%	45.00%
Subject 17	53.06%	47.22%	68.89%	55.28%
Subject 18	32.50%	60.00%	35.00%	50.00%
Subject 19	60.00%	42.50%	50.00%	57.50%
Subject 20	58.33%	56.25%	54.17%	50.00%
Cross-Participant	51.25%	49.17%	50.81%	45.20%

Appendix 8. Balanced accuracies from the RFCs trained on the neural correlate datasets

RFC	Human-Human Balanced Accuracy	Real-Time Sensing Balanced Accuracy	Classification Model Balanced Accuracy	Adaptive Probabilistic Balanced Accuracy
Subject 01	53.41%	46.59%	47.73%	45.45%
Subject 02	50.00%	50.00%	55.00%	42.50%
Subject 03	47.73%	56.25%	53.98%	53.41%
Subject 04	59.52%	50.00%	53.97%	58.73%
Subject 06	35.71%	50.79%	64.29%	50.79%
Subject 07	60.23%	53.98%	50.00%	53.98%
Subject 08	50.00%	57.66%	44.50%	47.37%
Subject 09	59.52%	50.00%	41.27%	49.21%
Subject 10	63.39%	45.09%	46.88%	55.80%
Subject 11	56.67%	60.00%	53.33%	53.33%
Subject 12	47.62%	58.73%	51.59%	50.00%
Subject 13	50.00%	53.98%	51.70%	50.00%
Subject 14	50.00%	50.00%	50.00%	77.08%
Subject 15	62.50%	42.50%	47.50%	37.50%
Subject 16	62.50%	45.00%	60.00%	37.50%
Subject 17	56.39%	40.00%	54.17%	58.61%
Subject 18	50.00%	47.50%	40.00%	52.50%
Subject 19	55.00%	50.00%	50.00%	50.00%
Subject 20	50.00%	43.75%	56.25%	56.25%
Cross-Participant	51.25%	48.79%	51.04%	50.11%

Appendix 9. Balanced accuracies from the ANNs trained on the neural correlate datasets

ANN	Human-Human Balanced Accuracy	Real-Time Sensing Balanced Accuracy	Classification Model Balanced Accuracy	Adaptive Probabilistic Balanced Accuracy
Subject 01	53.41%	50.00%	50.00%	50.00%
Subject 02	55.00%	47.50%	50.00%	52.50%
Subject 03	60.23%	50.00%	50.00%	50.00%
Subject 04	54.76%	50.00%	50.00%	50.00%
Subject 06	65.08%	50.00%	50.00%	45.24%
Subject 07	56.25%	50.00%	50.00%	50.00%
Subject 08	51.20%	50.00%	54.55%	44.74%
Subject 09	50.00%	50.00%	50.00%	47.62%
Subject 10	52.23%	58.48%	51.34%	50.89%
Subject 11	70.00%	56.67%	50.00%	50.00%
Subject 12	56.35%	50.00%	50.00%	47.62%
Subject 13	46.02%	50.00%	47.73%	50.00%
Subject 14	60.42%	50.00%	50.00%	50.00%
Subject 15	52.50%	37.50%	40.00%	45.00%
Subject 16	67.50%	50.00%	50.00%	47.50%
Subject 17	52.22%	50.00%	50.00%	50.00%
Subject 18	57.50%	50.00%	50.00%	50.00%
Subject 19	50.00%	50.00%	55.00%	50.00%
Subject 20	52.08%	50.00%	50.00%	50.00%
Cross-Participant	51.25%	47.81%	50.79%	50.56%

Bibliography

Ackermann, P., Kohlschein, C., Bitsch, J. A., Wehrle, K., & Jeschke, S. (2016). EEG-based automatic emotion recognition: Feature extraction, selection and classification methods. *2016 IEEE 18th International Conference on E-Health Networking, Applications and Services (Healthcom)*, 1–6. https://doi.org/10.1109/HealthCom.2016.7749447

Ajenaghughrure, I. B., Sousa, S. D. C., & Lamas, D. (2020). Measuring Trust with Psychophysiological Signals: A Systematic Mapping Study of Approaches Used. *Multimodal Technologies and Interaction*, *4*(3), 63. https://doi.org/10.3390/mti4030063

Akash, K., Hu, W. L., Jain, N., & Reid, T. (2018). A classification model for sensing human trust in machines using EEG and GSR. *ACM Transactions on Interactive Intelligent Systems*, *8*(4). https://doi.org/10.1145/3132743

Akash, K., Reid, T., & Jain, N. (2018). Adaptive Probabilistic Classification of Dynamic Processes: A Case Study on Human Trust in Automation. *Proceedings of the American Control Conference*, *2018-June*, 246–251. https://doi.org/10.23919/ACC.2018.8431132

Al-Fahoum, A. S., & Al-Fraihat, A. A. (2014). Methods of EEG Signal Features Extraction Using Linear Analysis in Frequency and Time-Frequency Domains. *ISRN Neuroscience*, *2014*(September), 1–7. https://doi.org/10.1155/2014/730218

Alós-Ferrer, C., & Farolfi, F. (2019). Trust Games and Beyond. *Frontiers in Neuroscience*, *13*(September), 1–14. https://doi.org/10.3389/fnins.2019.00887

Atkinson, D. J., & Clark, M. H. (2013). Autonomous agents and human interpersonal trust: Can we engineer a human-machine social interface for trust? *AAAI Spring Symposium - Technical Report*, *SS-13-07*, 2–7.

Beres, A. M. (2017). Time is of the Essence: A Review of Electroencephalography (EEG) and Event-Related Brain Potentials (ERPs) in Language Research. *Applied Psychophysiology Biofeedback*, *42*(4), 247–255. https://doi.org/10.1007/s10484-017-9371-3

Bhattacharyya, S., Khasnobish, A., Chatterjee, S., Konar, A., & Tibarewala, D. N. (2010). Performance Analysis of LDA, QDA, and KNN Algorithms in Left-Right limb movement classification from EEG data. *2010 International Conference on Systems in Medicine and Biology*, 126–131. https://doi.org/10.1109/ICSMB.2010.5735358

Bigdely-Shamlo, N., Mullen, T., Kothe, C., Su, K. M., & Robbins, K. A. (2015). The PREP pipeline: Standardized preprocessing for large-scale EEG analysis. *Frontiers in Neuroinformatics*, *9*(JUNE), 1–19. https://doi.org/10.3389/fninf.2015.00016

Binias, B., Myszor, D., & Cyran, K. A. (2018). A machine learning approach to the detection of pilot's reaction to unexpected events based on EEG signals. *Computational Intelligence and Neuroscience*, *2018*. https://doi.org/10.1155/2018/2703513

Chancey, E. T., Bliss, J. P., Yamani, Y., & Handley, H. A. H. (2017). Trust and the Compliance-Reliance Paradigm: The Effects of Risk, Error Bias, and Reliability on Trust and Dependence. *Human Factors*, *59*(3), 333–345.

https://doi.org/10.1177/0018720816682648
Chicco, D., & Jurman, G. (2020). The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics*, *21*(1), 1–13. https://doi.org/10.1186/s12864-019-6413-7
Cho, J. H., Chan, K., & Adali, S. (2015). A Survey on Trust Modeling. *ACM Computing Surveys*, *48*(2). https://doi.org/10.1145/2815595
Cohen, M. X. (2014). *Analyzing neural time series data theory and practice*. MIT Press. http://cognet.mit.edu/book/analyzing-neural-time-series-data
de Visser, E. J., Beatty, P. J., Estepp, J. R., Kohn, S., Abubshait, A., Fedota, J. R., & McDonald, C. G. (2018). Learning From the Slips of Others: Neural Correlates of Trust in Automated Agents. *Frontiers in Human Neuroscience*, *12*(August), 1–15. https://doi.org/10.3389/fnhum.2018.00309
de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied*, *22*(3), 331–349. https://doi.org/10.1037/xap0000092
Dhiman, R., Priyanka, & Saini, J. S. (2013). *Wavelet Analysis of Electrical Signals from Brain: The Electroencephalogram BT - Quality, Reliability, Security and Robustness in Heterogeneous Networks* (K. Singh & A. K. Awasthi (eds.); pp. 283–289). Springer Berlin Heidelberg.
Dimoka, A., Pavlou, P. A., & Davis, F. D. (2011). Research Commentary: NeuroIS: The Potential of Cognitive Neuroscience for Information Systems Research. *Information Systems Research*, *22*(4), 687–702. https://doi.org/10.1287/isre.1100.0284
Dvey-Aharon, Z., Fogelson, N., Peled, A., & Intrator, N. (2015). Schizophrenia detection and classification by advanced analysis of EEG recordings using a single electrode approach. *PLoS ONE*, *10*(4), 1–12. https://doi.org/10.1371/journal.pone.0123033
Euston, D. R., Gruber, A. J., & McNaughton, B. L. (2012). The Role of Medial Prefrontal Cortex in Memory and Decision Making. *Neuron*, *76*(6), 1057–1070. https://doi.org/10.1016/j.neuron.2012.12.002
Fareri, D. S., Chang, L. J., & Delgado, M. R. (2015). Computational substrates of social value in interpersonal collaboration. *Journal of Neuroscience*, *35*(21), 8170–8180. https://doi.org/10.1523/JNEUROSCI.4775-14.2015
Fu, C., Yao, X., Yang, X., Zheng, L., Li, J., & Wang, Y. (2019). Trust Game Database: Behavioral and EEG Data From Two Trust Games. *Frontiers in Psychology*, *10*(December), 1–6. https://doi.org/10.3389/fpsyg.2019.02656
George, A. (2012). Anomaly Detection based on Machine Learning Dimensionality Reduction using PCA and Classification using SVM. *International Journal of Computer Applications*, *47*(21), 5–8. https://doi.org/10.5120/7470-0475
Hare, T. (2014). Exploiting and exploring the options. *Science*, *344*(6191), 1446 LP – 1447. https://doi.org/10.1126/science.1256862
Hosking, G. (2002). Why we need a history of trust. *Reviews in History*.
Hu, W. L., Akash, K., Jain, N., & Reid, T. (2016). Real-Time Sensing of Trust in Human-Machine Interactions. *IFAC-PapersOnLine*, *49*(32), 48–53. https://doi.org/10.1016/j.ifacol.2016.12.188
James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). *An Introduction to Statistical*

Learning. https://doi.org/10.1007/978-1-4614-7138-7
Jung, E. S., Dong, S. Y., & Lee, S. Y. (2019). Neural Correlates of Variations in Human Trust in Human-like Machines during Non-reciprocal Interactions. *Scientific Reports*, *9*(1), 1–10. https://doi.org/10.1038/s41598-019-46098-8
Jurcak, V., Tsuzuki, D., & Dan, I. (2007). 10/20, 10/10, and 10/5 systems revisited: their validity as relative head-surface-based positioning systems. *NeuroImage*, *34*(4), 1600–1611. https://doi.org/10.1016/j.neuroimage.2006.09.024
Kramer, R. M. (1999). Trust and distrust in organizations: Emerging perspectives, enduring questions. *Annual Review of Psychology*, *50*, 569–598. https://doi.org/10.1146/annurev.psych.50.1.569
Lee, Y. Y., & Hsieh, S. (2014). Classifying different emotional states by means of eegbased functional connectivity patterns. *PLoS ONE*, *9*(4), 1–17. https://doi.org/10.1371/journal.pone.0095415
Leichtenstern, K., Bee, N., André, E., Berkmüller, U., & Wagner, J. (2011). Physiological measurement of trust-related behavior in trust-neutral and trust-critical situations. *IFIP Advances in Information and Communication Technology*, *358 AICT*, 165–172. https://doi.org/10.1007/978-3-642-22200-9_14
Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: an integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708
Nuamah, J., & Seong, Y. (2017). *Electroencephalography (EEG) Classification of Cognitive Tasks Based on Task Engagement Index*. http://techlav.ncat.edu/publications/2017/CogSIMA_2017_Final_Manuscript 01.pdf
Oh, S., Seong, Y., & Yi, S. (2017). Preliminary study on neurological measure of human Trust in autonomous systems. *67th Annual Conference and Expo of the Institute of Industrial Engineers 2017*, 1066–1072.
Oostenveld, R. (2014). *Improving EEG source analysis using prior knowledge PDF hosted at the Radboud Repository of the Radboud University Nijmegen This full text is a publisher ' s version* . (Issue July).
Oxley, B. C. (2017). *International 10-20 system for EEG electrode placement*. https://upload.wikimedia.org/wikipedia/commons/6/6e/International_10-20_system_for_EEG-MCN.svg
Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055
Piryatinska, A., Darkhovsky, B., & Kaplan, A. (2017). Binary classification of multichannel-EEG records based on the ϵ-complexity of continuous vector functions. *Computer Methods and Programs in Biomedicine*, *152*(December 2019), 131–139. https://doi.org/10.1016/j.cmpb.2017.09.001
Regulinski, T. L. (1962). The Air Force Institute of Technology. *IRE Transactions on Education*, *5*(2), 117–118. https://doi.org/10.1109/TE.1962.4322266
Riedl, R., & Javor, A. (2012). The biology of trust: Integrating evidence from genetics, endocrinology, and functional brain imaging. *Journal of Neuroscience, Psychology, and Economics*, *5*(2), 63–91. https://doi.org/10.1037/a0026318
Sauer, J., & Chavaillaz, A. (2018). How operators make use of wide-choice adaptable

automation: observations from a series of experimental studies. *Theoretical Issues in Ergonomics Science*, *19*(2), 135–155. https://doi.org/10.1080/1463922X.2017.1297866

Sayler, K. M. (2020). Artificial Intelligence and National Security – Economic Impacts and Considerations. *Congressional Research Service*, *June 2020*, 1–43.

Shmueli, B. (2019). Matthews Correlation Coefficient is The Best Classification Metric You've Never Heard Of. In *Towardsdatascience.Com* (pp. 1–8). https://towardsdatascience.com/the-best-classification-metric-youve-never-heard-of-the-matthews-correlation-coefficient-3bf50a2f3e9a

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making? *International Journal of Human Computer Studies*, *51*(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Takayama, L., Ju, W., & Nass, C. (2008). Beyond dirty, dangerous and dull: What everyday people think robots should do. *HRI 2008 - Proceedings of the 3rd ACM/IEEE International Conference on Human-Robot Interaction: Living with Robots*, 25–32. https://doi.org/10.1145/1349822.1349827

USAF. (2010). *Technology Horizons: A Vision for Air Force Science and Technology 2010-2030*. Office of Chief Scientist (AF/ST), Washington, DC.

Wang, M., Hussein, A., Rojas, R. F., Shafi, K., & Abbass, H. A. (2018). EEG-Based Neural Correlates of Trust in Human-Autonomy Interaction. *2018 IEEE Symposium Series on Computational Intelligence (SSCI)*, 350–357. https://doi.org/10.1109/SSCI.2018.8628649

Yuen, C. T., San, W. S., Seong, T. C., & Rizon, M. (2013). Classification of Human Emotions from EEG Signals using Statistical Features and Neural Network. *International Journal of Integrated Engineering*, *1*(3), 190. http://penerbit.uthm.edu.my/ojs/index.php/ijie/article/view/118

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 074-0188

PLEASE DO NOT RETURN YOUR FORM TO THE ABOVE ADDRESS.

1. REPORT DATE *(DD-MM-YYYY)*	**2. REPORT TYPE**	**3. DATES COVERED** *(From – To)*
25-03-2021	Master's Thesis	Sep 2019 – Mar 2021

TITLE AND SUBTITLE

Comparison of Machine Learning Techniques on Trust Detection Using EEG

5a. CONTRACT NUMBER

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)

Elkins, James R., Second Lieutenant, USAF

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAMES(S) AND ADDRESS(S)

Air Force Institute of Technology
Graduate School of Engineering and Management (AFIT/EN)
2950 Hobson Way, Building 640
WPAFB OH 45433-8865

8. PERFORMING ORGANIZATION REPORT NUMBER

AFIT-ENG-MS-21-M-033

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)

Air Force Office of Scientific Research
Dr. James Lawton
AFOSR Program manager, Information and Networks (703) 696-5999
james.lawton.1@us.af.mil

10. SPONSOR/MONITOR'S ACRONYM(S)

AFOSR/OTA

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

12. DISTRIBUTION/AVAILABILITY STATEMENT

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

13. SUPPLEMENTARY NOTES

14. ABSTRACT

Trust is a pillar of society and is a fundamental aspect in every relationship. With the use of automated agents in today's workforce exponentially growing, being able to actively monitor an individual's trust level that is working with the automation is becoming increasingly more important. Humans often have miscalibrated trust in automation and therefore are prone to making costly mistakes. Since deciding to trust or distrust has been shown to correlate with specific brain activity, it is thought that there are EEG signals which are associated with this decision. Using both a human-human trust and a human-machine trust EEG dataset from past research, within-participant, cross-participant, and cross-task cross-participant trust detection was attempted. Six machine learning models, logistic regression, LDA, QDA, SVM, RFC, and an ANN, were used for each experiment. Multiple within-participant models had balanced accuracies greater than 70.00%, but no cross-participant or cross-participant cross task models achieved this.

15. SUBJECT TERMS

Trust Detection, EEG, Machine Learning, physiological signals

16. SECURITY CLASSIFICATION OF:			**17. LIMITATION OF ABSTRACT**	**18. NUMBER OF PAGES**	**19a. NAME OF RESPONSIBLE PERSON**
a. REPORT	**b. ABSTRACT**	**c. THIS PAGE**			Dr. Brett Borghetti, AFIT/ENG
U	**U**	**U**	**UU**	130	**19b. TELEPHONE NUMBER** *(Include area code)* (937) 255-3636 x4612, (brett.borghetti@afit.edu)

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