THESIS

NEURAL NETWORK SECURITY AND OPTIMIZATION FOR SINGLE-PERSON AUTHENTICATION USING ELECTROENCEPHALOGRAM DATA

Submitted by

Naomi Andre

School of Biomedical Engineering

In partial fulfillment of the requirements

For the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2022

Master's Committee:

Advisor: Steve Simske

Jennifer Mueller
Michael Lyons

ABSTRACT

NEURAL NETWORK SECURITY AND OPTIMIZATION FOR SINGLE-PERSON AUTHENTICATION USING ELECTROENCEPHALOGRAM DATA

Security is an important focus for devices that use biometric data, and as such security around authentication needs to be considered. This is true for brain-computer interfaces (BCIs), which often use electroencephalogram (EEG) data as inputs and neural network classification to determine their function. EEG data can also serve as a form of biometric authentication, which would contribute to the security of these devices. Neural networks have also used a method known as ablation to improve their efficiency. In light of this info, the goal of this research is to determine whether neural network ablation can also be used as a method to improve security by reducing a network's learning capabilities to include authenticating only a given target, and preventing adversaries from training new data to be authenticated. Data on the change in entropy of weight values of the networks after training was also collected for the purpose of determining patterns in weight distribution.

Results from a set of ablated networks to a set of baseline (non-ablated) networks for five targets chosen randomly from a data set of 12 people were compared. The results found that ablated maintained accuracy through the ablation process, but that they did not perform as well as the baseline networks. Change in performance between single-target authentication and target-plus-invader authentication was also examined, but no significant results were found. Furthermore, the change in entropy differed between both baseline networks and ablated networks, as well as between single-target authentication and target-plus-invader authentication for all networks. Ablation was determined to have potential for security applications that need to be expanded on, and weight distribution was found to have some correlation with the complexity of an input to a network.

TABLE OF CONTENTS

INTRODUCTION

For anyone who has ever used a machine or program designed specifically for them, it is of the utmost importance to ensure that said machinery or programming cannot be misused by an adversary. For this reason, the application of security to biometrics, particularly authentication, is a priority for brain-computer interfaces (BCIs), and other similar devices. As electroencephalogram (EEG) signals are unique, ubiquitous, and are often used to operate BCIs, they are well-suited to being a focus for improving security (Abdullah et al., 2010; Bao, Wang, & Hu, 2009; Campisi & Rocca, 2014; Riera et al. 2007). EEG signals have also been used for subject classification before; in particular, the usage of visually evoked potentials (VEPs) have aided in EEG classification (Palaniappan, 2004). In this thesis, EEG classification and authentication (using VEPs to obtain data) will be explored in order to better understand classification methodologies, with brief exploration into Hamming classifiers before focusing on neural networks. This includes an examination of the distribution of the values of weights in neural networks used for classification, as well as using entropy measurements of these weight values in order to examine patterns. The thesis will also discuss ablation, which is the method of removing certain nodes from a neural network's hidden layers as a means of making neural networks more efficient, and consider the ways in which ablation could be combined with EEG authentication.

Combining these foci together poses the question of whether or not ablation be used to customize networks trained for authentication to improve security. For the purpose of this thesis, the security would include preventing the network from being re-trained to accept anyone new into the authentication group, while still providing authentication to those who are in the access control list for the network. This is done by ablating a certain number of nodes from a given neural network, while measuring accuracy and change in entropy of the weights along the way. These results are then compared to a baseline set of nets, which here are neural networks of the same size as the ablated

ones, but that are trained and tested as they are without being ablated from a larger network. Both of these sets of nets (baseline and ablated) are also re-trained along the way to examine their ability, or lack thereof, to allow for new members of the authentication group.

LITERATURE REVIEW

Part 1. Biometrics and Electroencephalogram

While the term "biometrics" usually refers to features such as fingerprint, iris, and gait, and can also include other chemical and behavioral means of assessing identity, additional means of biometric identification come in the form of bioelectrical signals such as the electroencephalogram (EEG) and the signals from the brain that it measures. EEG signals are useful for biometrics due to their ubiquity in fauna, a trait which doubles as a liveness detector, as well as being resistant to being faked (Abdullah et al., 2010; Bao, Wang, & Hu, 2009; Campisi & Rocca, 2014; Riera et al. 2007). EEG signals also change when the participant is undergoing emotional stress, which can prevent people from being forced into allowing an adversary authenticated access to something protected by an EEG authentication system (Marcel & Millan, 2007). When compared to other means of collecting signals from the brain, EEG has the advantage of being more mobile and user friendly than other devices (such as measuring brain function via fMRI or MEG). If implemented properly, identification and authentication systems can benefit from using EEG as a biometric instead of physical characteristics. Despite the promise that EEG shows for biometrics, there are still some obstacles barring EEG signals from becoming a commonplace biometric technology; most prominently, the barrier of gathering the data requires applying the EEG electrodes to the scalp along with gel to help conduct the signal through the skull, a process that is currently more time-consuming and potentially uncomfortable in comparison to other biometric measures. As time progresses, however, new EEG designs may allow for faster and easier data collection (Revett, Deravi, & Sirlantzis, 2010), such as having electrodes that do not need a gel to apply and using a helmet to easily place the device on one's head. Because of this, it is important to start exploring the security around using EEGs for identification and authentication so that high ease-of-use devices can be designed with security in mind.

Research exploring the usefulness of EEG signals in subject identification is currently focused on improving both data collection and data classification. It is worth noting that while the firing rate of neurons can reach 200 Hz, the most useful data falls within the range of 0.5 to 50 Hz, usually divided into four or five bands: the delta band (0.5-4 Hz), the theta band (4-8 Hz), the alpha band (8-13 Hz), the beta band (13-30 Hz), and the gamma band (above 30 Hz) (Campisi & Rocca, 2014). Please note that the ranges of these bands are not universally accepted, and that while the limits of the ranges are similar between interpretations, differences may be seen in various references (particularly in regard to the gamma band, which is sometimes not included and the frequencies of which are categorized as the beta band). In one of the earlier papers looking at subject identification via EEG waves in 1999, Poulos et al. used Kohonen's Linear Vector Quantizer, a neural net, to classify the alpha band (here 7.5-12.5 Hz) of the Fourier Transform of EEG data. For the four subjects in this study, the classification resulted in an accuracy of 72-84%, well above the guessing rate of 25%. Though commonly data from all bands is used in classification research, the gamma band (here 30-50 Hz) has also been used to classify subjects, due to its relation to brain functions such as thought and memory, as well as its prior usage to classify alcoholism in subjects (Palaniappan, 2004).

Though an example of how bands might be used in classification, Palaniappan's 2004 study may be considered more notable for its inclusion of visual evoked potentials (VEPs) in EEG subject identification. VEPs are defined as the potentials produced by the brain when looking at certain visual stimuli – for example, pictures from the Snodgrass and Vanderwart picture set, as was used in the aforementioned paper. VEPs have been used as analytical tools for many other types of classification such as glaucoma and dementia, and are used to elicit a stronger response from the brain than might be seen in a brain at rest (Watts, Good, & O'Neill, 1989; Wright, Harding, & Orwin, 1984). For Palaniappan (2004), the classification of the 20 subjects ultimately had an accuracy of 99%, using a neural network to classify. Palaniappan ran a similar study in 2007, with improvements including using the same VEP

stimuli but extending the range of the band used, using a different filter, and using the Davies-Boulder index to reduce the feature set. For a varying number of nodes in the hidden layer of the neural net used to classify, classification accuracy reached 98%. In 2009, Das et al. also explored using VEPs to classify subjects. Unlike Palaniappan's work, Das et al. didn't separate the EEG waves into different bands, and used two different machine learning techniques (Support Vector Machine and Linear Discriminant Analysis) to classify the data from 20 subjects. This study resulted in classification accuracies of 91-94% (for different numbers of samples in the training set: 2, 4, 5, and 10). The researchers also split the EEG data into pre- and post-stimulus data to examine whether important data occurred before the visual stimulus, and found that post-stimulus data led to much higher accuracy.

Though VEPs might be an important step in biometric identification, the usage of VEPs as a stimulus reduces the universalness of EEGs, as not all people have sight. Many studies also focus on EEG data that does not have a visual stimulus, often asking subjects to attempt to clear their minds, and differentiating between data taken while subjects have their eyes open and data from when they have eyes closed. The previously mentioned Poulos et al. (1999) was a study that had EEG recordings taken from both eyes open and eyes closed. Abdullah et al. (2010) also looked at EEG signals for 10 subjects, with a correct recognition rate of 78% for eyes closed, and 81% for eyes open, using four of the eight channels of the eight electrode EEG for classification. Despite this, the study reported no significant difference between classification for eyes open and eyes closed. Subjects in a study done by Paranjape et al. (2001) were also asked to remain at rest while EEG data was collected, both with eyes open and closed. The study looked at classifying between five and 40 subjects using autoregressive models of various degrees while eyes were open, and at the highest order reaching 99-100% accuracy for each set. When the data was split into a training and testing set, the accuracy of the testing set (for the highest order) was 82% for all 40 subjects. The results of these studies indicate that the resting state of subjects does provide enough information through EEG for classification accuracy above guessing (whether with

eyes open or eyes closed), though results may vary depending on other factors that will be mentioned later.

Further research into how EEG data is collected goes beyond VEPs and resting mental state, to other means of inciting specific or unique patterns. Such stimuli may include performing or imagining specific movements, of hands, feet, or tongue, or generation of words beginning with the same letter (Bao, Wang, Hu, 2009; Marcel & Millán, 2007 respectively). For the study looking at movement of both hands, feet, and tongue, the identification rate was between 81% to 90%, with the imagined tongue movement having the highest accuracy at 90%, while the other three options averaged to around 82% (Bao, Wang, Hu, 2009). Another study examined the difference between imagined movements of the left and right hands, as well as generation of words starting with the same letter, the results of which were half total error rates (HTER; false acceptance rate plus false rejection rate divided by 2) ranging from 6.6-26.1, with imagined left hand movements having the average lowest HTER (Marcel & Millán, 2007). These and other studies such as a 2019 study by Craik, He, and Contreras-Vidal show that there are a multitude of possibilities for EEG classification beyond VEPs and resting state which can still be explored and may prove to offer additional information for better classification.

Once the stimulation of potentials is determined, there exist additional ways in which data collection for EEGs can differ. The number of electrodes, or channels, on an EEG can differ from four to 256 electrodes, and of those any number can be used for the feature collection and classification. Eight electrode and 64 electrode EEGs are commonly used, though some of the studies here also look at classification using only one (Gui, Jin, & Xu, 2014) or two channels (Riera et al., 2007). As Gui, Jin, and Xu had accuracies between 70 and 99%, and Riera et al. had accuracies between 87 and 98%, these studies would seem to evidence that classification can be done for a limited number of channels. These results are comparable that of EEGs with six to eight channels, for which the accuracies of the studies mentioned here were within the range of 81% to 90% at their best showings (Abdullah et al, 2010, Bao,

Wang, and Hu, 2009, Paranjape et al., 2001). The 64 channels also have similar, though often higher-performing, results - between 94 and 99% for certain studies (Palaniappan, 2004, Palaniappan & Mandic, 2007, Das et al., 2009). Though these results imply that 64 channel EEGs produce more consistently higher accuracy, having smaller EEG options with comparable results is important as EEGs become more widely used, as EEGs with fewer electrodes are both cheaper and quicker to apply to a subject. Fortunately, while electrode number may be a factor in EEG classification, the standardization of electrode placement using the 10-20 electrode system means that electrode positioning is not (Jasper, 1958).

Following data collection, there are still more differences that remain in EEG subject classification; namely, how features are collected from the raw EEG data (if they are at all), and what system is used for classification. With regards to the former, both wavelet packet decomposition and Fourier transforms are commonly used to transform data, both of which offer easier access to the four to five bands that brain activity is divided into. The spectral power of the gamma band has also been used (Palaniappan, 2004), and the raw data could also be potentially used, though processing will still need to be done to remove blinks (if the subject had their eyes open) or other movements. With regard to the means of classification, a multitude of classifiers have been used in studies, but the most common classifier is the neural net, which in itself has a multitude of different variables. The number of nodes in the hidden layer is one such variable. Though both of those considerations (classifier multiplicity and settings) have additional sources of variability, the described variabilities already provide evidence as to how large a field EEG subject classification is and how much research remains. Wavelet packet decomposition and Fourier transforms are common, but not the only way to collect features, and neural net architecture is itself an important area of research. As using EEGs for biometrics becomes more viable, it becomes important to consider the security of such devices while they develop, rather than trying to insert security into a more complete product.

In order to further understand security with regards to EEG classification, it is worth exploring the difference between identification and authentication. Many of the studies explored up to this point have been focused on identification; namely, using the training data to determine what person the testing data came from. In comparison, authentication is a classification of whether or not the data came from someone in the authentication group. The security of authentication is applicable to multiple scenarios, the most common of which is the security of brain-computer interfaces (BCIs) (Huan & Palaniappan, 2004). Though BCI is a term used to refer to many interfaces with different functions, the basic definition of a BCI is a machine that obtains information from the brain of a participant (usually using an EEG), and then uses that information to determine some function of the interface (Wolpaw et al., 2000). In particular, for machines that are meant to offer aids to their user, either in the form of mobility or some other way, confirming that the person using the machine is the person intended to be the user is a requirement to prevent misuse. For this reason, certain studies have focused on authentication over identification. There are still issues with using EEGs for authentication, such as the time required to put the device on and the necessity of using conductive gel, which both costs money and may be uncomfortable and therefore undesirable to the user, though the possibility of "dry" electrodes and helmets made for easier application could alleviate some of those issues (Revett, Deravi, & Sirlantzis, 2010). The application of EEGs for authentication continues to be studied, however. Studies that focus on authentication over identification include Marcel and Millán, who in 2007 focused on authentication using EEG data, as mentioned previously, as well as Riera et al., also in 2007, and Gui, Jin, and Xu in 2014. Not all of these studies use accuracy percentage as a measure of success, but they do all show results that classify greater than guessing point, as was seen earlier.

While the studies examined thus far contribute heavily to the potential use of EEGs in biometrics, there are several common points of weakness between them. Sample size, as is often the case, is usually quite small in these studies, the largest of those mentioned here being 40 subjects. This

is understandable, due to the difficulty there often is in recruiting subjects for studies, but nevertheless reduces the generalizability of these studies (though the number of various studies on identification and authentication that occur somewhat makes up for this). Most of these studies also have the setback of using data taken from only one session, meaning that even though the studies show that EEG data can be used to classify people, there is less evidence that this will remain so over time. However, due to the at least partial genetic element of EEG brain waves, it is likely that they will stay stable over time, barring any significant brain damage (Näpflin, Wildi, & Sarnthein, 2007). It has also been noted that EEG waves have a heritable element to them, which may make identification between two family members, or in particular identical twins, more difficult (Poulos et al., 1999; Mohammadi et al., 2006). However, other biometrics such as fingerprints also have a heritable element to them and are still considered to be a successful biometric; fingerprint identification has also been achieved between identical twins, though not at as high an accuracy as non-related persons (Jain, Prabhakar, & Pankanti, 2002). Determining how many differences exist between people that make them more or less classifiable is also worth studying and is less likely to be mentioned in studies on identification or authentication. Research tends to focus on the overall results of a set.

For all studies mentioned here, the accuracy or other measurement of classification of the set being classified has occurred more than the accuracy of single people in that set. Also of note is the fact that none of the studies had any mention of the possibility of overtraining their nets, meaning that the net can identify certain traits or samples so well that a new sample introduced may be classified incorrectly due to the lack of that trait just by chance. Many of the accuracies of these studies were quite high; this can be indicative of a good classifier, but it can also be indicative of a classifier that has been overtrained. This is related to the security of a biometric in a couple of ways: first, that a biometric that is supposed to authenticate a person will not, but also second, that the biometric will then be more

likely to not authenticate a person who shouldn't be. Balancing these could be key in determining the security of an EEG based biometric system.

Part 2. Neural Network Pruning and Security

As one might imagine from their name, neural networks are networks designed to work in a similar way as the brain, with "neurons" that connect to each other and can cause each other to activate or deactivate through their own activation and strength of connection. Because of this, it may not be surprising that research tactics used in neuroscience are also applied to neural networks. In particular, the method of ablation is worth discussing due to how it can be applied to neural networks in ways different from how it is applied in neuroscience. Ablation studies of the brain are used to determine the function of the section of the brain being ablated from the rest of it, by determining what function is changed, reduced, or missing when it is gone (Reale, Brugge, & Chan, 1987; Kanold et al., 2003). The same principle can be applied to neural networks, which would then use ablation of specific neurons or sometimes specific layers to determine what those features contribute to classification – that of either a single class, or the overall accuracy, as some units will affect one or the other (Meyes et al., 2019). Though this is important to understand neural networks, ablation can also be used to improve the efficiency of a network. This is done by selecting and removing neurons from a trained network, in combination with training again afterwards, in order to retain its accuracy while reducing the power needed to compute. This approach is also known as pruning.

The concept behind pruning a network assumes that some of the neurons created are redundant and can be removed without greatly affecting the accuracy of the network. It is more difficult, however, to determine which neurons are redundant, and so there are a variable number of ways to determine the important of a neuron. One such way is to look at the magnitude of the weights of the neurons after training, as in theory weights with a larger magnitude will exert a greater effect on

the neurons that they act upon. In 2015, Han et al. used this method along with retraining the network post-pruning to great effect, maintaining similar accuracies even after reducing the number of weights by 9-12 times across various neural nets and image sets. Pruning has been continuously used in research, often with variations added to neuron selection for ablation to attempt further improvements to accuracy or other aspects of the net. For instance, one method removes single neurons from a layer at a time, managing to prune more than 85% of neurons from a network with originally over one million neurons while still maintaining accuracy (Babaeizadeh, Smaragdis, & Campbell, 2016). Other alterations include a means of reducing need for computational power via trained quantization, wherein neurons share weights, or with Huffman coding for further compression (Han, Mao, & Dally, 2015). Other features of a neural network that have been pruned include kernels and channels (Li et al., 2016; Zhao et al., 2019). The effect of applying low magnitude pruning to kernels can be seen prominently in certain studies, with accuracy being measurably reduced by only 0-1.5% (and sometimes increasing) following pruning approximately 47% of kernels for three nets and four datasets, of which one is text-based and the others are image-based (Wu et al., 2019). Beyond removal of neurons with the lowest weights from a network, there has also been extensive research into other ways of determining a neuron's worth or lack thereof. One study measures the importance of neurons by change in loss after removing a kernel, approximated via Taylor series expansion (Molchanov et al., 2019). Another focuses on pruning weights by layer based on the energy consumption they require, so as to further reduce the need for computational power (Yang, Chen, & Sze, 2017). Methods for determining neural importance are numerous, and while there are a multitude of possibilities beyond those mentioned (testing for redundancy, removing neurons whose removal have a low impact on accuracy, etc.), it is worth noting for the sake of this thesis that removing neurons with the lowest magnitude weights is both easy to implement and continues to be a competitive pruning method for retaining neural network accuracy with higher efficiency (Vadera & Ameen, 2022).

While the benefits of neural network pruning are plentiful, there are concerns that are crucial to consider with smaller neural networks. Cybersecurity concerns, and the question of whether or not a neural network can remain robust to attacks once it has been pruned to a smaller size, is paramount if pruning is to be applied as a security measure for authentication. Many studies surrounding this question look at adversarial robustness, or the ability of a neural network to stand against adversarial attacks, wherein small perturbations are added to inputs in order to disrupt the classification of them. In 2019, Sehwag et al. discussed the effects that pruning had on adversarial robustness, finding that low pruning ratios in networks that had been pruned but not retrained did not alter robustness a large amount, but higher pruning ratios did. It was also seen that robustness could be preserved with retraining of a network, in a similar way to accuracy. Further studies explore the ways in which pruning can be implemented in a network for adversarial robustness, and continue to show that robustness can be maintained after the pruning process (Gui et al., 2019; Ye et al., 2019). It has been noted that adversarial training requires having a larger net to begin with, both for improving robustness and accuracy, and thus pruning becomes more important for retaining both those features while maintaining a practical level of computational power (Ye et al., 2019; Sehwag et al, 2020). One study in 2018 by Frankle and Carbin suggests that pruning might not be a necessary step if the architecture that a pruned net produces could be found immediately (and thus a smaller net could be trained to the same levels as a larger net). However, it only examines the accuracy of the network, suggesting that beginning with a larger net would be more worthwhile for training for security as well as accuracy. Because of this, it remains important to explore the ways in which pruning could be concurrent with, or even beneficial to, security in neural networks.

In order to explore how security can mesh with networks being used for authentication, and in particular for authentication using EEG signals, one must first understand how pruning has been used in EEG research to date. For the most part, pruning has been utilized in the same ways as have been

previously discussed. In 2021, Vishnupriya et al. researched pruning for efficiency in the classification of motor imagery. The use of magnitude-based pruning proved effective, reducing the network to 90% of its original size. Similarly, pruning was shown to be promising in reducing energy usage for networks used to predict seizures, removing up to 75% of the nodes while retaining comparable accuracy to the baseline (Zhao, Yang, & Sawan, 2021). Duggal et al. (2020) show that pruning can be used in EEG sleep monitoring while maintaining adversarial robustness, affirming the same results seen previously for image classification. In addition, pruning has also been used as a method of discerning what modalities contribute to classification (of the three that were examined in the study: EEG, electrooculogram, and electromyogram), though this has less bearing on security and the study presented in this thesis (Ellis et al., 2021). While these studies extend the ways in which network pruning has been adapted to various networks and datasets, none of them explicitly look at subject identification or authentication. Indeed, it is difficult to find any work that looks at any biometric identification or authentication and neural network pruning. Though this may be because networks for the purpose of identification and authentication aren't usually large enough to warrant methods of compression, it also signifies potential for research in the field.

The way in which security is to be applied in this study does not involve adversarial training or robustness, despite the abundance of studies surrounding those techniques. Instead, the idea of neural network saturation will be introduced, and methods by which it can be applied to cybersecurity will be discussed. Neural network saturation is here defined as a point in which new changes to inputs fail to change the values of neuron weights, thus suggesting that the network will not train any more. Network saturation is generally considered a bad thing for a network, similar to overtraining, in which the network does not accommodate for new inputs that are not sufficiently similar to previous inputs, causing them to be misclassified. For instance, saturation has been shown to negatively affect networks using particle swarm optimization as a training algorithm (Rakitianskaia & Engelbrecht, 2015). It might

seem peculiar to introduce such an idea in the context of contributing to cybersecurity, but it is worth exploring for a few reasons. Many of the studies examining the effect of pruning on neural network efficiency fail to include insight on whether or not the pruning contributes to overtraining, focusing instead on the results for the given dataset. Since pruning is intended to remove unnecessary weights, and low magnitude pruning in particular gives importance to weights more strongly associated with specific pathways, it seems likely that pruning would more quickly lead to network saturation. It is also worth exploring network saturation as a method of security simply to add to the selections that one can have on hand. The concept behind applying saturation as security is two-fold: first, it suggests that a network that is saturated can't be easily retrained to authenticate new inputs that one does not want to be authenticated, and second that, should a neural network be retrained for a new input, the accuracy of the network will change drastically enough that the tampering will be obvious. Whether or not these concepts hold true is the focus of the study herein.

With this, the parts of the study described here have all been stated. From chapter 1 it is seen that signals taken from EEG readings have been established as unique and useful biometrics that can be utilized in both identification and authentication. Various parts of data collection and processing that can influence the final product, including whether data was taken from resting potentials or evoked potentials, which classifier was used, what brainwave band is being used to classify, etc. EEG subject authentication is important for the usage of BCIs (brain-computer interfaces), and therefore requires security to prevent misuse. From this chapter it can be observed that neural network pruning is a method that has been used for efficiency and is compatible for security. Saturation has been discussed in light of its potential contribution to security. It is also noted that the security usually combined with pruning is adversarial training, such as is commonly used for image classification and large datasets. With this in mind, it is suggested here that neural network pruning can be applied to a network trained

for EEG subject authentication to bring the network to a state of saturation or near-saturation, such that retraining the network to authenticate an adversary will either be impossible or readily detectable.

METHODS

Part 1. Data Collection

Unless otherwise mentioned, all reported data are from the same dataset. This dataset is comprised of EEG data collected from a 64 channel EEG recording for 12 subjects, downloaded from PhysioNet on September 9, 2019 (Goldberger et al., 2003). Data was gathered from the subjects during conditions described in Citi et al., 2010. EEG data was gathered for 20-21 trials per subject in the original data set, from which 20 trials per subject were used for training and testing in order to balance the data. Assignments of the dataset to training and testing samples were all done using the random module in Python.

Part 2. Classifier Comparisons

Various methods of classification were explored before focusing on neural network approaches. Data from each channel of the EEG was transformed using a fast Fourier transform in Python, then normalized. For each subject, ten samples were randomly assigned to the training set, with the remainder being assigned to the testing set (half of the subjects had 21 total samples, meaning 11 samples were sorted into the testing set – this was corrected for classifiers that are discussed later, such that both training and testing had 10 samples for each subject). Values of each channel for the entire training set were averaged to create a reference mean for each channel. In order to classify the testing dataset, testing samples were transformed in the same way as the training samples, with the values of each channel then compared to the previously found reference means. The test samples were then classified based on which subject had the smallest sum (L1) difference in means across all channels.

A classifier was also created that made use of a computed Hamming distance to classify subjects. Training and testing samples were randomly assigned, as described above, with 10 trials from

each of the 12 subjects assigned to either group. In this classifier, the raw EEG time series was divided into 500 sections (for example, a 2 second signal would be divided into 500 four millisecond long sections), which were assigned either a 1 or a 0 depending on whether the mean of the section was above (1) or below/equal (0) to the mean of the entire signal. Then, for each testing sample, the Hamming distance was found for each training sample of each subject. Hamming distance is a comparison of two binary strings wherein differences between the two strings are counted (so, if each string has the same number in the same position, the sum does not increase, but if there is a 0 in one string and 1 in the other, the sum increases by one; see Table 1). In this case, that means that every channel in the new testing sample is compared to the same channel from each of the training samples, and the Hamming distances are counted. The testing sample is then classified to the subject that it has the lowest over Hamming distance with.

Table 1. A demonstration of the Hamming distance between two arbitrarily chosen binary strings.

Sequence 1	0	1	1	0	1	Hamming Distance:
Sequence 2	0	0	1	1	1	
	+ 0	**+ 1**	**+ 0**	**+ 1**	**+ 0**	**= 2**

The next classifier to be compared is one that uses a method similar to Hamming distance. However, in this case, the EEG signal is not converted to a binary number but is instead converted to a nonary number. In this classifier, the range of the amplitude of the EEG signal is divided into nine parts, and the 500 sections of the data are assigned a number one through nine based on where in the nine parts of the range the number falls. A modified Hamming distance is calculated the same as in the previous classifier, except with a value between 0 and 8 assigned for the distance, calculated by the absolute value of the difference between the values, and the subjects are classified in the same way.

To compare the accuracy of the aforementioned classifiers to that of a neural net, features first had to be collected from the EEG data. Data was transformed using a Fast Fourier Transform (FFT) in

Python and was run through a low-pass filter of 40 Hz, by removing the FFT coefficient values above 40. It was then assigned into four signal ranges: delta waves (0-4 Hz), theta waves (4-8 Hz), alpha waves (8-14 Hz), and beta waves (14-40 Hz). The integral, mean, standard deviation, median, and magnitude of each of these sections of the FFT, as well as the magnitude of the entire signal, were collected as features for a total of 21 features per channel. These features from all EEG channels were then stored as the features for that subject for that trial. In total, for the 70 channels in each trial there were a total of 1,470 features per subject.

Part 3. Understanding the Neural Net

Features are not the only necessary component to set up a neural net. Each neural net can have multiple differences between them, including the number of hidden dense layers between the input and output layer, the number of nodes in those layers, the number of epochs used to train the net, the size of the kernel, the optimizer, and the number of filters. All of the networks used in this thesis will also have a single convolutional layer before any hidden dense layers, though the number of convolutional layers was not included in the feature set. In order to determine the optimal choice for each of these, an evolutionary algorithm was developed and utilized, as described below.

The purpose of using an evolutionary algorithm is to take advantage of the system by which biological creatures evolve, and thus, through multiple repetitions, determine a set of features that perform well for the neural network they are implemented into, while saving time by avoiding testing every possible permutation on its own. In an evolutionary algorithm, the features that perform well stick around, while those that perform poorly leave the population. Though measures of performance vary between applications, this here means that if a net achieves high classification accuracy, it has a higher chance of having its features “passed” to the next generation (here meaning copied into a new neural

net). Additionally, mutations and crossover events like those that might be observed in genetics are part of the generational survival process for the population of "genes" that are used in the algorithm.

For the purpose of this thesis, a population of 10 genes existed, where each gene consisted of a set of features randomly chosen from the predetermined possibilities for each feature. A convolutional neural network (CNN) was trained on the EEG data with the features from each gene in the population. The accuracy was recorded, then accuracies for the entire population were normalized for that population, giving each gene a percentage of "survival" based on how well their corresponding network performed. The range between zero and one was then divided into ten parts to reflect this, where each gene was assigned a range with a length of their chance of survival (so networks with lower accuracies would have smaller ranges, and those with larger would have larger ranges), and ten random numbers between zero and one were chosen. For each of those random numbers, the gene whose range the number fell within "survived" that generation and was copied into the next population, until a new population of 10 genes was recorded. Therefore, depending on performance and the random numbers chosen, some genes may not have been copied into the new population and some genes may have been copied more than once, with genes that performed better being more likely to be copied and more likely to be copied multiple times. The new population was then subjected to computational imitations of mutation and crossover events, which happened at a given frequency (in this case, each had a chance of happening 9% of the time). If a mutation occurred, the feature that "mutates" was replaced by a randomly chosen option for that feature. If a crossover event occurred, two members of the population exchanged one or two of their features. By the end of this process, there was a new population with 10 "genes" in it that was related to, but not exactly like the previous population, and the process was repeated. In general, genes that performed better were copied more often and eventually made up the majority of the next population. This process was designed to run either 250 times, or until the sum of the means of the final three runs equaled 2.5, thus giving a set of features that offered accuracy on

average above 83% without testing each permutation of features. Note that the neural networks were trained for a 12 class classification, rather than the binary classification that will be discussed in part 5 of the materials and methods.

The results of this algorithm offered a good idea of how to set up a neural network for classifying the EEG data, though it was not the final stopping place. From it, neural networks were set up with one hidden dense layer with 384 nodes in it. Epoch numbers from 50 to 100 showed up towards the latter stages of the algorithm; 60 was eventually chosen as the standard for future usage. Multiple optimizers also performed well. Of those that did, both "adamax" and "rmsprop" were used in future applications of the net. Batch number tended to range from 20-40 (in steps of five), resulting in 25 being chosen as the default option for future nets. Kernel numbers from the algorithm ranged from 175 to 400 (in steps of 25), but the majority were between 175 and 275. The number chosen as default was 250. Differences in filter number rarely affected the accuracy of the resulting nets; however, high filter numbers caused a noticeable increase in runtime of the neural net, so the evolutionary algorithm only permitted up to 12 features (from 1), and 12 filters were eventually chosen as the default.

Using the standards set for a neural net as noted above, the accuracy of subject classification was again tested using a neural net, with 10 trials (50%) per subject being used as training, and 10 trials (50%) per subject being used as testing. Due to the resulting accuracy of the neural net being much higher than any other tested classifier, the focus of study for EEG classification was shifted to solely include neural nets, and research mentioned further in this paper will no longer include results from any other classifier.

Part 4. Neural Network Input Sensitivity

To ensure that the neural net was performing as intended, a series of sensitivity tests were developed and utilized. New wave series data was created for this test, rather than using the previously

used data. Five tests were applied to binary classification and categorical classification. To create data for these tests, different classes were assigned different sinusoidal waves with prime number frequencies but the same amplitude to derive features from. For the binary classification, two classes were created, one with a frequency of 5 cycles per unit and the other with a frequency of 7 cycles per unit. For categorical classification there were a total of four classes, with respective frequencies of 3, 5, 7, and 11 cycles per unit. The binary classification classes copied those signals into samples to be split into training and testing; Test 1 had four samples in both training and testing, Test 2 and 3 had three samples in training and testing each, and Tests 4 and 5 had six samples in both training and testing. For all tests run on categorical classification, there were 10 samples in both training and testing. Features for use in the neural net were collected in the same way as described above for the EEG, but with differing band wave sections (here, the sections were 0-3, 3-5, 5-7, 7-11, and 11-13 cycles per unit). The net itself also differed in the number of layers (one convolutional for both binary and categorical, four hidden layers for binary and three for categorical, and one output layer for both) The tests are as follows:

Test 1: A very basic sensitivity test in which each the data above was not changed. For both binary and categorical classification, each class had a single unique frequency. This test was meant to test that the neural net was classifying properly, as it should easily achieve 100% accuracy with this data. An example of this test can be seen in Table 2.

Table 2. An example of the categorical version of Test 1. Each subject has a unique frequency to their wave series data.

	3 Hz	**5 Hz**	**7 Hz**	**11 Hz**	**13 Hz**
Person A	1.00				
Person B		1.00			
Person C			1.00		
Person D				1.00	

Test 2: Over multiple runs the strength of the original frequency decreased by 0.01 (1%) and the strength of the other frequencies increased by 0.01 (X in Table 3; see Table 3). For binary classification, this was only one other frequency, and for categorical it was three other frequencies. The first of several tests examining how a neural net responds to noise, this test would ideally show accuracies of or close to 100% during the first few runs when the original frequency is the strongest, but would have accuracy decreasing to guessing accuracy (50% for binary, 25% for categorical) when the strengths of all frequencies become closer together during the later runs.

Table 3. A categorical example of Test 2. X represents the increase of noise in steps of 0.01, from 0.00 to 0.25.

	3 Hz	**5 Hz**	**7 Hz**	**11 Hz**	**13 Hz**
Person A	1.00-3X	X	X	X	
Person B	X	1.00-3X	X	X	
Person C	X	X	1.00-3X	X	
Person D	X	X	X	1.00-3X	

Test 3: Here, each class had a consistent signal with their original frequency at lower amplitudes than as seen in Test 1: 0.50 amplitude for binary classification classes and 0.25 amplitude for categorical classes. They also, over multiple runs, had the other frequency(ies) added in variable strengths between 0.0 and 0.50 at random, with a mean of 0.25 between the three for the categorical classifier (see Table 4). This test ideally checked for the net's ability to classify based on a consistent signal, though it was not expected for all or even most neural nets to reach 100%.

Test 4: This test had a similar concept to Test 3 in that it added in other frequencies besides the original ones with varying strengths while the original frequency had a strength of 0.50 or 0.25 (binary or categorical respectively); however, in this test, the first class also had a new frequency (13 Hz) which increased in strength from 0.0 to 0.50 in steps of 0.01 over multiple runs (see Table 5). This test looked

Table 4. A categorical example of Test 3. Here the signals for the subjects remains consistent as noise increases. For categorical testing, X is chosen randomly (for each individual X in the table) from a distribution of 0.00 to 0.50, such that the X's of each row have a mean of 0.25.

	3 Hz	**5 Hz**	**7 Hz**	**11 Hz**	**13 Hz**
Person A	0.25	X	X	X	
Person B	X	0.25	X	X	
Person C	X	X	0.25	X	
Person D	X	X	X	0.25	

Table 5. A categorical example of Test 4. Here the first subject (Person A) has an extra frequency in their wave series data. X here, like Test 3, is chosen randomly from a distribution of 0.00 to 0.50, such that the X's of each row have a mean of 0.25.

	3 Hz	**5 Hz**	**7 Hz**	**11 Hz**	**13 Hz**
Person A	0.25	X	X	X	Y
Person B	X	0.25	X	X	
Person C	X	X	0.25	X	
Person D	X	X	X	0.25	

at the sensitivity of the net to a new frequency amidst noisier signals, and would ideally show an increase in accuracy as the new frequency grows stronger.

Test 5: A similar test to Test 4, but now with noise removed and with classes having the same original frequency (i.e. 5 cycles per unit), with one having a new frequency added with increasing strength over multiple runs (0.0 to 0.050 in steps of 0.001; see Table 6). While low strengths of the new frequency

Table 6. A categorical example of Test 5. Here all subjects have the same data excepting Person A, who has an additional frequency. Y increases from 0.000 to 0.050 in steps of 0.001.

	3 Hz	**5 Hz**	**7 Hz**	**11 Hz**	**13 Hz**
Person A	1.00				Y
Person B	1.00				
Person C	1.00				
Person D	1.00				

would mean the accuracy of the neural net would likely be close to guessing accuracy, as the new signal grows stronger the accuracy of the net would approach 100% for binary and 50% for categorical. This difference in accuracy is due to one class being consistently classified accurately and the other three only being able to be classified with guessing accuracy.

Part 5. Neural Net Weight Distribution and Entropy of Weight Distribution

Once the neural net had been shown to be operating as expected, the focus shifted towards how to combine knowledge of EEG signals and neural net classification for security and privacy applications. The idea of security through neural net saturation was proposed, wherein neurons would be removed from a trained network, followed by retraining the network, to ensure an inability to train it past the original application. To accomplish this, the weights of the neural net had to be explored. The weights of the neural net include the weights of the nodes of each layer as well as the weights of the connections between the nodes in each layer (edge weights).

To begin with, the ways the weights changed between training and testing was examined. Using the sensitivity tests from above as the basis for the neural net, the trained values of the weights were graphed against the initialized weight values and labeled with the accuracy of the net to discern if any pattern arose as accuracy changed. Furthermore, the initialized weight values of the neural net before training were subtracted from the weight values after they had been trained to compare the changes made between them, and again, see if any pattern arose.

Observation of the weights and their distribution proved difficult to quantify for comparison, so a new method of comparison was proposed, in which instead of simply looking at the difference between the weights from before and after training the entropy of those weights would be calculated and compared. Entropy was calculated using Equation 1, where $H(X)$ is the entropy of set X, P is the probability of an occurrence of weight value x_i, and n is the total number of values in set X. Since

entropy is a measurement of distribution, where a larger entropy indicates a larger distribution of values across their range and a smaller entropy indicates that values fell into certain ranges more often, the

$$H(X) = -\sum_{i=1}^{n} P(x_i) log P(x_i) \qquad \textit{Eq. 1}$$

comparison of the entropy of weights prior to and post training of the net would reveal any changes in distribution of those weights. This was done using the sensitivity tests as a basis, and the distribution of the weights was also sorted into a histogram with the number of the weights in each bin being measured. The entropies and differences in entropy were then compared to the accuracy of the net; again, looking for patterns such as entropy decreasing after training.

From here, the same concept above was applied to the classification of the original EEG data using neural nets. Heretofore, the entropy examined was the overall entropy of every node and edge in the neural net being used. For further understanding, this group was broken down to examine certain part of the structure of the neural net. Comparisons were made not only between the entirety of the pre-trained and post-trained weights, but also between the three layers outside of the input layer (convolutional, hidden dense, and output) as well as between the edges and the nodes. This revealed that the entropy statistics of the overall weights most closely resembled those of the edge weights, due to the exponential ratio between edges and nodes. It also revealed that the hidden dense layer showed the greatest change in entropy between training and testing, and that nodes showed greater change than edges.

Part 6. Ablation Studies

Due to graphs from the statistical tests indicating that weight distribution, and therefore entropy, was unlikely to be correlated with neural network performance, a new question was asked: how entropy could be used to indicate other useful information about a neural net. By combining the

study of the entropy of a net's weights and ablation studies with which to create a more efficient net, a research question formulated. Could a neural net be brought to a point of saturation via ablation of hidden layer nodes, here meaning that it is trained so efficiently with the given weights for a certain type of classification that it cannot be retrained for any additional information, and does entropy relate to this point of saturation?

In order to begin to explore this question, a baseline set of neural nets was needed, which would be trained with a variable number of nodes in the hidden dense layer (with no ablation) as a comparison to the ablated nets. In order to better apply this method to security applications as well, the neural net classification was switched from categorical classification (classifying each subject from their own data) to binary classification (having a single person be in the 'authentication group' and classifying them from the other eleven people). The number of nodes in the hidden dense layer included three and four, and then ranged from 5 to 245 in steps of five. For each of these neural nets, they were preliminarily trained for one subject (to represent the person that a system hypothetically should authenticate; hereafter referred to as the 'Target') and then later retrained to include one to five more people in the authentication group (for a total of six out of twelve people by the last retraining, representative of hypothetical attackers to a system; as such, these five additional people will be referred to as 'invaders' for the purpose of this paper). For each of these nets and each of these runs, overall accuracy was collected along with the accuracy of the first subject in the authentication group, the accuracy of the entire authentication group, the specificity, the entropy of the trained nodes, and the change in entropy of the weights of the edges both leading into and away from the hidden dense layer. For the purpose of figures in the results section, the accuracy was also adjusted (normalized) to better reflect the binary classification of the results, as seen in Equation 2, where A is the adjusted accuracy, T_P is the percent of correctly classified authentication group trials, and T_N is the percent of correctly classified out group trials. Because the authentication group had only 10 trials in it (without

invaders) and the out group had 110, assigning all trials to the out group would result in an accuracy of 91.67%, even though no authentication trials would be properly classified. After adjustment, the accuracy of the same classification would be 50%, or guessing accuracy. This method was then repeated again, classifying the first subject, and then twice for another four times for four new people from the subject set, so that five people each had two baseline sets of neural networks.

$$A = \frac{(T_P + T_N)}{2}$$

Eq. 2

With a comparison set established, the same methods were then repeated for ablation, with the exception that the smaller nets than the 245 were created by ablating the 245 node network (as described ahead). Each ablation step consisted of removing five nodes with the smallest magnitude of weights from the hidden dense layer of the neural network (or one node, once the neural net reached size 5, to match the baseline nets), while the other nodes retained the weights they had after training (a simplified depiction of ablation can be seen in Figure 1). After removal, the neural net was retrained for the singular subject in the authentication group, and measurements were taken again. All measurements mentioned above were collected. Similarly to above, invaders were added into the authentication group at each step after retraining, and measurements collected. The nets retrained to add invaders into the authentication group were not then the networks ablated – only the networks trained on one subject in the authentication group had nodes ablated. Furthermore, only nodes from the hidden dense layer were ablated, as entropy studies from the previous section indicated that this is where the most processing occurs, and as the number of nodes in the input and output layers are determined by the shape of the input and the number of classes being classified respectively. The two processes for baseline and ablated networks were repeated twice for each of five randomly chosen targets, and the results from the ablated set was then compared to the baseline set for each target.

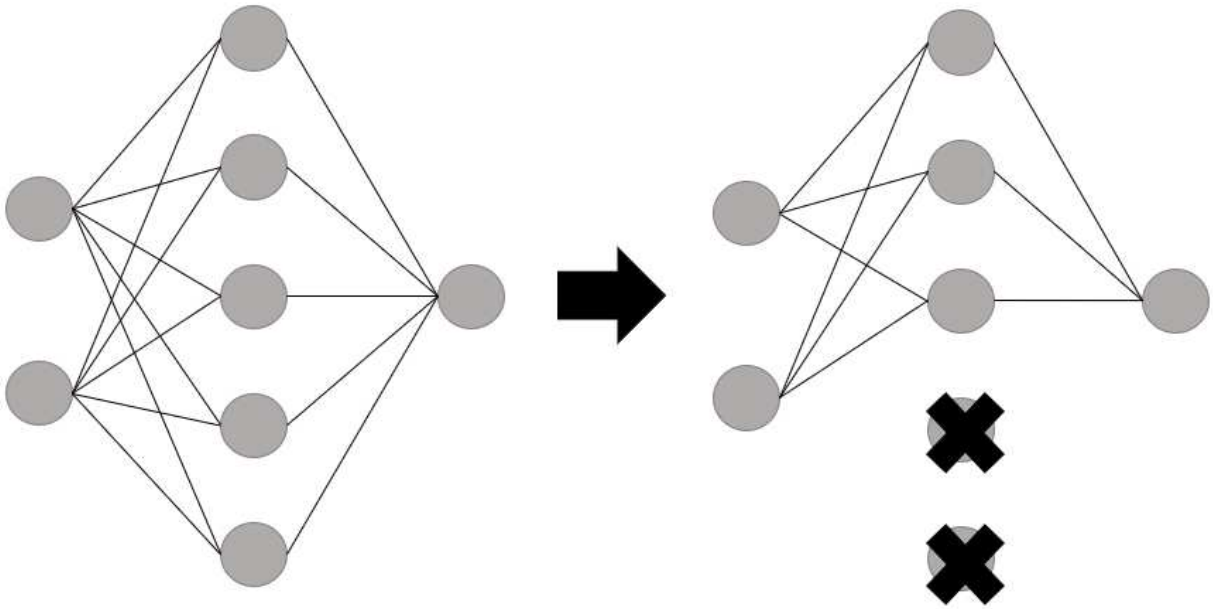

Figure 1. A simplified depiction of a neural network being ablated. The network on the left is a fully connected neural network with one hidden dense layer. The network on the right is the same network, but two nodes from the hidden dense layer have been ablated (removed from the network, along with the edges connecting them to the other layers).

RESULTS

Since not every part in Methods produced results, the numbering of parts in the Results will not directly match Methods section. However, the corresponding Methods section will be noted in each section of this Results section for reader clarity.

Part 1. Classifier Comparisons (Methods Part 2)

For the classifier that used Fourier transforms and compared the means of each channel to determine which subject a test sample belonged to, the classifier resulted in an accuracy of 83.2%. This accuracy is much larger than the guessing accuracy of 8.33%. However, due to later high performance and greater applicability of neural networks, this method was not explored for future questions.

The Hamming classifier (using a binary comparison) resulted in a classification accuracy of 36.8% (see Figure 2a). While higher than guessing accuracy (8.33%), this accuracy would likely not suffice for security applications. The adjusted Hamming classifier (using a nonary comparison) resulted in an accuracy of 35.0%, a lower accuracy than the original Hamming classifier. Figure 2b shows the comparison of actual to predicted class. While certain subjects were more accurately identified using the adjusted Hamming classifier (namely subjects 3, 9, and 11), 42 of the 120 total test samples were classified as subject 3 and 40 were classified as subject 9, indicating that the success of the classification of those subjects was likely not due to an ability of the classifier to distinguish their samples from other samples.

Part 2. Evolutionary Algorithm (Methods Part 3)

The parameters of a neural net have already been discussed in Methods. A neural net using these parameters resulted in an accuracy of 95.8% when classifying the EEG data categorically – much

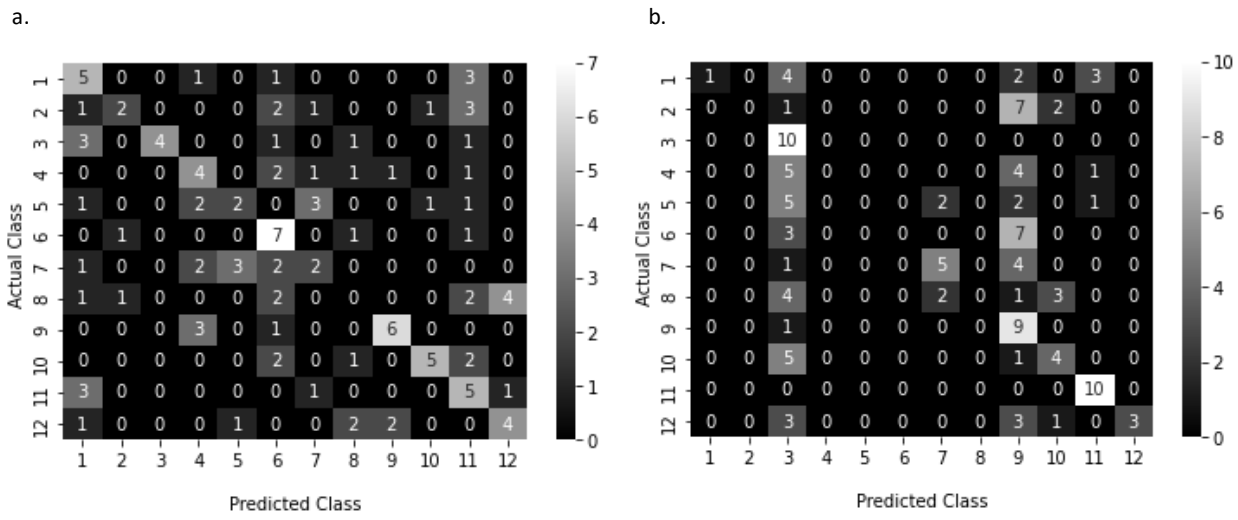

Figure 2. a. The confusion matrix for Hamming distance between binary string representations of the EEG signals from the 12 subjects being classified. Subject 6 was most accurately classified with 7/10 samples being correctly classified; however, many subjects were not classified well, including subject 8 which had no test samples correctly classified. b. The confusion matrix for the modified Hamming distance classifier. The majority of test samples were classified as either subject 3 or 9, while subjects 2, 4, 5, 6, and 8 had no samples classified as belonging to them at all.

higher than the 8.33% that is guessing accuracy, and higher than the percentages of either Hamming classifier. For this reason, neural networks became the focus of the research for the rest of the thesis.

Part 3. Neural Network Input Sensitivity (Methods Part 4)

In order to ensure that the neural networks used were operating properly, a series of input sensitivity tests were performed as described in the Methods. While the results of those tests are discussed here, the primary purpose of these tests were for calibration, and as such there was no expectation of novel discovery from them.

Binary Sensitivity Testing:

The results of the first sensitivity test applied to a binary classification neural net showed an accuracy of 100%, with five out of five samples being correctly classified. This is as expected, and

indicates that the neural net is operating properly. Were the accuracy to be less than 100% for this first test, this would indicate that the net cannot reliably be used to classify wavelengths, as it would be mistaking what should be two readily distinguishable inputs.

The second sensitivity test revealed that minor perturbations to the input wavelengths could obscure classification for a binary class, and also that obscurity becomes more likely the more noise added to the signals, as was expected. Figure 3a shows the mean accuracy over the course of ten runs of test 2. The pattern mentioned previously can be observed here, with accuracy remaining consistently high until added noise eventually reduces it to the expected results for guessing (here, 50.0%). Accuracy was reduced because the original frequency was reduced as noise was added, to the point where noise equaled the original frequency at strength 0.50.

Sensitivity test 3 varied from test 2 in that the constant signal for each class was overall weaker, and noise followed no pattern and occasionally was as strong as the original signal. This resulted in accuracies ranging between less than 40.0% and 100% and followed no discernable pattern. Though one of the classification results was below the guessing point, and one was at the guessing point, many of the accuracies of the 20 different as seen in Figure 3b were above the guessing point, suggesting that the constant signal may have contributed to overall classification. However, this test indicates that there can be no guarantee of good binary classification for noise that is similar to noise seen in other subjects.

The usage of a new signal in sensitivity test 4 contributed to overall higher accuracies, with accuracies falling between 68.0-100%, as seen in Figure 3c. The addition of a unique frequency helped with classification, as all except one of the accuracies averaged to be 80.0% or above. Despite the overall increase in accuracy, however, the neural net never achieves consistent results of 100% accuracy, suggesting that it is not solely dependent on the additional signal but instead also relies on the original signal and the noise, and thus may still misclassify samples.

The fifth sensitivity test was set to confirm that the presence of an entirely new signal contributed to the accuracy. Test 5 shows that, though confusion can remain while the strength of the new signal is small, as the signal gets stronger, the accuracy remains at 100%. A graph showing the mean accuracies taken over ten runs can be seen in Figure 3d. This test reveals that the neural net, in the absence of other differences between subjects, can be very sensitive to the addition of a unique frequency.

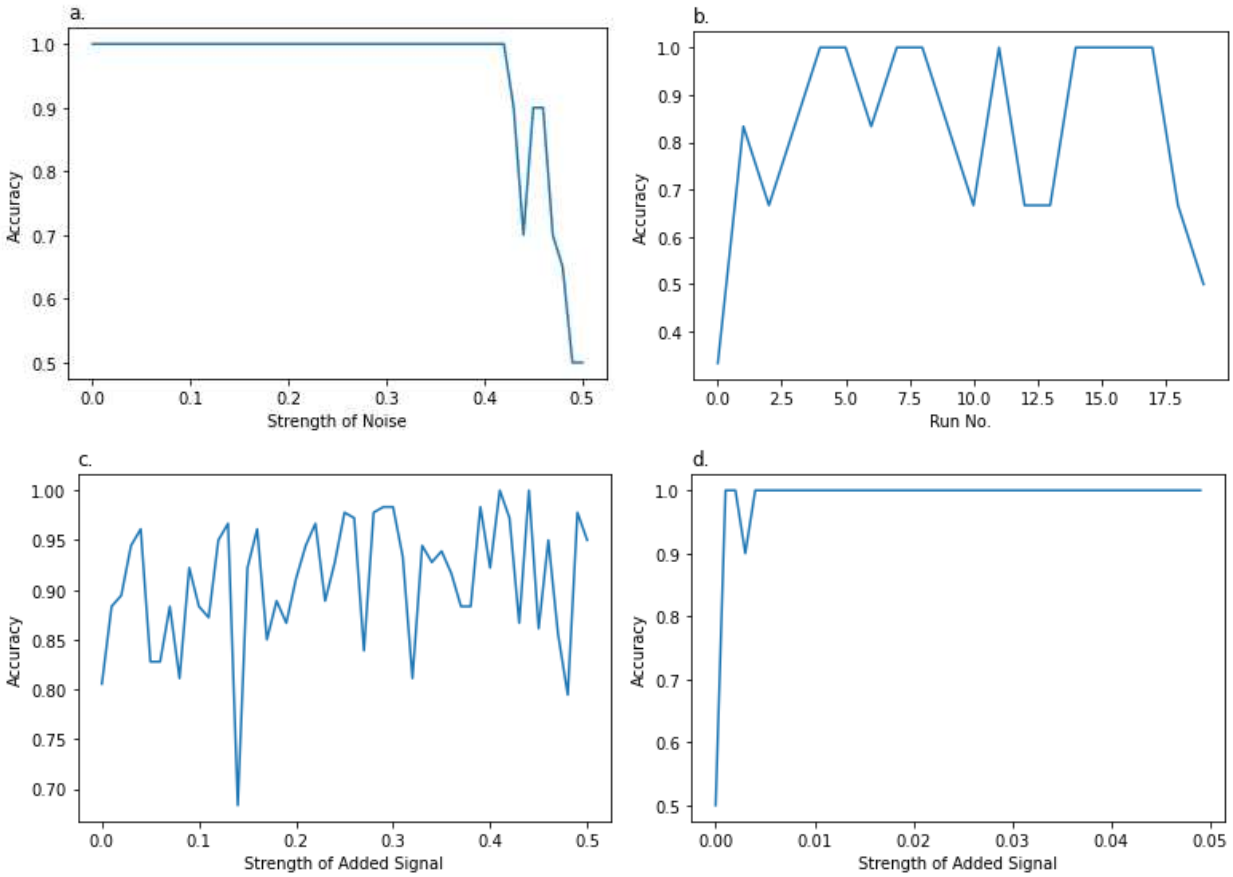

Figure 3. a. The results of Test 1 applied to a binary classification set, as averaged over 10 runs of the test. b. A set of accuracies from 20 runs of Test 2, as applied to the binary classification set. Due to the x-axis variable not being the independent variable, these were not averaged. c. The results of Test 3 applied to the binary classification set, as averaged over 10 runs. d. The results of Test 4 applied to the binary classification set, as averaged over 10 runs. Mean of the runs is shown at each value along the x-axis.

Categorical Sensitivity Testing:

Like the binary classification, categorical classification successfully proved that the neural net worked as expected by achieving a 100% accuracy for the first sensitivity test for each of 10 separate runs. Having shown that the neural net is unlikely to classify subjects based solely on random criteria, the neural net then proceeded to test 2.

In test 2, the expected outcome was, as before, that accuracy should vary inversely with the amount of noise added, eventually reaching the expected accuracy of guessing (here, 25.0%). This is what was seen in the results, which were averaged over ten runs and can be viewed in Figure 4a. The average of these results shows that accuracy drops as the strength of the noise reaches roughly 0.23, or 23.0% of the total strength of the frequency. As the strength of the original signal reaches 0.25 when noise does, this is the point when the net should not be able to classify higher than guessing point, a phenomenon which is observed in Figure 4a. This test denotes that the neural net can use even subtle differences between EEG signals to classify them, given that the signals consistently contain those differences.

Test 3 had a large variance of accuracies, the range of which extended from below 25% to above 55%. The results of 20 different runs of Test 3 can be seen in Figure 4b. The range of accuracies suggests that, despite noise of random strength being added to the signals, the consistency of the original frequency might occasionally provide a means of classification, given many of these runs resulted in accuracies above the guessing point. However, there was no guarantee of this, showing that occasionally multiples of the same frequency can confuse the neural net, which was the expected outcome. As such, this test gave an example of what to expect when the noise within a signal is too large or overlaps too much with common signals from other subjects.

This knowledge of how a neural net reacts to noise is then challenged in test 4, where one of the classes gains an additional frequency in their signals. As can be seen in Figure 4c, the general trend of

this test is that the stronger the new signal, the higher the accuracy of the classification, though the trend appears to plateau when the strength of the signal hits about 0.30. This is likely due to the fact that, since only a single subject receives the new unique signal, the accuracy can only rise so long as samples of that subject are not being classified correctly. Once they are all correctly classified, the plateau occurs, with variation in the accuracies occurring due to variation in the noise. This test, along with the following test 5, indicate that the neural net is capable of discerning unusual signals amongst signals that are otherwise difficult to distinguish.

As mentioned, test 5 covers a similar idea to test 4, and the trend line of it (as seen in Figure 4d) is more pronounced. Again the accuracy starts at 25.0% when there is no extra signal, as would be the most likely outcome for four identical classes, but the accuracy quickly reaches a level close to a plateau at 50.0%. Unlike the trend graph for test 4, for which the line of best fit would have a more gradual increase, test 6 seems to follow a more exponential growth, showing how the noise from test 4 can still interfere with new signals that would be distinct in a test 6 environment. Despite this, test 5 again contributes to evidence supporting the neural net's capabilities of distinguishing unique features of a class, even in small amounts.

Combined, both binary and categorical classification show that the neural net that is being used to classify subjects in later sections works as intended. The neural net easily distinguishes between different signals and cannot distinguish between largely overlapping noise, unless there is a unique signal also contributing to classification. As the frequencies of EEG signals all tend to fall in the same range, this is important to note for EEG classification of subjects as well.

Part 4. Neural Network Weight Distribution and Entropy (Methods Part 5)

The comparison of the initial weights to the trained weights for Test 2 suggests that as the test progresses, the changes in weight distribution become more drastic. Figure 5 shows two examples from

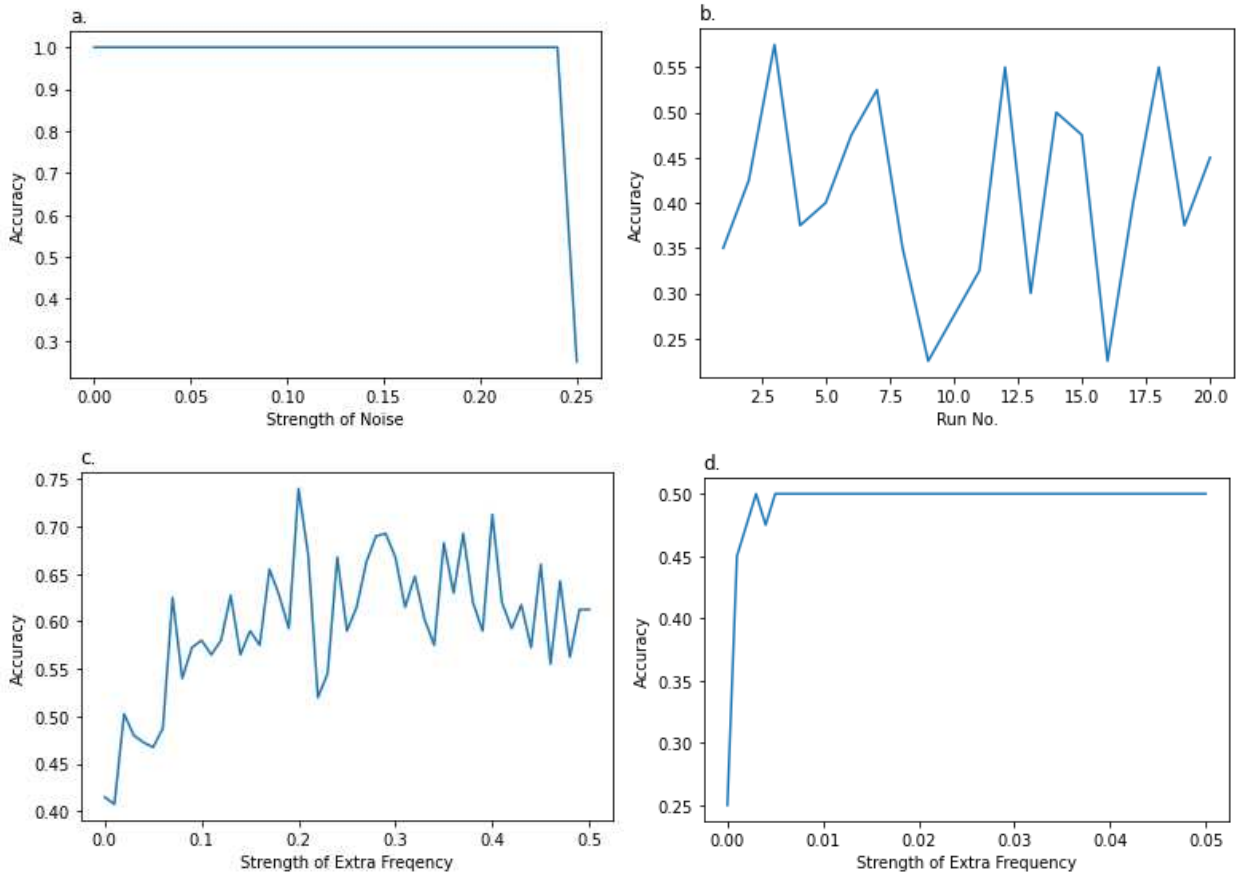

Figure 4. a. Accuracy of Test 2 applied to a categorical classification set. b. Accuracy of Test 3 applied to the same set. Note that the units of the x-axis are arbitrary, so no obvious pattern should be discerned here. c. Results of Test 4 applied to the categorical classification set. d. Results of Test 5 applied to the categorical classification set.

Test 2, with 5a and 5b being the results of the first step, wherein noise is zero (and accuracy is 100%), and 5c and 5d being graphs from the final step, wherein noise is 0.25 (and accuracy is 25.0%). All four graphs indicate that the overall range of weights is -0.3 to 0.3, and 5b and 5d show that the weights that change the most initially fall between roughly -0.16 and 0.16. However, the comparison between 5b and 5d reveals that, as noise is added and accuracy reduces, the maximum change in weights increases by a very large magnitude (particularly of a small number of weights that began at zero). Though only the first and last step of the test have been provided for reference, this concept holds mostly true for the steps in-between, with some exceptions where noisier steps may have smaller ranges of change in

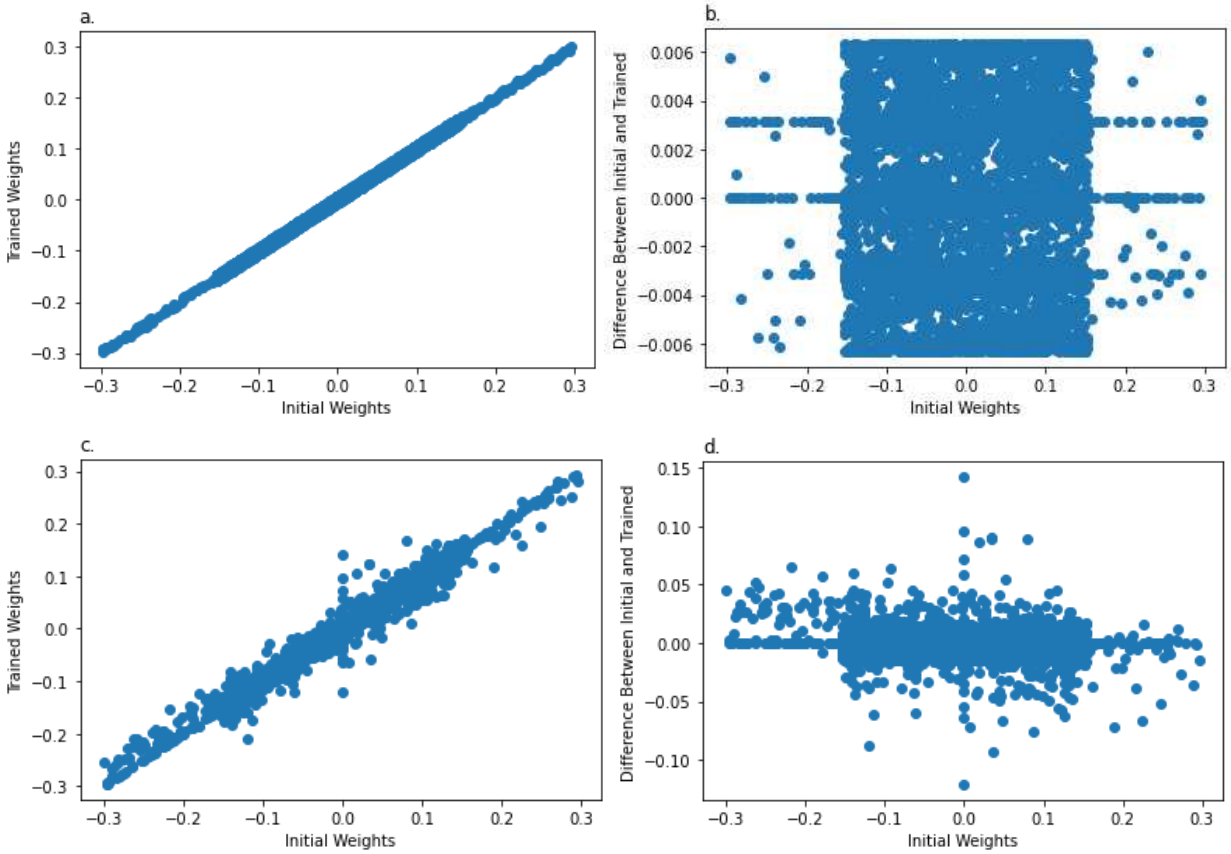

Figure 5. a. A comparison of weight values from a Test 2 neural network prior to testing to post-testing, where no noise was added and accuracy was 100%. b. A comparison of weights prior to testing and the difference in their values after testing, from the same net as 10a. c. A comparison of weight values pre- to post- testing for a net with noise added to the point where classes were indistinguishable (resulting in a 25% accuracy). d. A graph of how much initial weight values changed for the same net as 5c.

weight than previous steps. It may also be worth noting that the final step (the step related to Figures 5c and 5d) is the only step with an accuracy other than 100% for this run of Test 2.

To explore this further, similar graphs were obtained from Test 4 to see if similar phenomena could be observed with the addition of constant noise. These graphs can be viewed in Figure 6. Figure 6a displays the difference in weights for a network from Test 4 that achieved an accuracy of 25.0%, and Figure 6b shows the same but for a network from Test 4 that achieved an accuracy of 87.5%. Instead of the clear difference seen from the figures derived from Test 2, these graphs show a similar distribution in the change of weights and have similar ranges, though Figure 6a seems to have more negative change

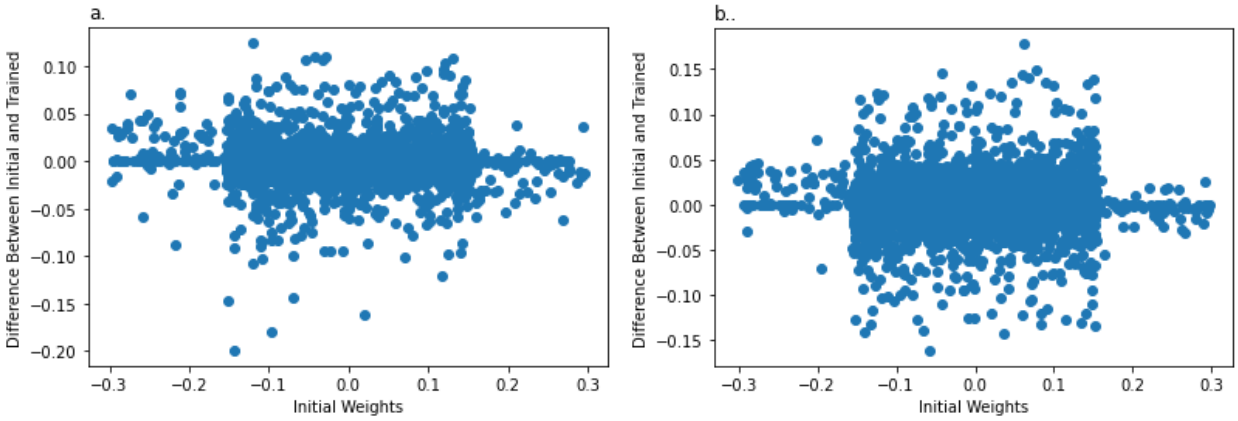

Figure 6. a. A plot of the amount that the initial weights changed after training for a network used in Test 4 (achieved 25.0% accuracy). b. A similar plot depicting how much the weights changed for a Test 4 network that achieved an accuracy of 87.5%.

than Figure 6b. This would suggest that the scattering of weight distribution seen in Figures 5d, 6a, and 6b. is more likely to be related to the noise introduced into the network, and visual estimation thereof is unlikely to be able to be useful for determining how well a network is adapted to a given set of data.

To attempt to add a more quantitative aspect to this measurement of weights, the entropy of the weights prior to testing, post-testing, and the difference between those two numbers was calculated for Tests 2 and 4. These were then organized by the overall accuracy that the net achieved, falling into one of four categories (20.0-39.9%, 40.0-59.9%, 60.0-79.9%, and 80-100%). The ranges of weight values for the weights pre- and post-testing were collected, along with the range of differences in those weights and the mean difference and standard deviation of the differences. These results can be seen in Table 7. Note that these entropy data were calculated by dividing the weight values into 30 bins, meaning that the maximum entropy (where a value is equally likely to fall into any of those bins) is 4.907 (and the minimum entropy is, of course, 0.0).

For the most part, the patterns seen in Table 7 indicate a high overall entropy for both pre- and post-training weights, with generally minimal change in entropy during training. The mean

Table 7. A table examining the entropy of the of weights of the neural nets used in Tests 2 and 4. From left to right, the columns are: range of entropies pre-training, range of entropies post-training, the range of the differences between the previous two columns, the means of those differences, and the standard deviation of the differences. For Test 2, only one net achieved an accuracy between 20.0% and 39.9%, and none did for the 40.0-59.9% and 60-79.9% ranges.

	TEST 2					TEST 4				
	Pre	Post	Diff.	Diff. Mean	Diff. St.Dev.	Pre	Post	Diff.	Diff. Mean	Diff. St. Dev
20.0-39.9%	4.087	4.110	0.0234	x	x	4.072-4.082	3.991-4.096	-0.091-0.024	-0.033	0.058
40.0-59.9%	x	x	x	x	x	4.070-4.124	3.957-4.168	-0.128-0.090	-0.012	0.069
60.0-79.9%	x	x	x	x	x	4.069-4.139	3.967-4.162	-0.114-0.053	-0.019	0.048
80.0-100%	4.072 - 4.088	4.072-4.116	-0.006-0.034	-0.033	0.058	4.079-4.081	4.018-4.166	-0.063-0.087	0.012	0.075

change in entropy suggests that the change is usually negative, but the z-scores calculated for a left-tailed test suggest no significant difference between pre-training scores and post-training scores ($z = 1.14$, $p = 0.13$; $z = 0.35$, $p = 0.36$; $z = 0.79$, $p = 0.21$; $z = 0.32$, $p = 0.37$ for each range respectively, starting with 20.0-39.9%) Comparing the entropies of Test 2 and Test 4 suggest that differences between the two tests are negligible.

Once preliminary studies were performed using the sensitivity tests, a neural network was trained to classify the EEG data categorically. Graphs as described above were generated to examine the distribution of weights for these nets. An examination of how weights compare post-training to pre-training and the differences in magnitude can be seen in Figures 7a and 7b. These graphs are similar to the graphs seen in the sensitivity tests, but with a different distribution of initial weights. The distributions were then divided by nodes and edges, though it became clear that due to the enormous disparity of the amount of nodes and edges (of 5634384 total weights, 408 were nodes and the other 5633976 were edges), the graphs representing the weight distribution of the edges only were identical to those of all the weights. Furthermore, since the weights of all nodes are initially set to zero prior to

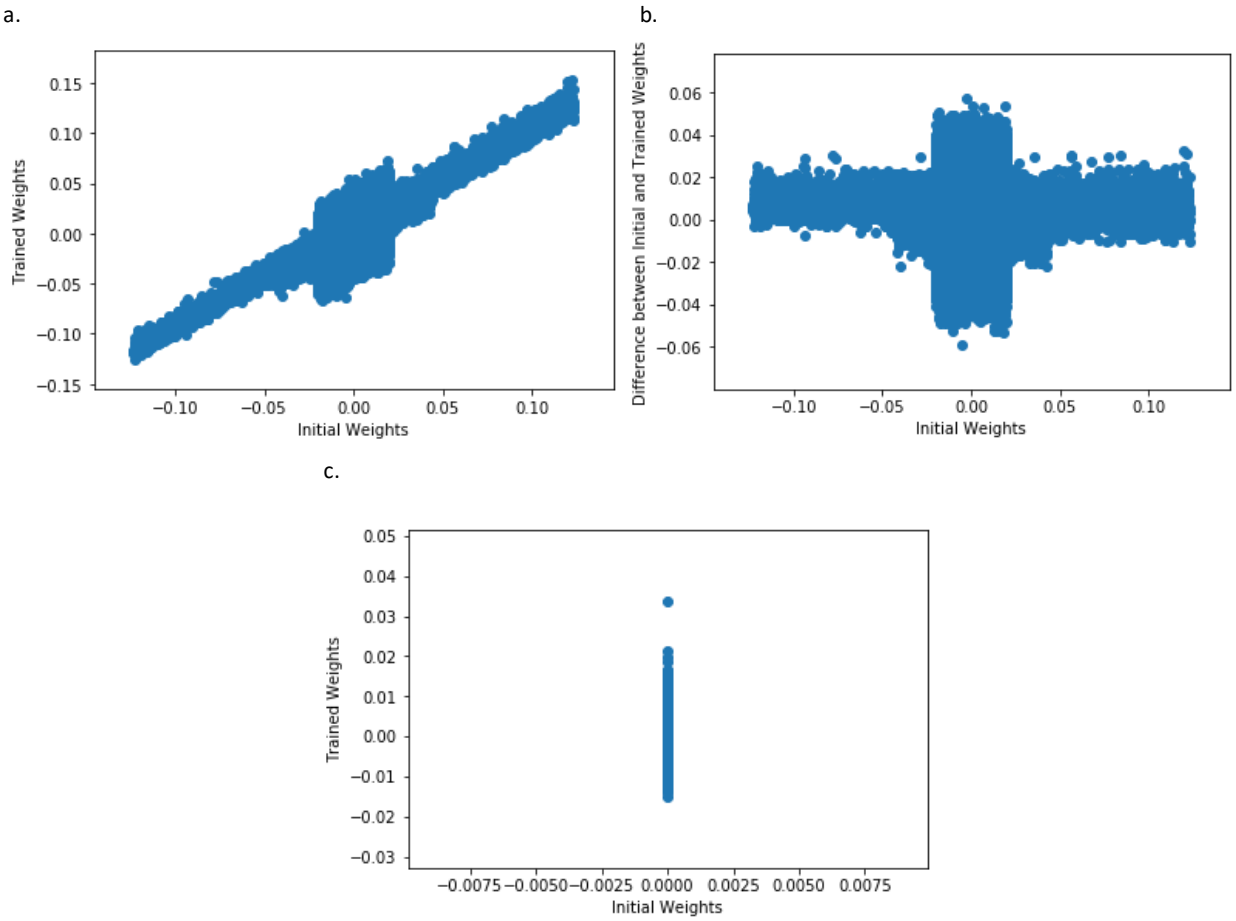

Figure 7. a. A comparison of weight values before training to the weights after training. The majority of change occurs between -0.02 and 0.02. b. A comparison of the initial weight values to how much they changed after training. c. A specific comparison of the weight values of the nodes of a neural network to their values after training.

training, the graph for the comparison of node weight pre-training to post-training is identical to the graph comparing pre-training weights to the difference after training. This can be seen in Figure 7c.

Like the sensitivity tests, once a general idea was established of how weights are distributed as a whole, entropy was calculated prior to and after training a neural net, to see how the distribution might change. This was performed to compare between nodes and edges, as well as between the three layers of the neural network (input, hidden, and output). These results can be seen in Table 8. Note that for these calculations the number of bins used to calculate entropy is unknown, meaning that while the

Table 8. A table showing entropy calculations for the weights a neural network. The top two rows show the entropy for the neural network weights as divided into nodes and edges, and the bottom three rows show the entropy for the neural network weights as divided by layer (convolutional, dense, and output), not including the input layer as the weights of the input layer are reliant on the input to the network.

	Pre-training Entropy	Post-training Entropy	Difference
Nodes	0.0	1.293	1.293
Edges	2.013	2.464	0.451
Conv	3.226	3.221	-0.005
Dense	2.000	2.376	0.376
Output	4.661	4.660	-0.001

numbers for this net can be compared to each other, they should not be compared to previously calculated entropies for the sensitivity tests, or future calculated entropies. The results here show that the change in entropy in the nodes is much greater than that of the edges, likely in large part due to the fact that node entropy begins at zero, since the weights of all nodes begin at zero. It can also be seen that the hidden dense layer has the greatest change in entropy out of the three layers by a large margin, a result that is also to be expected, as dense layers are where the most processing in a neural network should be performed. Though the results here are promising in terms of offering a way of observing a neural network as it learns, they are taken from a singular neural network and no other results were recorded for multiple nets. Despite this, the observation that entropy seems to change the most in the hidden dense layer will be taken into account, and further results will focus on the hidden dense layer in part because of it, along with the fact that the number of nodes in the input and output layers are decided based on the shape of the inputs and the number of classes being classified, respectively.

Part 5. Ablation Studies (Methods Part 6)

Having studied neural networks and their weights to gain a better understanding of them, the focus now turns to the application of that knowledge towards creating efficient and secure neural

networks. To compare the performance of the baseline networks to the ablated networks, the mean of the difference between the two lines was measured, the results of which can be seen in Table 9. A larger number indicates a larger difference between the baseline values and the ablated values, and a negative number indicates that the mean of the ablated network was larger than that of the ablated network.

Table 9. The mean of the difference between the baseline networks and ablated networks for the five targets. The numbers are calculated by subtracting the ablated values from the baseline values, so that positive values indicate a higher baseline mean and negative values indicate a higher ablated mean.

	Target 1	Target 2	Target 3	Target 4	Target 5
Mean Difference	0.407	-0.022	-0.008	0.239	0.016

Graphs were generated comparing the average performance of the ablated networks to the baseline networks for each of the five subjects, which can be seen in Figure 8. As each subject had two sets of both baseline and ablated nets, the graphed values are the mean of the two nets for each. Overall the ablated networks appear to have provided lower accuracy than the baseline networks with regards to general classification. Note that these figures do not show any information related to the invaders, but instead focus only on the adjusted accuracy for the first authentication subject. It can be seen from Figure 8a that the ablated network for Target 1 never fully trained to classify Target 1, as and such the mean difference is the largest of all the networks at 0.4. The networks for Targets 2 and 3 had negative values, indicating that the ablated network may have worked better overall. However, Figure 8b reveals a curious trend in accuracies for higher number of nodes, where the accuracy doesn't appear to settle until 145 nodes and Figure 8c shows a large similarity between the baseline and ablated networks. Figure 8 also shows that Target 3 retains a higher accuracy for larger range of nodes than any of the other four targets (within this dataset). With the data for each of the targets in mind, the same graphs and mean difference values were collected for the invaders as well. Table 10 shows the complete set of values for the mean differences, including the values seen in Table 9. With the exception of Target 1, all values increase after retraining for one invader, and increase by more than 0.08 for Targets 2, 3, and 5.

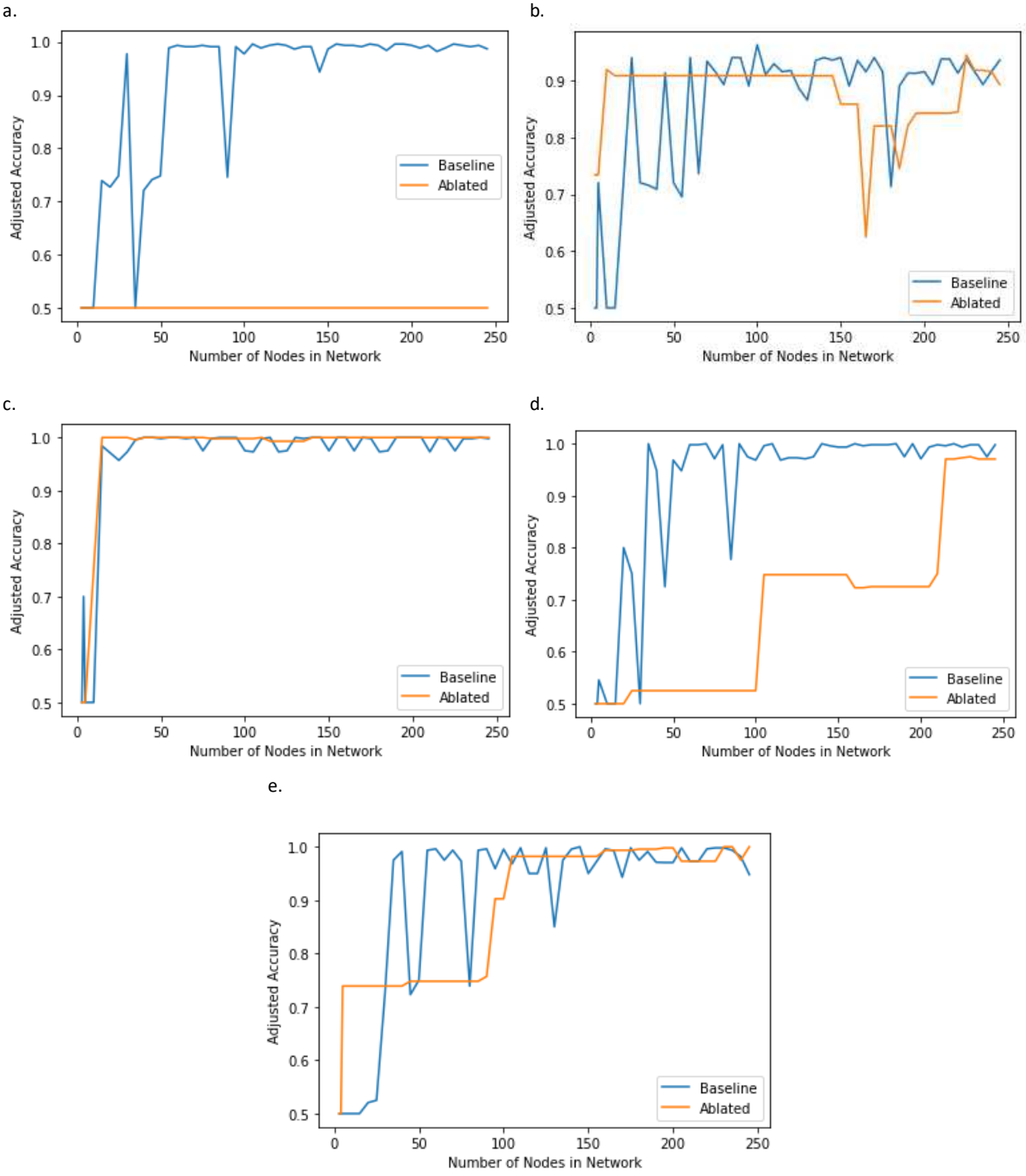

Figure 8. A comparison of the means of the two baseline neural networks to the mean of the two ablated neural networks for each of the five targets. a. The comparison for Target 1. b. The comparison for Target 2. c. The comparison for Target 3. d. the comparison for Target 4. e. The comparison for Target 5.

Table 10. The mean difference between the baseline network values and the ablated network values for each subject as well as the networks after invaders were retrained into the authentication group. The column under Target contains the same values as in Table 9.

Mean Difference per Subject/Invader	Target	1 Invader	2 Invaders	3 Invaders	4 Invaders	5 Invaders
Subject 1	0.407	0.354	0.292	0.241	0.213	0.198
Subject 2	-0.022	0.218	0.197	0.196	0.184	0.154
Subject 3	-0.008	0.094	0.150	0.163	0.175	0.159
Subject 4	0.239	0.250	0.240	0.243	0.210	0.195
Subject 5	0.016	0.103	0.114	0.146	0.138	0.167

To demonstrate the difference between no invaders and one invader, graphs comparing the number of nodes to the adjusted accuracy of the authentication group containing one invader can be found in Figure 9. Though the ablated network of Target 1 performs the same as for no invaders (due to the fact that it never trained beyond guessing accuracy), the ablated networks of the other targets appear to perform worse than their no-invader counterparts. Though the baseline nets also appear to perform more erratically, the numbers from Table 10 indicate that the overall difference between the two is greater (excepting aforementioned targets).

The relation between baseline network and ablated network accuracy was explored further by creating graphs examining the performances of the networks for authenticating the Target only versus the addition of a single invader, as can be seen for Targets 3, 4, and 5 in Figure 10. Although all prior graphs were an average of two sets of baseline and ablated networks, these graphs are taken only from one set for each. From the graphs alone, the accuracies of the single-target authentication networks (in blue) appear to be higher than target-plus-invader networks (in orange). However, the ablated networks (all graphs on the right) appear to have a larger difference between the accuracies than the baseline networks. The differences for all targets was calculated and a two-tailed paired t-test was run to test this, but did not produce significant results ($t = -1.603$, $p = 0.184$). Because of this, no conclusions can be drawn from the data provided here.

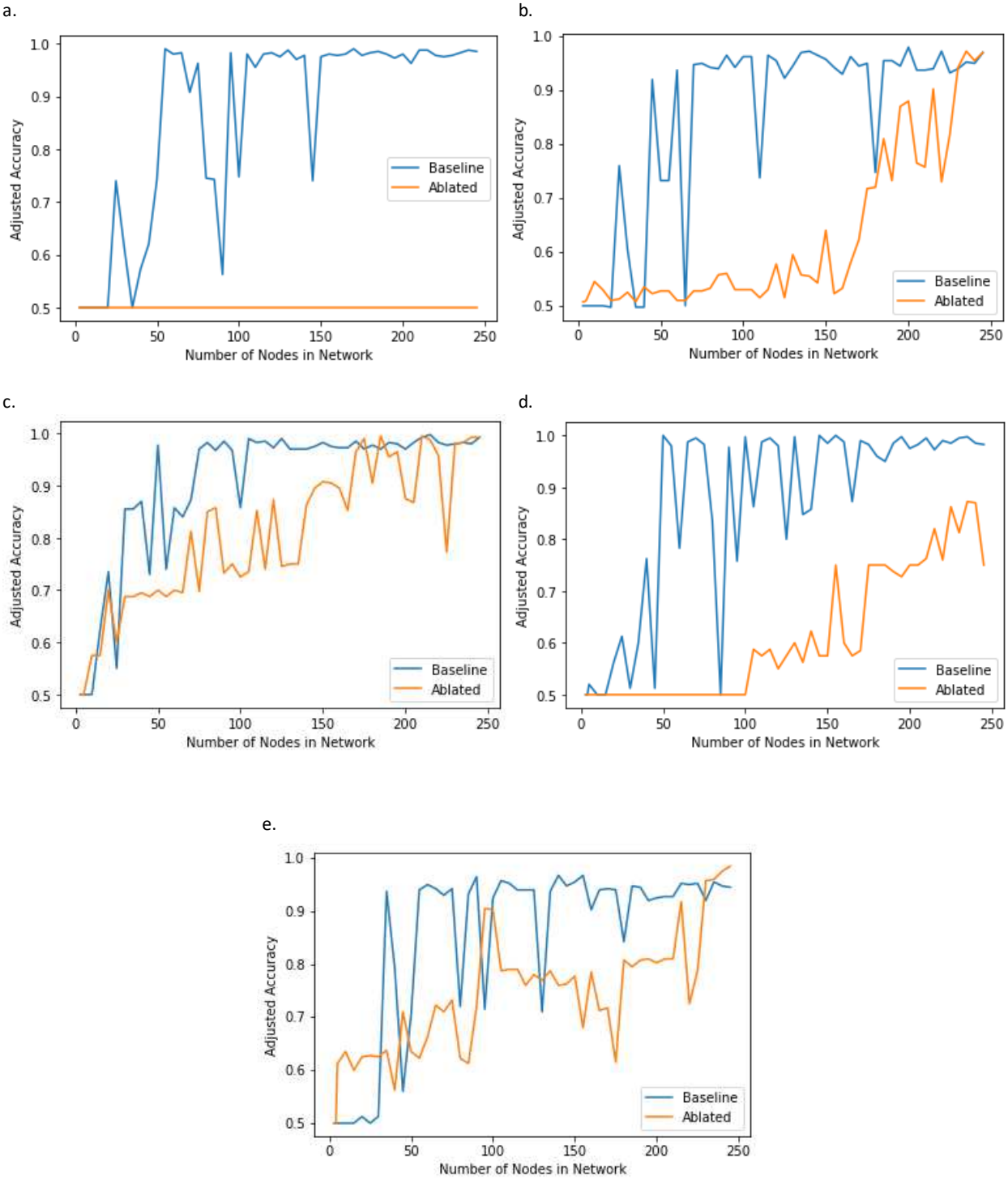

Figure 9. A comparison of the adjusted accuracy of the baseline networks and ablated networks for an authentication group with the target and one invader. a. The comparison for Target 1. b. The comparison for Target 2. c. The comparison for Target 3. d. The comparison for Target 4. e. The comparison for Target 5.

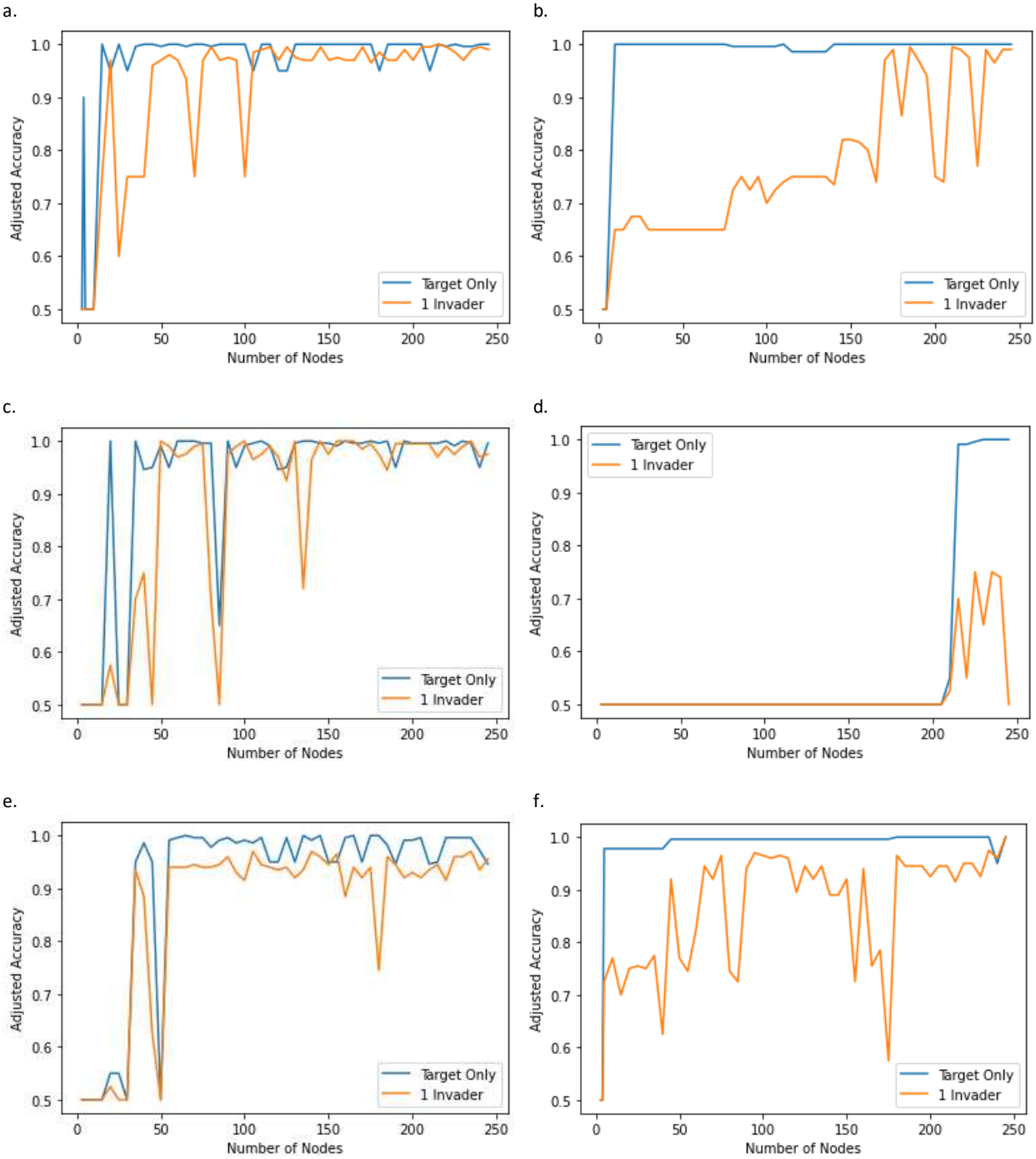

Figure 10. For all graphs, the blue line indicates accuracies of networks authenticating only the target, and the orange line indicates accuracies of networks authenticating the target and one invader. a. A comparison of baseline network accuracies for Target 3. b. A comparison of ablated network accuracies for Target 3. c. A comparison of baseline network accuracies for Target 4. d. A comparison of ablated network accuracies for Target 4. e. A comparison of baseline network accuracies for Target 5. e. A comparison of ablated network accuracies for Target 5.

As the change in entropy for the weights leading into the hidden dense layer and exiting the hidden dense layer was also collected, similar graphs and data were generated to examine the behavior of the change in entropy. Figures 11a and 11b examine the mean change of entropy for edges leading into the hidden dense layer for the two networks trained to authenticate only the target per Targets 2 and 3 respectively. Figures 11c and 11d show the mean change of entropy for networks trained Targets 2 and 3 respectively as well as five invaders. The first two graphs reveal that, while the change in entropy moves towards zero at lower node numbers, it remains close to zero for most of the nodes for

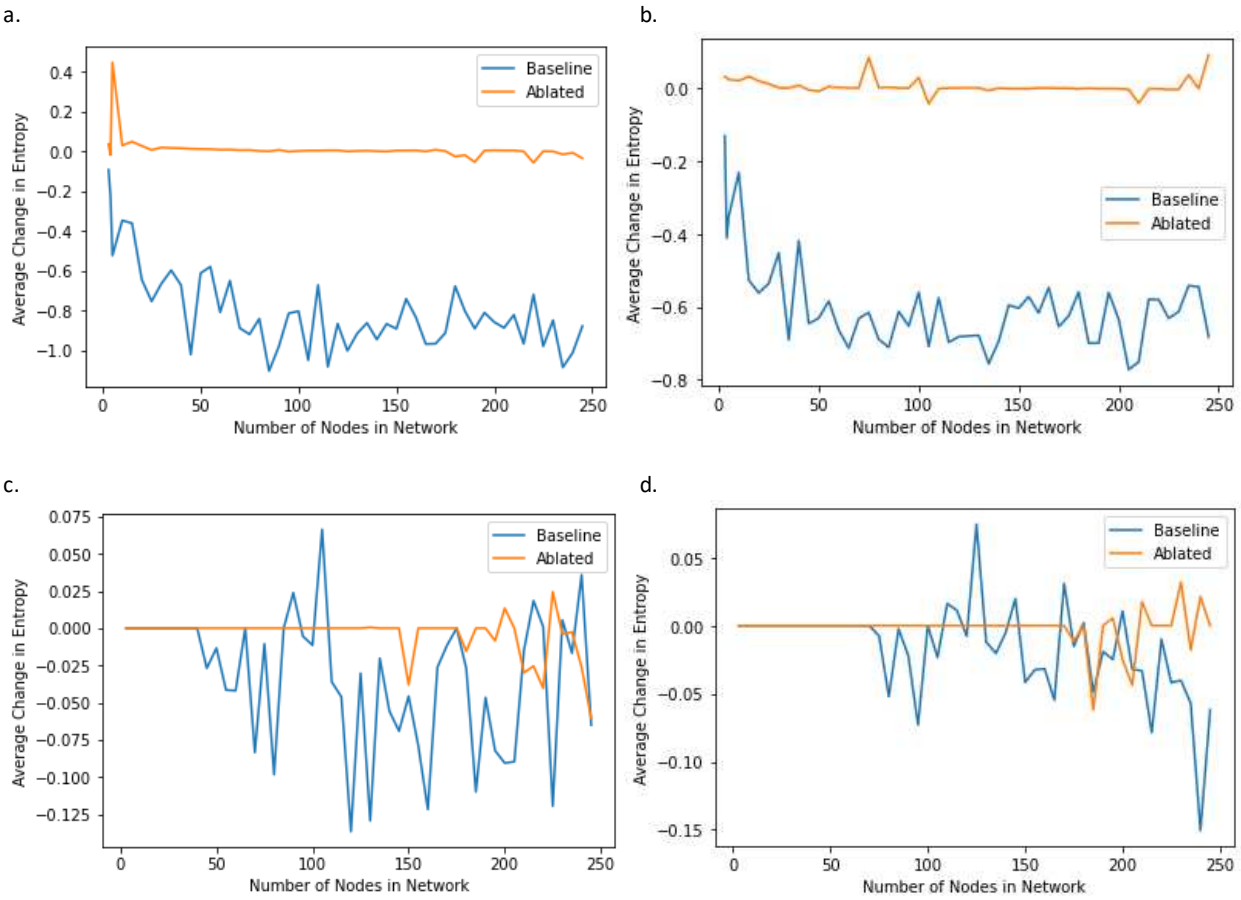

Figure 11. A comparison of the change in entropy for edges leading into the hidden dense layer to network size. a. Mean change in entropy for Target 2, trained only for the target. b. Mean change in entropy for Target 3, trained only for the target c. Mean change in entropy for Target 2, trained on the Target and 5 Invaders. d. Mean change in entropy for Target 3, trained on the Target and 5 Invaders.

the ablated networks, with more variance at higher node numbers, while the change varies more for baseline networks at any node number. It is worth noting that low entropy at lower nodes can also be explained by a fewer number of weights in the networks. Since the entropy for all of these networks was calculated with 30 bins, networks with fewer than 30 edges with weights to measure will naturally have a smaller entropy than those with more than 30 edges. The latter two graphs (11c and 11d) show that baseline values have moved closer to zero. While the two graphs appear to have more variance, the y-axis scale suggests that the values, particularly for the baseline networks, are actually much smaller. Similar trends can be seen for all Targets; Targets 2 and 3 were chosen to be representative of the ways in which change in entropy can vary. Table 11 shows the mean difference between the change in entropy for nodes leading into the hidden dense layer for all Targets and their networks, similar to Table 10. Since the calculations for this table use ablated numbers subtracted from baseline numbers, the

Table 11. The mean change in entropy difference between the baseline network values and the ablated network values for each subject as well as the networks after invaders were retrained into the authentication group.

Mean Difference per Subject/Invader	Target	1 Invader	2 Invaders	3 Invaders	4 Invaders	5 Invaders
Subject 1	-0.934	-0.236	-0.103	-0.095	-0.054	-0.025
Subject 2	-0.809	-0.141	-0.145	-0.136	-0.036	-0.028
Subject 3	-0.606	-0.403	-0.098	-0.052	-0.036	-0.015
Subject 4	-0.941	-0.136	-0.107	-0.053	-0.037	-0.071
Subject 5	-0.961	-0.076	-0.053	-0.164	-0.047	-0.025

negative status indicates that, for all of these, the mean of the ablated values was greater than the mean of the baseline values. There is a large increase in number after the first invader joins the authentication group, indicating that the ablation values and baseline values become closer together. The graphs from Figure 11 suggest that this is likely due to baseline values moving closer to zero, or becoming larger than the ablated values. Trends seem to be similar for all Targets, including Target 1, for whom the ablated network did not train.

Figures 12a and 12b show the trend for mean change in entropy of weights moving out of the hidden dense layer and into the output layer, for Targets 4 and 5 (no invaders), while 12c and 12d show the mean change in entropy for the same targets plus 5 invaders. The graphs in Figure 12 show that

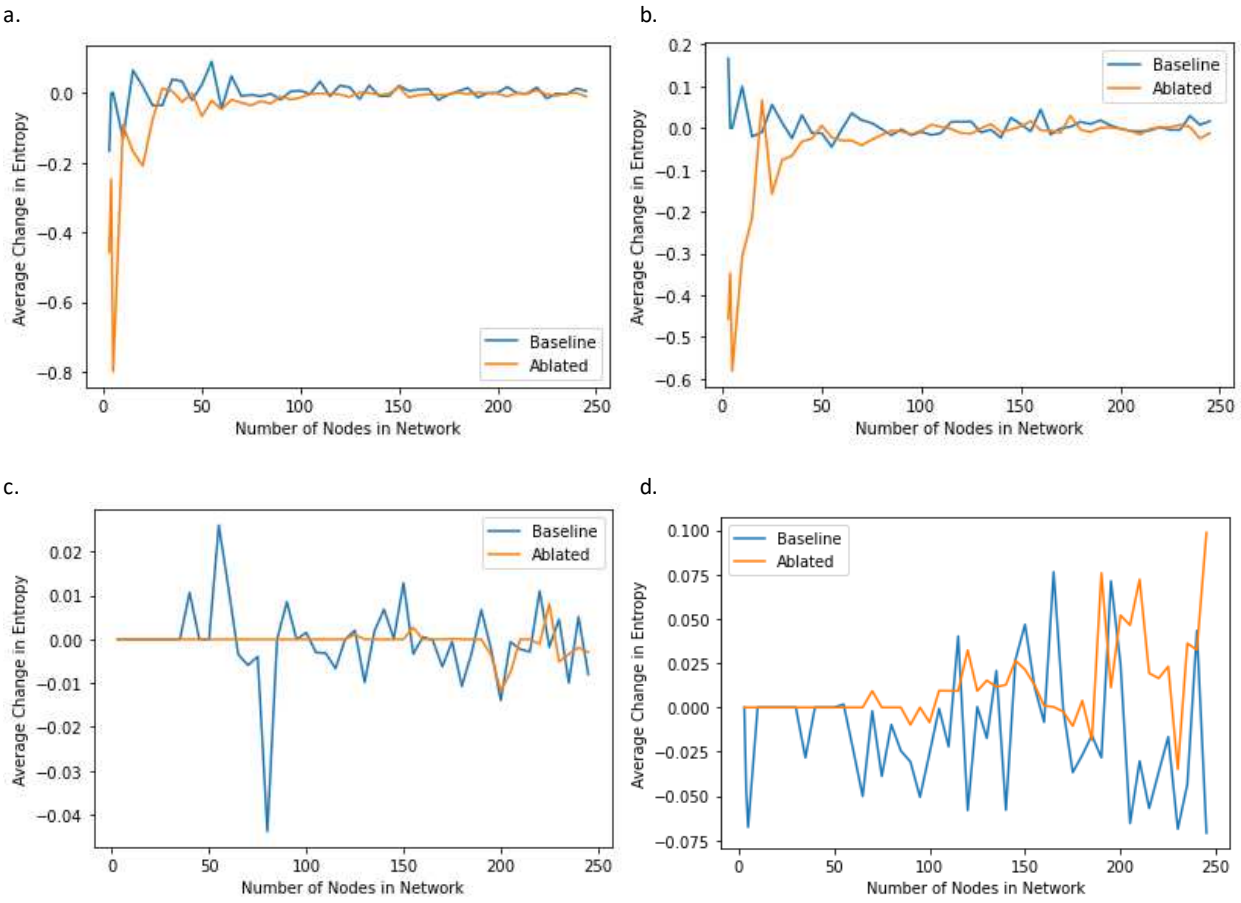

Figure 12. A comparison of the change in entropy for edges leading out of the hidden dense layer for to network size. a. Mean change in entropy for Target 2, trained only for the target. b. Mean change in entropy for Target 3, trained only for the target c. Mean change in entropy for Target 2, trained on the Target and 5 Invaders. d. Mean change in entropy for Target 3, trained on the Target and 5 Invaders.

both ablation and baseline values tend to oscillate around zero, particularly for larger node numbers, but that the mean change in entropy of the ablated set becomes more strongly negative at lower node numbers for the Target-only trained networks. Apart from that exception, differences between ablated values and baseline values seem diminished compared to the change in entropy for edges leading into the hidden dense layer. Table 12 again shows the mean difference between the baseline and ablated

Table 12. The mean change in entropy difference between the baseline network values and the ablated network values of nodes leading out of the hidden dense layer for each subject as well as the networks after invaders were retrained into the authentication group. If a number is marked <0.001 or >-0.001 it indicates that the number would round to 0.000 but is not 0.000.

Mean Difference per Subject/Invader	Target	1 Invader	2 Invaders	3 Invaders	4 Invaders	5 Invaders
Subject 1	0.058	0.002	<0.001	0.005	>-0.001	<0.001
Subject 2	0.047	<0.001	-0.003	-0.001	-0.002	<0.001
Subject 3	0.049	0.003	0.007	0.004	-0.003	-0.002
Subject 4	0.047	-0.001	>-0.001	-0.001	0.005	>-0.001
Subject 5	0.055	-0.006	0.001	0.001	-0.003	-0.003

values for the change in entropy of nodes leading out of the hidden dense layer. On the whole the magnitude of the number are much smaller than those of the edges leading into the hidden layer. The difference between baseline and ablated values decreases after re-training a network to include at least one invader for these values as well, and there do not appear to be any significant differences between the different subjects. In examining the relationship between change and entropy and saturation, the behavior between the number of invaders and the change in entropy should also be examined. This will focus on networks for Targets 4 and 5, to show the breadth of behaviors exhibited in these graphs. Figures 13a and 13b show the behavior for baseline nets of sizes 235, 240, and 245 nodes, Figures 13c and 13d show the behavior for baseline nets of size 115, 120, 125, and 130 nodes, and Figures 13e and 13f show the behavior for baseline nets of size 3, 4, 5, and 10, with Target 4 nets being on the left (13a, c, and e) and Target 5 nets being on the right (13b, d, and f). Note that these changes in entropy are not means of the two network sets for each target, but are instead just representative of one set of networks. Both changes in entropy of edges leading into the hidden dense layer and edges leading away converge towards zero as more invaders are trained into the authentication group, with edges leading into generally having a larger change in entropy than edges leading away, as was seen in

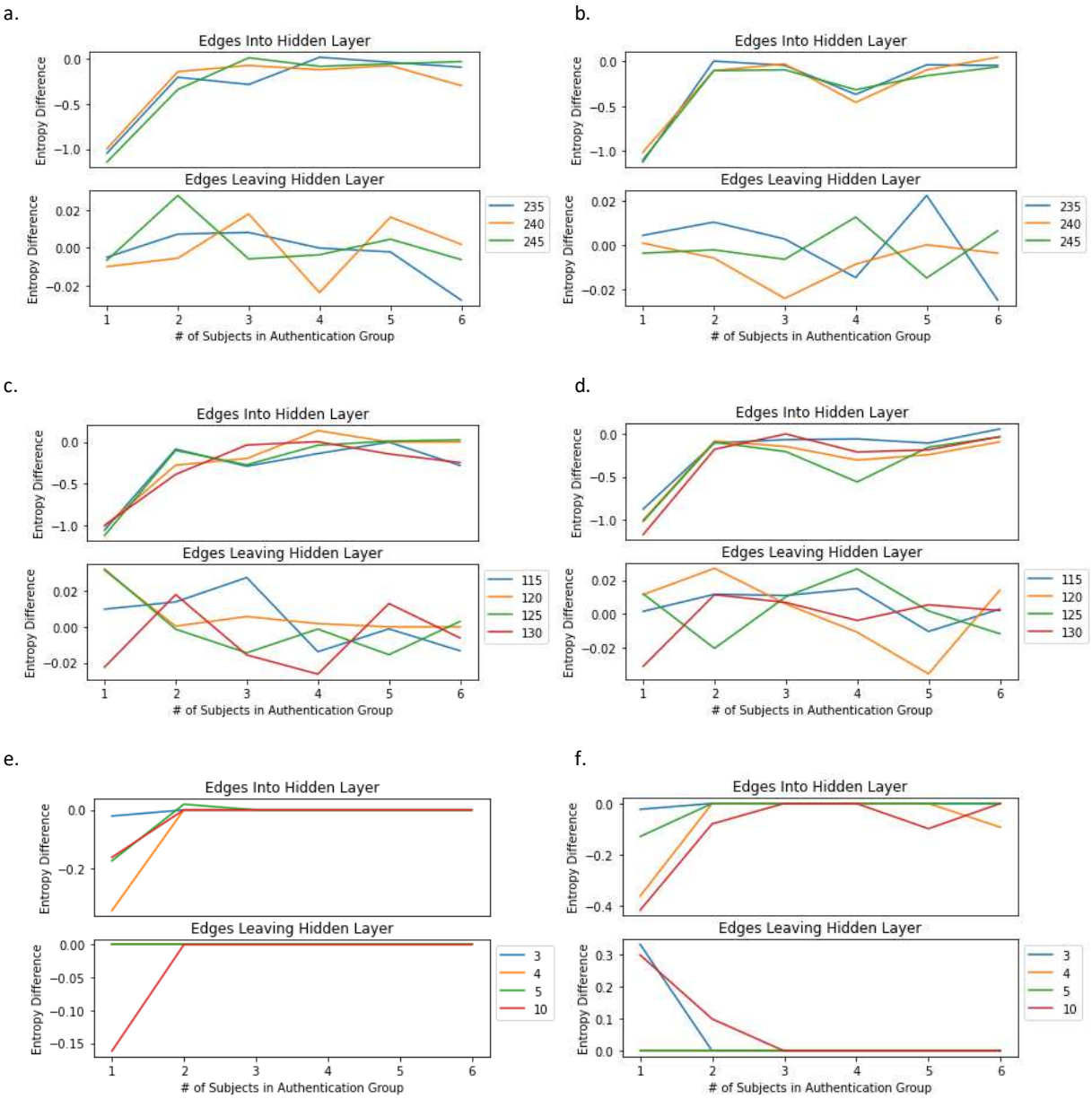

Figure 13. Graphs looking at the change in entropy for edges leading into and away from the hidden dense layer for baseline networks for Targets 4 and 5. 1 on the x-axis represents networks that are only trained for the target, and higher numbers represent the invaders. a. Change in entropy for multiple subjects in the authentication group (1 Target and up to 5 Invaders) for networks sized 230-245 trained for Target 4. b. Change in entropy for multiple subjects in the authentication group for networks sized 230-245 trained for Target 5. c. Change in entropy for multiple subjects in the authentication group for networks sized 115-130 trained for Target 4. d. Change in entropy for multiple subjects in the authentication group for networks sized 115-130 trained for Target 5. e. Change in entropy for multiple subjects in the authentication group for networks sized 3, 4, 5, and 10 trained for Target 4. f. Change in entropy for multiple subjects in the authentication group for networks sized 3, 4, 5, and 10 trained for Target 5.

comparing Tables 11 and 12. An exception to this can be seen in Figure 13f, where the change in entropy after training for just the Target extend to 0.3 or more for networks with 3 and 10 nodes.

Figure 14 has the same setup as Figure 13, but with ablated networks instead of baseline. The changes in entropy for both networks, both leading into and away from the hidden dense layer look to converge to zero as the number of invaders increases, but while the convergence occurs more quickly for both targets at lower node numbers, Target 5's networks have more variance at the higher node numbers. However, as can be seen in Figure 14d, a network's edges leading into a hidden dense layer reaching zero change in entropy in edges does not prevent the entropy from changing after training it again for a future invader. The same phenomenon for edges leading away from the hidden dense layer can be seen in Figures 14a, 14b, and 14f.

As a further observation from the data, for some networks there was an abrupt shift at five invaders wherein the authentication group, previously all misclassified to the off-group, would suddenly be reclassified to the authentication group (though, without an accompanying rise in accuracy). The current theory is that, as at 5 invaders there are an equal number of subjects in each group (six), the outcome of classifying every subject as in the authentication group or in the out-group is equal in terms of accuracy (50%, or guessing accuracy). As such, occasionally the network will switch to classifying all subjects as authenticated, rather than as the out-group, as it was doing previously, perhaps due to some decision-making from the learning algorithm.

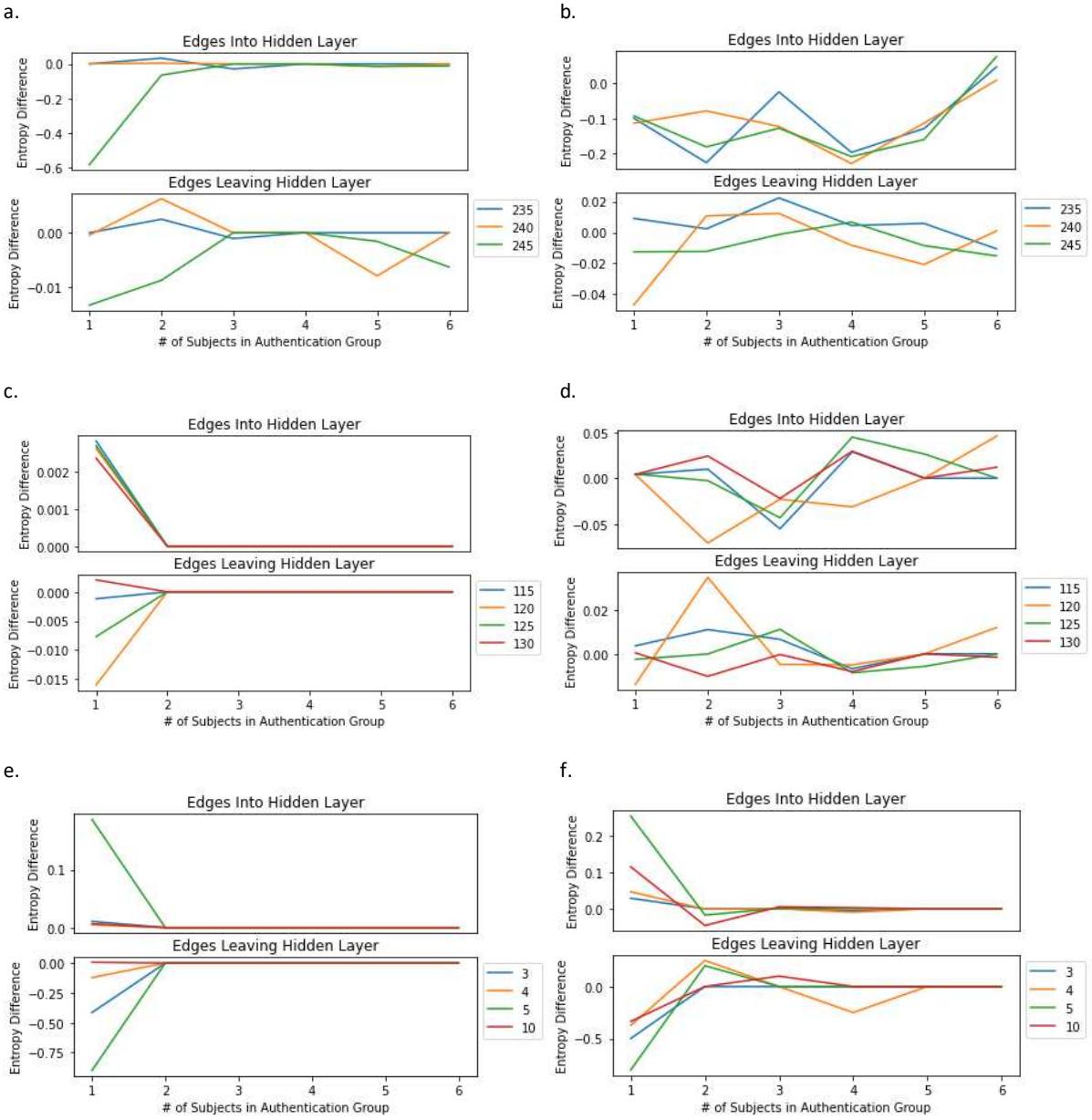

Figure 14. Graphs looking at the change in entropy for edges leading into and away from the hidden dense layer for ablated networks of various sizes for Targets 4 and 5. The values at 1 on the x-axis are values for networks that are only trained for the target, and further subjects added represent the invaders. a. Change in entropy for multiple subjects (target and invaders) in the authentication group (1 Target and 5 Invaders) for networks sized 230-245 trained for Target 4. b. Change in entropy for multiple subjects in the authentication group for networks sized 230-245 trained for Target 5. c. Change in entropy for multiple subjects in the authentication group for networks sized 115-130 trained for Target 4. d. Change in entropy for multiple subjects in the authentication group for networks sized 115-130 trained for Target 5. e. Change in entropy for multiple subjects in the authentication group for networks sized 3, 4, 5, and 10 trained for Target 4. f. Change in entropy for multiple subjects in the authentication group for networks sized 3, 4, 5, and 10 trained for Target 5.

DISCUSSION

Having examined the results of this thesis, it now becomes necessary to explain how this data plays a role in the larger field of neural network research. The brief foray into mean-based and Hamming distance-based classifiers is only a starting point in examining their suitability for cybersecurity purposes. The calculated accuracies were not large enough to be applied to classifying groups of people much larger than the small group seen here. Furthermore, the two Hamming classifiers were readily surpassed in accuracy by neural networks. Still, noting that some of the subjects were attractors (6 for the binary classifier, 3 and 9 for the adjusted classifier), such that many of the other subjects were classified as them, could be useful for future approaches involving multi-stage classification wherein the first stage involves clustering subjects.

In testing the neural networks, the sensitivity tests behaved as expected, justifying the usage of neural networks for the rest of the research. The more interesting results, for the purpose of this thesis, are those pertaining to the change in entropy for the tests. The results suggest that entropy, though it has no observable correlation with accuracy, may have a relationship with the novelty or complexity of the information that the neural network is learning. It was mentioned that the change in entropy increased as noise was added to the inputs for Test 2, despite the fact that the accuracy didn't change until the final step. This concept will be revisited later on in this section.

For now, the discussion will move on to the comparison between the ablated networks and their corresponding baseline networks. Figure 8 in the Results section shows the relation between the mean baseline nets and mean ablated nets for each of the five targets, without any invaders. The expectation of the ablation methodology was that ablated nets would largely retain their original accuracy, an accuracy expected to be comparable to the baseline accuracy for the largest network (245 nodes), past the point where the baseline nets performance would decline. Qualitatively, this behavior is somewhat

followed by Targets 2, 3, and 5. Ideally, for security, the graphs would show the orange ablated accuracies to be higher than the blue baseline accuracies, and show their accuracies to decrease at a smaller number of nodes than the baseline. As the ablated nets for Target 1 failed to train to authenticate the target, they could not show this behavior, and the ablated networks for Target 4 consistently performed worse than the baseline nets. The reason for the network failing to train could be due to being an insufficient size to train, but also may include a similarity between the target and other subject in the data set, as well as the training mechanism chosen to train. Ablation networks performed more consistently and better than baseline networks at the lowest node numbers for Target 2, but they had worse performances at higher node numbers. Ablated network performance decreased for Target 5 before those of the baseline networks did, though accuracy was maintained past the decrease of baseline accuracy. In Tables 9 and 10, this relation is quantified, and we see that only Targets 2 and 3 had ablated nets with mean accuracies above the baseline, and there it is by a difference of 0.022 and 0.008 (bearing in mind that the difference is almost certainly going to fall in the range of 0 to 0.5 for these comparisons, as 1.0 is the highest accuracy and 0.5 is guessing accuracy). In comparing to other ablation studies, only Targets 2 and 3 had adjusted accuracies above 0.9 past 100 nodes, while the baseline nets were above 0.9 at 100 nodes for all targets. This does not match the previous studies where accuracy remained similar to the initial accuracy as nodes were removed (Wu et al., 2019). There are a few reasons why that could be, including that the ablation method doesn't work as well with EEG classification. However, it is more likely to be both that some EEG data requires more nodes to learn to authenticate a target and that with more starting nodes a trend would appear so that some subjects will be easily trained into a network and others do not have enough distinguishable features for a network to readily recognize them.

There are more promising results with respect to authentication security when the graphs in Figure 9 and the other columns in Table 10 are considered. The initial aim of this research was to

improve security of EEG authentication by preventing re-training of a network for invaders, or to make it obvious when it occurred. Columns 2-6 in Table 10 show that, as invaders are introduced, the difference between the baseline accuracies and the ablated accuracies increases for Targets 2-5, with ablated accuracies performing more poorly than the baseline. Target 1 shows a decrease in difference after adding an invader, though this is likely due to the ablated networks maintaining guessing accuracy for every net, and thus the difference represents the fall in accuracy of the baseline nets. The increase is also quite small for Target 4, which is also the target that had the largest difference between baseline and ablated accuracies (with the exception of Target 1). The implication is that for nets that were more likely to have trained well to authenticate their Target, adding an invader causes a larger relative drop in accuracy compared to non-ablated nets than otherwise. However, the amount that the accuracy drops is subjective to the targets being authenticated. Furthermore, with no significant results being produced from the comparison of single-target accuracies to target-plus-invader accuracies, it cannot be said decisively that ablation significantly impacts this drop in accuracy. However, these results could be impacted by features specific to the small sample size, including the failure of the ablated networks to classify Target 1, and there is potential for further research to provide more data towards this hypothesis.

Having examined and compared the accuracies of the ablated neural networks and baseline neural networks, it becomes prudent to look at the change in entropy, as seen in Figures 11-14, and Tables 11 and 12. Graphs (a) and (b) in Figure 11 suggest that the general change in entropy for the ablated networks remains closer to zero than the change in entropy for the baseline nets when only the target is being classified, though baseline changes in entropy appear to move towards zero at lower node networks. Compared to graphs (c) and (d), the change in entropy for the baseline larger when authenticating the target versus adding in five invaders. The difference between the change in entropy for ablated nets versus baseline nets can be seen in Table 11, which reveals that all nets show a

reduction in entropy difference after adding a single invader to be classified. As before, the change in entropy of ablated networks is greater than that of the baseline for all values observed in Table 11. To explain this, the theory mentioned previously should be considered – that complexity of an input will increase change in entropy. If this theory is expanded to suggest that novelty of an input is also a measure of input complexity to a network, then it would make sense for networks created by ablating other networks to have a smaller change in entropy than networks initialized with the information each time. The ablated nets retain some amount of the learning from their predecessor, meaning that the inputs are not novel. This does not explain why the change in entropy would be close to zero for the ablated nets with 245 nodes, however. Table 12 shows a similar trend to Table 11 in that after adding invaders the difference between the ablated networks and baseline networks grows smaller, but intriguingly, the change in entropy for the baseline nets is higher than that of the ablated nets when only the target is being authenticated. Looking at Figures 12a and 12b, this seems to especially be the case at lower nodes. The reasoning behind the difference between how the entropy of ablated and baseline would decrease with the introduction of invaders evidences this theory as well, since re-training for each invader includes the input data for the target and any previous invaders, meaning that the input will not be as novel for each of those inputs for the baseline nets. The graphs in Figures 13 and 14 support this as well, as for both baseline and ablated edges leading into the hidden dense layer, the change in entropy decreases after adding invaders. On the whole, however, the original scale of change in entropy is larger for baseline nets (Figure 13) than ablated nets (Figure 14). Given the smaller scale of edges leading out of the hidden dense layer, it is more difficult to tell if they follow the same trend. Because of this, as well as the reversal of change in entropy for the edges mentioned above, edges leading out of the hidden dense layer may not be as associated with input novelty or complexity.

One of the original goals was also to attempt to saturate a network, and to hopefully use the change in entropy to observe the saturation. While the potential connection between change in entropy

and input novelty/complexity would suggest that a neural network is no longer learning once the change in entropy reaches zero, there was some evidence pointing towards the possibility of this not being a point of saturation. Namely, after some of the smaller networks reached a point of zero change in entropy for the first three or four invaders, there was an occasional small change in entropy for the fifth invader. Computationally, this usually indicated a network formula switching from assigning all testing subjects to the out-group to assigning them all to the authentication group, as this would give 50% accuracy either way. Practically, it suggests that there can be an occasional non-zero change in entropy for a very small or unimportant change in the network's learning, that does not indicate learning anything new about the input.

There are several things about the research performed in this thesis that could be improved upon in future research. Ablation is often used for optimization; however, many other settings of the neural networks used here were not optimized. For instance, the number of epochs was higher than was necessary or common to use (60, though it was observed that training often finished between 30 and 35 epochs). The number of features used for each trial was also quite high, with no studies performed relating to which features might be more useful than the others. Both of these settings can be altered for practical usage, though fortunately the smaller number of subjects being classified/authenticated allowed for some level of lack of optimization. These considerations should be included in future research though, particularly should some future research include a larger number of subjects to authenticate from. The 12-person sample size here is small for the purpose of authenticating a target from a global population, though it should still be applicable for the purposes of brain-computer interfaces (BCIs) where only a few people would ideally be using the technology, and research with a larger sample size would provide further data for the purpose of authentication in other settings. For such research, the set of metrics collected should be considered carefully, as raw accuracy can often be misleading when authenticating a single person from a large group. The adjusted accuracy used here

helps with this issue, and sensitivity would similarly be a more useful metric. Furthermore, multiple runs per target allow for a better idea of how common trends are, as would exploring more than five targets to further examine possible differences in authenticating different people.

Since the different target often had varying results for the runs mentioned here, it would also be useful to begin with a larger number of nodes for both baseline and ablated nets, since the likelihood of a net to train for the target varies from subject to subject. For instance, Target 3 had neural networks with high performance for nets as small as five nodes, whereas Target 1 only trained for the baseline nets and failed to train for the ablated nets, even with 245 nodes. Starting with a larger number of nodes and ablating larger numbers might change the effectiveness of the method. T Relatedly, the original concept was to use the 245 baseline network as the basis for the ablation to ensure that issues like that of Target 1 didn't occur, but circumstances and time restraints mean that the baseline nets for each target had to be re-run and no time remained to re-run the ablated nets afterwards. This will be addressed in a future study.

While addressing the aforementioned issues, there are also plenty of interesting directions to take this research in the future. The EEG data used here was taken from a single session, and while research by Näpflin, Wildi, & Sarnthein (2007) suggests that EEG signals remain stable over time, data taken from multiple sessions (especially across multiple days) would ensure that any fluctuations on a day-to-day basis would be taken into account. There would also be worth in research focusing on EEG data taken from activities that might be more relevant to security – resting brain potentials, or evoked potentials that could be used as passwords. As the data here was taken from subjects performing a task, the signals received by the EEG might not be as similar to signals that would be received from someone being authenticated for usage of a BCI, or other similar device. There may also be some interest in fields outside of security, including classification in general and clustering specifically. A hybrid system of clustering and authentication, for example, may overcome some of the authentication limitations of this

study while also providing a scalable design. Given that some subjects are more difficult to classify than others, and may be near impossible to classify at all, it would suggest that there are some features that some people have that allow them to be easily classified, or classified into a cluster, whereas others may be in more of a gray area. This may also have an impact on the field of subject classification in general, as the required starting conditions may differ for specific subjects in a classification problem. Though this question was not explored in this thesis, possible contributions to other areas of research should not be overlooked.

CONCLUSION

Though the research and results discussed in this thesis offer several branches of focus, there are a few takeaways that should be considered the core of what has been achieved. Though the ablation method used here did not produce the anticipated results in alignment with previous research, it did contribute to the security of the authentication. Ablated networks had a (non-statistically significant) larger decrease in accuracy after adding invaders than the baseline nets did, as evidenced by the difference between their adjusted accuracies. Network performance was not found to have a relationship with the change in entropy of its weights. However, the change in entropy of the weights due to training suggest that differences in weight distribution likely have some correlation to the novelty or complexity of an input to the neural network. These two findings have potential for future research, both in replicating findings to ensure their generalizability, as well as exploring future possibilities using these ideas.

REFERENCES

Abdullah, M. K., Subari, K. S., Loong, J. L. C., & Ahmad, N. N. (2010). Analysis of the EEG signal for a practical biometric system. *World Academy of Science, Engineering and Technology*, *68*, 1123-1127.

Babaeizadeh, M., Smaragdis, P., & Campbell, R. H. (2016). A simple yet effective method to prune dense layers of neural networks.

Bao, X., Wang, J., & Hu, J. (2009, June). Method of individual identification based on electroencephalogram analysis. In *2009 international conference on new trends in information and service science* (pp. 390-393). IEEE.

Campisi, P., & La Rocca, D. (2014). Brain waves for automatic biometric-based user recognition. *IEEE transactions on information forensics and security*, *9*(5), 782-800.

Citi, L., Poli, R., & Cinel, C. (2010). Documenting, modelling and exploiting P300 amplitude changes due to variable target delays in Donchin's speller. *Journal of Neural Engineering*, *7*(5), 056006.

Craik, A., He, Y., & Contreras-Vidal, J. L. (2019). Deep learning for electroencephalogram (EEG) classification tasks: a review. *Journal of neural engineering*, *16*(3), 031001.

Das, K., Zhang, S., Giesbrecht, B., & Eckstein, M. P. (2009, September). Using rapid visually evoked EEG activity for person identification. In *2009 Annual International Conference of the IEEE Engineering in Medicine and Biology Society* (pp. 2490-2493). IEEE.

Duggal, R., Freitas, S., Xiao, C., Chau, D. H., & Sun, J. (2020, April). Rest: Robust and efficient neural networks for sleep monitoring in the wild. In *Proceedings of The Web Conference 2020* (pp. 1704-1714).

Ellis, C. A., Zhang, R., Calhoun, V. D., Carbajal, D. A., Sendi, M. S., Wang, M. D., & Miller, R. L. (2021, October). A Novel Local Ablation Approach For Explaining Multimodal Classifiers. In *2021 IEEE 21st International Conference on Bioinformatics and Bioengineering (BIBE)* (pp. 1-6). IEEE.

Frankle, J., & Carbin, M. (2018). The lottery ticket hypothesis: Finding sparse, trainable neural networks. *arXiv preprint arXiv:1803.03635*.

Gui, Q., Jin, Z., & Xu, W. (2014). Exploring EEG-based biometrics for user identification and authentication. In *2014 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)* (pp. 1-6). IEEE.

Gui, S., Wang, H., Yang, H., Yu, C., Wang, Z., & Liu, J. (2019). Model compression with adversarial robustness: A unified optimization framework. *Advances in Neural Information Processing Systems*, *32*.

Goldberger, A. L., Amaral, L. A., Glass, L., Hausdorff, J. M., Ivanov, P. C., Mark, R. G., ... & Stanley, H. E. (2000). PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *circulation*, *101*(23), e215-e220.

Han, S., Mao, H., & Dally, W. J. (2015). Deep compression: Compressing deep neural networks with pruning, trained quantization and huffman coding. *arXiv preprint arXiv:1510.00149*.

Han, S., Pool, J., Tran, J., & Dally, W. (2015). Learning both weights and connections for efficient neural network. *Advances in neural information processing systems*, *28*.

Huan, N. J., & Palaniappan, R. (2004). Neural network classification of autoregressive features from electroencephalogram signals for brain–computer interface design. *Journal of neural engineering*, *1*(3), 142.

Jain, A. K., Prabhakar, S., & Pankanti, S. (2002). On the similarity of identical twin fingerprints. *Pattern Recognition*, *35*(11), 2653-2663.

Jasper, H. H. (1958). The ten-twenty electrode system of the International Federation. *Electroencephalogr. Clin. Neurophysiol.*, *10*, 370-375.

Kanold, P. O., Kara, P., Reid, R. C., & Shatz, C. J. (2003). Role of subplate neurons in functional maturation of visual cortical columns. *Science*, *301*(5632), 521-525.

Li, H., Kadav, A., Durdanovic, I., Samet, H., & Graf, H. P. (2016). Pruning filters for efficient convnets. *arXiv preprint arXiv:1608.08710*.

Marcel, S., & Millán, J. D. R. (2007). Person authentication using brainwaves (EEG) and maximum a posteriori model adaptation. *IEEE transactions on pattern analysis and machine intelligence*, *29*(4), 743-752.

Meyes, R., Lu, M., de Puiseau, C. W., & Meisen, T. (2019). Ablation studies in artificial neural networks. *arXiv preprint arXiv:1901.08644*.

Mohammadi, G., Shoushtari, P., Molaee Ardekani, B., & Shamsollahi, M. B. (2006). Person identification by using AR model for EEG signals. In *Proceeding of World Academy of Science, Engineering and Technology* (Vol. 11, No. CONF, pp. 281-285).

Molchanov, P., Mallya, A., Tyree, S., Frosio, I., & Kautz, J. (2019). Importance estimation for neural network pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 11264-11272).

Näpflin, M., Wildi, M., & Sarnthein, J. (2007). Test–retest reliability of resting EEG spectra validates a statistical signature of persons. *Clinical Neurophysiology*, *118*(11), 2519-2524.

Palaniappan, R. (2004). Method of identifying individuals using VEP signals and neural network. *IEE Proceedings-Science, Measurement and Technology*, *151*(1), 16-20.

Palaniappan, R., & Mandic, D. P. (2007). EEG based biometric framework for automatic identity verification. *The Journal of VLSI Signal Processing Systems for Signal, Image, and Video Technology*, *49*(2), 243-250.

Paranjape, R. B., Mahovsky, J., Benedicenti, L., & Koles, Z. (2001, May). The electroencephalogram as a biometric. In *Canadian Conference on Electrical and Computer Engineering 2001. Conference Proceedings (Cat. No. 01TH8555)* (Vol. 2, pp. 1363-1366). IEEE.

Poulos, M. et al. (1999, September). Person identification based on parametric processing of the EEG. In *ICECS'99. Proceedings of ICECS'99. 6th IEEE International Conference on Electronics, Circuits and Systems (Cat. No. 99EX357)* (Vol. 1, pp. 283-286). IEEE.

Rakitianskaia, A., & Engelbrecht, A. (2015, May). Saturation in PSO neural network training: Good or evil?. In *2015 IEEE Congress on Evolutionary Computation (CEC)* (pp. 125-132). IEEE.

Reale, R. A., Brugge, J. F., & Chan, J. C. (1987). Maps of auditory cortex in cats reared after unilateral cochlear ablation in the neonatal period. *Developmental Brain Research*, *34*(2), 281-290.

Revett, K., Deravi, F., & Sirlantzis, K. (2010, September). Biosignals for user authentication-towards cognitive biometrics?. In *2010 International Conference on Emerging Security Technologies* (pp. 71-76). IEEE.

Riera, A., Soria-Frisch, A., Caparrini, M., Grau, C., & Ruffini, G. (2007). Unobtrusive biometric system based on electroencephalogram analysis. *EURASIP Journal on Advances in Signal Processing*, *2008*(1), 143728.

Sehwag, V., Wang, S., Mittal, P., & Jana, S. (2019). Towards compact and robust deep neural networks. *arXiv preprint arXiv:1906.06110*.

Sehwag, V., Wang, S., Mittal, P., & Jana, S. (2020). Hydra: Pruning adversarially robust neural networks. *Advances in Neural Information Processing Systems*, *33*, 19655-19666.

Vadera, S., & Ameen, S. (2022). Methods for pruning deep neural networks. *IEEE Access*.

Vishnupriya, R., Robinson, N., Reddy, R., & Guan, C. (2021, November). Performance Evaluation of Compressed Deep CNN for Motor Imagery Classification using EEG. In *2021 43rd Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)* (pp. 795-799). IEEE.

Watts, M. T., Good, P. A., & O'Neill, E. C. (1989). The flash stimulated VEP in the diagnosis of glaucoma. *Eye*, *3*(6), 732-737.

Wolpaw, J. R., Birbaumer, N., Heetderks, W. J., McFarland, D. J., Peckham, P. H., Schalk, G., ... & Vaughan, T. M. (2000). Brain-computer interface technology: a review of the first international meeting. *IEEE transactions on rehabilitation engineering*, *8*(2), 164-173.

Wright, C. E., Harding, G. F. A., & Orwin, A. (1984). Presenile dementia—the use of the flash and pattern VEP in diagnosis. *Electroencephalography and clinical neurophysiology*, *57*(5), 405-415.

Wu, L., Yue, H., Chen, P., Wu, D. A., & Jin, Q. (2019). A novel dynamic network pruning via smooth initialization and its potential applications in machine learning based security solutions. *IEEE Access*, *7*, 91667-91678.

Yang, T. J., Chen, Y. H., & Sze, V. (2017). Designing energy-efficient convolutional neural networks using energy-aware pruning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 5687-5695).

Ye, S., Xu, K., Liu, S., Cheng, H., Lambrechts, J. H., Zhang, H., ... & Lin, X. (2019). Adversarial robustness vs. model compression, or both?. In *Proceedings of the IEEE/CVF International Conference on Computer Vision* (pp. 111-120).

Zhao, C., Ni, B., Zhang, J., Zhao, Q., Zhang, W., & Tian, Q. (2019). Variational convolutional neural network pruning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 2780-2789).

Zhao, S., Yang, J., & Sawan, M. (2021). Energy-efficient neural network for epileptic seizure prediction. *IEEE Transactions on Biomedical Engineering*, *69*(1), 401-411.