Using Consumer-Grade Brain-Computer Interface Devices to Capture and Detect Unaware Facial Recognitions

by

Christopher Bellman

A thesis submitted in partial fulfillment
of the requirements for the degree of

Master of Science

in

Computer Science

University of Ontario Institute of Technology

Supervisor: Dr. Miguel Vargas Martin

August 2017

Abstract

The brain's natural reaction to viewing and processing faces in an aware manner is an area of research that has been explored for previously, however the brain's unaware reactions to these stimuli prove to be fairly less explored. An experiment was performed where recruited participants viewed images of individuals' faces while their brains' electroencephalography signals were recorded using a consumer-grade BCI device. The chosen images were assigned one of three classes of recognition, corresponding with what we expect the images to be recognized as: No Recognition, Possible Unaware Recognition, and Possible Aware Recognition. Using modern filtering and analysis techniques, it was found that, in effect, using consumer-grade brain-computer interface devices, the three previously-defined classes of recognition are easily identified, both with the human eye and machine learning tools, and previous efforts to detect unaware/subconscious facial recognition have been improved on using a variety of methods for data manipulation.

Acknowledgements

I would like to thank my fantastic supervisor Dr. Miguel Vargas Martin of the Faculty of Business and Information Technology. From the writing and editing of this thesis, to the sometimes daily questions, Dr. Martin was always willing to help and guide me through all the challenges and pitfalls of my work. With the work completed surrounding and including this thesis, and all the questions about publications and academia in general, I feel more than well prepared to continue my academic career.

I'd also like to thank the rest of the ERC2100B research team (in alphabetical order by surname): Ruba Al Omari, Dr. Ramiro Liscano, Shane MacDonald, and Amit Maraj. Being able to discuss my work and bounce ideas around was invaluable to my work and I have learned so much from their input.

Finally, a huge thank-you to my family and friends for all the encouragement and support throughout the past two years of my research. My work would not have been possible without all the support you have given me.

Portions of the research in this paper use the FERET database of facial images collected under the FERET program, sponsored by the DOD Counterdrug Technology Development Program Office.

Christopher Bellman

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Overview

This thesis describes the use of consumer-grade Brain-Computer Interface (BCI) devices in detecting unaware facial recognition. In this chapter, some of the background of facial recognitions and the technologies used throughout this thesis, along with some of the populations that this work is more specifically targeting to give the reader a better understanding of the direction and positioning of this work is covered. This section includes some context-building parts such as the motivations, objectives and questions to answer, and some of the contributions that this work gives to the field of human-computer interaction and computer science.

1.2 Background

Facial recognition is a fairly well-studied field and a number of experiments have taken place to further our understanding of how the brain recognizes faces. While the conscious or aware side of facial recognition has seen much research, understanding

the brain's reaction to faces at an unaware level is a topic that has room to grow and be further explored.

One of the ways that has become a useful method for studying how the brain reacts to seeing faces is through the use of Electroencephelography (EEG). EEG consists of using electrodes placed on the scalp to measure the electrical output from neurons firing within the brain [56]. Modern advancements in computer devices and technologies have linked the modern computer with brain analysis methods (e.g. EEG) to create what is now called a BCI. BCIs are an interface, quite often in the form of a headset or cap of sensors, that reads input signals from the brain (in the case here, EEG signals) and interfaces with a computer for collection and recording [56]. In recent years, BCIs have become more popular due to the low costs of recording equipment and computers (whether mobile or desktop-based), and have been made accessible at a consumer-grade level, meaning the average consumer could purchase one of these devices at an almost off-the-shelf interaction. These consumer-grade devices generally feature applications more useful to consumers such as games or productivity tools. An example of this is the Emotiv Epoc headset [16], which has a number of applications available to it on Emotiv's web store such as mind-controlled Tetris or drone flying [15] [19]. Another cheaper and more consumer-friendly BCI headset is the Interaxon MUSE, which is primarily targeted as a meditation assistant and features a mobile companion application for its use [28]. Both headsets mentioned here have comprehensive developer tools that allow for external developers to build new applications and make use of the data being streamed from the headset [18] [27]. This additional functionality above the pure consumer-grade applications available to them allow for researchers and application developers to access the raw data and make use of these devices for a variety of applications that may not be originally intended by the manufacturers. In this thesis, the application of these consumer-grade

BCI devices in detecting unaware facial recognitions is focused on.

Previous works (see Chapter 2) have undertaken research based on facial recognition in subjects who have unhealthy or damaged brains which are unable to perform the task of facial recognition, and generally focused on more of a medical side of the uses for this technology. The work presented in this thesis focuses on the healthy brain's ability to recognize faces at an unaware level. Focusing on this large population may provide results than can be viewed in a more general light, providing a far greater possibility for applications in the future. For example, an envisioned application of unaware facial recognition may lay in law enforcement fields where witnesses of crimes can assist law enforcement agencies in identifying faces of criminals, even if the face was seen only for a brief period of time. Other potential applications may exist with the general public's utilization in mind, thus giving a far broader future impact.

1.3 Motivation

Previous research using Brain-Computer Interface devices quite often makes use of more advanced devices that utilize upwards of 100 sensors and require sophisticated wearable caps and advanced recording interfaces. While this works well for lab, research, or medical-grade users who can afford more expensive and intricate devices, those who do not have the budget or means to purchase and make use of these devices are left at a disadvantage. This is one of the reasons why this research is focused on consumer-grade devices. They are cheaper and generally more accessible for end-users to interact with. In the past, the more expensive headsets have been used in facial recognition research, but utilizing consumer-grade devices would be of great value to organizations or individuals looking to use applications or conduct research

of facial recognitions on a tighter budget, or in a more practical application-based environment (e.g. out in "the field", an office building requiring portability, etc.).

1.4 Objectives

This research was carried out to improve on existing efforts of unaware facial recognition in healthy brains using consumer-grade BCI devices. The reason this research was done using consumer-grade devices is that an increasing number of BCI devices have been seen coming onto the market in the past few years and consumers are starting to get their hands on these devices in larger numbers. With the increasing ubiquity of the headsets comes a larger segment of the population that has the means to make use of brain-based applications. This research area is two-fold: on the one side, efforts studying consumer-grade devices can develop new techniques and methods that may directly benefit consumers in the future, but on the other side, finding that we can utilize consumer-grade devices for more advanced applications such as facial recognition can also be of benefit to users in a variety of fields such as medical or law enforcement.

Machine learning is a rapidly growing field, and with advances in computing technologies, new methods for analyzing data have become available to use for extremely cheap. The costs associated with machine learning are generally just the cost of the computer hardware that machine learning software is run on, however more expensive and powerful computers are able to vastly out-perform generic non-purpose-built computers. Fantastic software libraries have become free or open-source [40] [1], allowing for a reduced barrier of entry. The ability to classify unaware facial recognitions using the human eye has been shown to be possible using specific techniques and at a small scale (see Section 2.5.2), but doing this for applications of unaware

facial recognition or on a larger scale will be very time consuming and potentially tedious. Utilizing machine learning to classify these facial recognitions for us will save us time, effort, and costs, and allow for a more rapid introduction of unaware facial recognition processing techniques into applications that will make use of them.

This thesis looks to determine if the human brain can recognize a human face at an unaware level, and if the recognition can be recorded using consumer-grade brain-computer interface devices. To this end, following question is answered:

> Can the combination of consumer-grade BCI headsets and modern out-of-the-box machine learning tools be used to accurately detect and classify unaware facial recognitions automatically, and with greater accuracy than previous work? (1)

1.5 Hypothesis

The goal of the experiment outlined in this thesis (Chapter 3) was to determine if new methods for dataset manipulation can be used to accurately detect and classify unaware facial recognition in the human brain's EEG signals, and determine, using modern out-of-the-box machine learning tools, if previous efforts can be improved upon.

My hypotheses for the results of this experiment are:

1. Using this experimental design, the Emotiv Epoc BCI device will be able to accurately capture EEG data from the brain in such a way that each recognition class is recognizable and unique.

2. Machine learning tools, however out-of-the-box, will provide adequate accuracies to be able to classify each recognition class. The naivety of using out-of-

the-box classifiers may produce adequate results, but further modification and tuning may provide greater accuracies.

3. Of the three methods for dataset manipulation (Chapter 4), the combined datasets (part of the contribution of this work) will outperform previous efforts.

1.6 Contributions

With the work done in this thesis, contributions to the general field of Human-Computer Interaction (HCI), the field of Computer Science, and the more niche field of facial recognition, insights mostly on unaware facial recognition are provided, however this work could be easily extended to cover aware facial recognitions as well. This is done through the following specific contributions:

- Determined three methods for analyzing and classifying unaware facial recognitions using EEG data, and conducting comparisons between methods and modifications to provide the greatest classification accuracies of unaware facial recognitions.
- Improved on previous efforts in classifying unaware facial recognition.
- Applying consumer-grade equipment to the field of unaware facial recognition for a more ubiquitous use in further cost-effective research projects and future consumer-grade facial recognition applications.

1.7 Structure

The goal of this thesis is to address the question (Question 1) asked in Section 1.4 and determine if:

> The utilization and combination of consumer-grade BCI devices, modern machine learning techniques, and new methods for dataset manipulation provide accurate detection and classification of unaware facial recognitions, and can improve on previous efforts to detect unaware facial recognition.

This thesis outlines the work done in background literature of related fields and topics, experiment design, experiment results, and discussion regarding the goals set in this chapter. Through this, the documentation of the development and execution of an experiment which allows for the accurate analysis of unaware facial recognitions are done, and three methods for detecting and classifying these unaware facial recognitions with the highest accuracies are compared under the guide of the research objectives.

This thesis will be structured in the following way:

- Chapter 2 - Related Work: A brief look at some previous work that influenced the work presented here in this thesis including topics such as Event-Related Potentials, aware and Unaware facial recognitions, EEG data processing techniques and methods, and other miscellaneous topics. Included in this section is a brief overview of my previous works and how they contributed to the creation of this thesis.

- Chapter 3 - Experiments: An in-depth discussion of the experiment that was performed leading to the results presented in this thesis. This section contains

numerous sub-sections where the experiment is described in detail, and the methods for pre-processing and processing the gathered participant data are explained. My hypotheses regarding the outcomes of the experiment are offered here.

- Chapter 4 - Analysis & Results: This chapter covers an in-depth look into the methods of analysis that were used to investigate the outcomes of the experiment (as outlined in Chapter 3). It explores the three methods of data manipulation used in classifying facial recognition, along with the benefits and drawbacks of the proposed methods. Each method discussed shows the results of classification from a variety of classifiers and manipulations to these classifiers to produce the best results. Among these manipulations are a look into boosting and bagging ensemble methods that combine classifiers to produce theoretically better results than the classifiers on their own, and a comparison between the two is done to highlight the benefits of each method when compared to the base-line classifiers explained in the analysis. Along with these classifiers and methods being used, a brief look into applying neural networks to unaware facial recognition is done, with a comparison of the classification results.

- Chapter 5 - Discussion: This chapter covers a discussion regarding the results of the experiment, consisting of a brief look into each classifier and their implications for the results of each dataset used in classification, a comparison of the three datasets consisting of each set's benefits and drawbacks in the context of improving classification accuracy of unaware facial recognition, and a look into some areas of note that may have influenced the results of the experiment. This section concludes with a look at the implications this work has on the field of HCI, BCI, and using machine learning to find unaware facial recognition that

takes place within the human brain.

- Chapter 6 - Conclusions and Future Work: This chapter summarizes the work presented in this thesis. An overview of this work in the context of unaware facial recognition is covered and a summarized explanation of the experiment conducted and results found in this thesis is explained. This chapter concluded with a discussion of some areas of potential future work.

Chapter 2

Related Work

2.1 Outline

In the following sections, previous works in related fields that influenced the work described in this thesis are provided and analyzed, as well as some insight into the relationship to this work. This chapter begins with a look into unaware facial recognition in both healthy and unhealthy brains, including the most related work which informed this experiment design. This moves into a discussion about EEG and the signals recorded during this experiment, and how previous works have handled EEG data. Next, a look at some related works that have implications in this work, but do not fit into the previously-mentioned categories, including topics such as implicit learning and Event-Related Potential (ERP) analysis. Finally, an overview of my previous works, including how they have influenced this work and some lessons learned which helped to improve the design of this experiment.

2.2 Unaware Facial Recognitions

To my knowledge, while most work done regarding facial recognition is on aware recognitions, there is little work done on analyzing unaware facial recognitions. Work in this field use the term "subconscious", however "unaware" was chosen to describe facial recognitions that are taking place without conscious knowledge of the brain. This was done to help differentiate the state of a person where "subconscious" could be defined as either "to be unaware of", or in the more physical sense, "to be fully unconscious" as-in "passed out" or "asleep".

Fairly recently, Martin et al. [63] ran a study to determine if unaware facial recognitions were able to be detected after participants viewed images of famous individuals and had their EEG data recorded. The experiment tasked participants with viewing faces of these famous individuals under the assumption that approximately 80% of faces would not be recognized, and the other 20% would be fully recognized. Given the ubiquitous nature of famous people in the media, another assumption was made assuming that a number of faces in the 80% set would have only been seen in passing, thus allowing for an unaware recognition to take place from this subset of faces. Unlike the work presented here, Martin et al. divided the recorded EEG data into separate epochs (nine total: 50-90ms, 130-200ms, 190-600ms, 200-300ms, 200-350ms, 250-500ms, 300-500ms, 500-750ms, and 0-999ms) and trained a Support Vector Machine (SVM) with each epoch's data separately. An average classification accuracy for unaware facial recognitions across all epochs of 64.89% was achieved with the highest accuracy of 67.16% being found in the 0-999ms epoch, which represented the entire time that an image was shown to the participant. This finding gave the suggestion that a one-second viewing time for images would be useful for analysis, so the experiment here was designed with one-second image viewing windows. This was

an interesting finding as it showed that greater unaware facial recognition accuracies can be achieved by training and testing the full set of data for an individual image rather than dividing the image up into a number of previously-defined epochs. It is this knowledge that helped inform the decision to consider the data collected in this experiment as a whole rather than dividing it like Martin did. Another area of note in Martin et al.'s work that helped inform this experiment's design was the choice of images presented to participants in the experiment. In their experiment, participants were shown images of celebrities and the assumption is that they already recognized roughly 20% of the faces. In this work's experiment, it was chosen to use completely new (assumed) faces so that participants would have no recognition to any images shown to them. As described in later chapters (Chapter 3), this experiment was split into two days so that the first day could be used to train participants rather than relying on an assumption that certain faces will be recognized. This work is possibly the most influential to the design of this experiment as it follows similar design and goals, but differing in overall design along with a safer method with less assumptions going into the experiment.

2.3 EEG and EEG Data Processing

In the past, many studies have been completed using human EEG data as the study topic and to further understand the brain's function and potentially reasons for said functions. This section highlights prior works that use interesting EEG data processing techniques that helped inform the research done in this thesis.

When considering EEG data for processing and classification, there are many directions to go with regards to post-processing, data manipulation for classification, and classification methods. Work done such as Shashibala and Gawali's [56] pro-

vide a great overview of existing technologies and techniques including EEG and BCI background information, interface styles and devices, and various classification methods and algorithms that could be utilized in classifying EEG data. A couple of the methods discussed in their work include SVM classifiers and neural networks, which were chosen to be included in this thesis for the analysis of participant EEG data (Section 4). Along with the overview of technologies and techniques are a discussion about BCI applications and what sorts of ways you can implement these technologies. Almost all of the applications discussed in their work are about assistive technologies, which would act generally as an assistant for any activity such as prosthetic control, computer text input via thoughts rather than physical button presses, and assisting drivers in detecting their alertness levels, providing a safer driving environment for both the driver of the car and others on the road [56]. Many applications for BCIs are assistive in nature and tends to be a draw for many new technologies. Detecting unaware facial recognitions, or as an extension to this work, detecting fully aware facial recognitions could, as Shashibala and Gawali discuss, act as an assistive device for a variety of applications and fields such as law enforcement or health-care. Not only would this be a useful application, but the increased usage of BCI technologies helps to promote the ubiquity of these devices for assisting people, thus promoting further research into the field.

A more modern approach to machine learning is in the utilization of neural networks. Neural networks were first conceptualized in very primitive forms back in the 1800s, but did not begin to take shape in the form of unsupervised and supervised learning until around the 1940s [46]. Modern advances in computer hardware have allowed for far more complex neural networks to be constructed and used, making neural networks a common choice for machine learning. They make use of a simulated design of the way a brain would work by constructing "neurons" which perform

calculations based on data they receive. Depending on the design and model, a neural network could have many layers of these neurons, consisting of potentially hundreds of neurons at each layer, with any given layer passing information through weighted connections to the next layer [46]. After a set of data is passed through the network and all nodes at each layer have made passed their information to the subsequent layers, some sort of decision is made at an output layer. In the case of classification tasks, the class label may be output from an input data sample with an unknown label. Anderson and Sijercic [2] used neural networks instead of more traditional classification algorithms to try to classify five different cognitive tasks performed by participants. They conducted an experiment where, over their total participant count of four, two participants averaged around 70% accuracy for mental task classification while the other two participants averaged around 33% and 45%. This suggests that, while some participants may provide strong classification, there does not seem to be a one-size-fits-all design in this work. Keeping this in mind, this thesis makes brief use of a neural network (Section 4.9) with a structure used by Subasi and Ercelebi [57], but results of the neural network are not heavily relied upon as it is not the main focus of the work, and, given the unstable results of Anderson and Sijercic, results are not expected to be exceptionally high.

In the late 80s, Keirn and Aunon [30] ran a study to determine if they can, as they say, "establish an alternative mode of communication between man and his surroundings" using EEG. As a precursor to many more modern studies such as Lee and Tan's work [31] (below), this study helped to show that classification of EEG data based on identification of mental tasks is possible. Participants in the experiment were given four tasks to complete, including tasks like geometric figure rotation, letter composition (postal letter–not an alphabetic character), complex problem solving, and visual imagination, and had their EEG signals recorded from six sensors: O1,

O2, P3, P4, C3, and C4 (according to the 10-20 system). For feature selection, recorded data was split into four unique frequency bands consisting of delta (0-3Hz), theta (4-7Hz), alpha (8-13Hz), and Beta (14-20Hz) and each power value of each sensor across these four frequency bands were used as features, similar to how in this thesis the voltage values are used as features. Along with these values, an asymmetry ratio (defined by Ehrlichman and Wiener [13] as $(R-L)/(R+L)$, with "R" and "L" being the area under the spectral density curve for right and left sides of the brain, respectively) was computed to compare the difference between the sensors on either side of participants' heads (left vs. right sides) and these values were included as features along with the power values. These features combined to form a total of 60 features for each sample in classification. This data was validated using the leave-one-out method where a classifier is trained on $N-1$ samples, and tested on the single remaining sample, and rotated through N number of times until each sample has been individually tested (Chapter 5.3 [23]). As a result of these pre-processing and validation strategies, they achieves classification accuracies ranging from 84.7% to 92% for the various tasks that were assigned. In the data processing and analysis done for this thesis' experiment, data was filtered using a bandpass filter from 0.5-12Hz, which encompasses many of the well-known frequency bands that Keirn and Aunon used, but does not make use of each frequency band as a separate entity for classification purposes. While the power values were used as features along with other identifiers (mostly in the frequency-based analysis area), this work focuses specifically on EEG voltage as features.

A study done by Lee and Tan [31] showed the use of machine learning tools in classifying tasks performed by participants, much like Keirn and Aunon's work [30]. These participants were assigned tasks in two experiments. In the first experiment: "rest" tasks where participants were not actively doing anything, "math" tasks where

participants were asked to solve basic math problems in their head, and "rotation" tasks where they were asked to rotate images in their mind. In the second experiment, the "rest" activity remained, a "solo" task where participants moved a character around the world in a game's world without engaging in any activities such as fighting or solving problems, and a "play" task where participants were instructed to play the game against an expert player. Recorded data was split into a number of slightly overlapping windows (two-second windows of data, one-second overlap between each). These windows of data were later used for training in a Bayesian Network classifier. Compared to the experiment presented in this work, Lee and Tan's feature selection and engineering was for more related to the features of the recorded signal rather than analysis of the produced voltage. Among the features that were used in their work, a few stood out as more useful for analysis of the voltage readings that were explored in this experiment and are highlighted in the Experiments section. Using the Bayesian Network classifier, when classifying all three tasks (within each experiment– rest, math, rotate in the first, rest, solo, and play in the second), an average of 68.3% was attained for classification accuracy in the first experiment, which performed far greater than the random-guess accuracy of 33%. When input data was reduced to a binary classification sets of "math vs. rotate", "rest vs. math", and "rest vs. rotate", far greater accuracies were achieved of 83.8%, 86.5%, and 82.9% across all eight subjects for each set, respectively. For the second experiment, an average classification accuracy of 92.4% was achieved for the three-task set, while binary classifications achieved higher average accuracies of 97.6%. The EEG equipment used in both of these experiments were not consumer-grade headsets like the one used in this work, but instead used a reduced number of electrodes placed at the P3 and P4 locations (according to the 10-20 system, much like [13] and [30]). Lee and Tan's work is valuable in this field as it shows how the brain's EEG output can very clearly

be used to differentiate between tasks that a participant is undergoing. In this work, participants view images of faces, which is obviously different than playing games or imagination tasks, but the techniques they used to detect this can be leveraged for facial recognition. The core mechanic used to determine the difference between the tasks is a Bayesian Network classifier that attempts to classify the individual tasks, whatever they may be. Using machine learning and modern classification techniques, a similar method can be used in detecting facial recognitions when paired up with the knowledge that the brain will produce varying EEG signals based on what it is doing at the time of recording. Another interesting technique used in Lee and Tan's work that was not used in this work is the idea of using overlapping windows for classifier training and testing. In this work, the entire one-second of data that an image was displayed on the screen for is used in training and classification whereas in Lee and Tan's work, an image is divided nine times with an overlap of 50% of each window (each window being two seconds with one second of that overlapping into the previous and next window). This was done to keep all data together as a single signal and guarantee that each section of the signal (un-windowed) would provide classification accuracy–for better or worse–to the type of recognition.

2.4 Miscellaneous

This section serves as a home for related works that do not fall into any of the above categories, but still influenced the work presented here in this thesis.

Schacter defines implicit memory as "... information that was encoded during a particular episode is subsequently expressed without conscious or deliberate recollection" [45]. One of the main pillars of the experimental design for this thesis is the assumption that participants are implicitly learning the faces shown to them in the

first phase. Based on the results of the experiment, it can be seen that participant are in-fact learning the faces and are remembering them in an unaware manner (as is shown by visual data graphs and classification accuracy). Tseng and Li [62] performed an experiment where participants searched a screen for a target image and found that pre-search queuing elements that participants were unaware of assisted participants in the search, thus providing evidence that implicit learning was taking place. Chun and Jiang [10] found, after conducting a searching experiment similar to Tseng and Li's [62], that participants' attention to specific areas of a scene or image can be guided by previously-learned contexts in implicit memory. Goujon et al. [24] ran three experiments on the subject of contextual queuing and found that knowledge and learning of patterns happens first at an unconscious level. These three prior works help identify how early repetition and training can produce an implicit knowledge of a subject before it becomes explicit knowledge ("conscious recollection of recently presented information" [45]). As defined in Chapter 3, participants are shown a number of faces multiple times to allow for implicit learning to take place and, hopefully, allow the brain to subconsciously remember the faces for the second and third phases of the experiment.

An area that was considered for additional study was the brain's reactions to more specific stimuli rather than just an overall analysis of voltage over time. Event-Related Potentials (ERPs) are a period of time after a stimuli that one expects to see some sort of elicited potential whether it be an activity in the brain's EEG signal, or lack thereof [25]. A number of studies have made use of ERP analyses for various purposes. While the results of this thesis do not make use of ERP-based analysis directly, prior works using a more ERP-based analysis featuring epoch windowing have influenced design and analysis of this thesis' experiment, and are discussed in this chapter where they apply.

An experiment by Shalgi and Deouell [55] studied human reaction to both conscious and unconscious errors forced by having participants bet money on the outcome of questions that they had to attempt to answer. They define an Error-Related Negativity index (ERN) which acts as a value for error processing within the brain and was based on how much money a participant wagered on a question. After analyzing participant EEGs, they found that ERNs were related to how aware a participant was of an error being made and that the ERN's elicited signal had a higher amplitude if their answer was more confident. While not quite focused on recognition, Shalgi and Deouell defined an ERP which was found to be related to both conscious and subconscious brain activity, and shows that the ERP will be elicited more obviously for errors that the brain is aware of, whereas the ERP will appear to be the same as a correct response for an unaware error. This shows that the brain is capable of processing information at an unaware level. While this finding has to do with error processing within the brain, it helped inspire the idea that other stimuli could also be processed subconsciously (e.g. faces, as in this work, or other stimuli, possibly leading to other avenues of future work).

An interesting application for facial recognition is in individuals who suffer from a condition known as Prosopagnosia, which impacts the brain's ability to identify faces [6]. Since the inability is due to a medical condition, I would not describe any reaction to a face as an "unaware" facial recognition, however authors such as Bobes et al. [6] refer to them as "covert" recognition. Bobes et al. [6] performed a series of experiments with a patient suffering from prosopagnosia indicating that a variety of facial recognition processes (e.g. P300 and N710 ERPs) still functioned at near-normal levels despite the fact that a facial recognition test resulted in random-guess levels of accuracy (guessing familiar face vs. non-familiar face). They find that the patient they worked with had their early facial recognition processing functions

still performing well and covert recognitions were quick. In contrast, Németh et al. [36] performed a similar experiment and suggested that prosopagnosia and the inability to perform facial recognition is caused by a missing early encoding of facial structure. Eimer et al. [14] suggest that covert facial recognition may exist within prosopagnosia patients, but they may be missing the link between visual memory and later stages of facial processing. Some of the works here suggest that some, but not all, prosopagnosia patients may still elicit core facial recognition processes, but without the ability to finish the entire process. Since facial recognition processes still work at a covert level for many prosopagnosia patients, utilizing facial recognition techniques using BCI devices may be possible, however the study of unhealthy brains and reactions are not within the scope of this work. The experiments and analysis conducted here are only considering assumed healthy brains.

2.5 My Previous Work

Over the course of my master's degree I was fortunate enough to be able to publish a number of works on the topic of BCI devices and unaware facial recognition. These works helped inform the direction of my thesis, but also played a significant role in shaping the experiment and analysis performed here.

2.5.1 Image Tagging

My first published work, entitled "Challenges in the Effectiveness of Image Tagging Using Consumer-Grade Brain-Computer Interfaces" [4] began during the Summer before the start of my master's degree after being awarded the Undergraduate Student Research Award (USRA) at UOIT. Working with Dr. Martin, we decided to tackle the application of BCIs for image tagging. This was my first exposure to using BCIs

and, more specifically, the Emotiv Epoc, which gave me great experience using the technology and the knowledge of how to retrieve and analyze data from it. We focused on the issues that we faced using the technology for image tagging, which appealed more to the Augmented and Virtual Reality field, allowing me to present the work at the Salento AVR conference in Otranto, Italy.

In terms of experimental design and instrument use, this work provided an insight into how the technology works, what issues would/could be faced, and how any problems can be alleviated along the way using techniques learned while conducting the image tagging experiment. Many of these lessons would carry over into my work in unaware facial recognition, and many of the challenges faced in that work are still faced in current work, which are addressed further in the thesis.

2.5.2 Unaware Facial Recognition

Following the acceptance of our image tagging paper, our research team and myself had designed and built a new experiment that would become the first version of my thesis experiment (outlined in Chapter 3). While being only preliminary work, my first work on unaware facial recognition titled "Excuse Me, Do I Know You From Somewhere? Unaware Facial Recognition Using Brain-Computer Interfaces" [5] had been submitted and was accepted to the Hawaii International Conference on System Sciences (HICSS) 2017. In this work, the experiment conducted is outlined (very similar to the one in this thesis) and the preliminary results of said experiment. Using the absolute value of pre-processed data, we found that each of the three recognition classes (NR, PUR, and PAR) can be easily differentiated by the human eye, suggesting that a wider participant base and machine learning techniques would be able to classify each of the three recognition classes with little trouble. It is in this experiment where we determined that the brain's EEG output, when confronted with a face that

is unrecognized, is significantly different than the EEG signals produced when viewing a face that is recognized at an aware or unaware level, thus prompting the majority of the investigation to be on determining the differences between aware and unaware recognitions. Unfortunately, after scaling this method to a larger participant base and using machine learning, the method of using absolute-value data does not, on average, improve classification accuracy over non-absolute data (see Chapter 4).

In addition to the HICSS work, this thesis' experiment, analysis, and results were prepared for publication and features a reduced analysis and discussion. Topics in this thesis that are not explored in the previously mentioned paper include an additional method of dataset creation, ensemble classification methods, neural networks, and a more in-depth discussion of the experiment and results.

2.6 Summary

In this section, a number of works that have influenced this thesis in either experiment design, analysis methods, or general considerations in the work have been covered. First, being that facial recognition is the focus of this work, a look into previous unaware facial recognition work is done. In this section, while not being the most explored topic, a few works are explored and how their methods or results influenced the design here. The largest topic discussed was about EEG data and EEG data processing. The data used in the experiment discussed later in the work makes use of participant EEG data as a base, so a number of works that cover this topic were explored. A number of more miscellaneous works were explored including a look into some implicit learning implications for the work here. This includes a look at Event-Related Potentials which, while not directly involved with the analysis conducted here, have been shown in previous work to help improve analysis by breaking recorded data

down into smaller chunks for a more granular look at the data. Finally, a brief look at my previous works that helped inform the design of this experiment and analysis, and some of the lessons learned to help improve the work produced here.

Chapter 3

Experiments

3.1 Overview

Previously, Martin et al. [63] had explored the same topic with a different approach to experiment design (see Section 2.2), which heavily influenced the design of this experiment. Like this work and others before, the use of consumer-grade BCI devices in non-medical topics highlighted the potential for taking a technology that is considered more medical in nature, and using it for more application-based purposes such as detecting unaware facial recognitions like in Martin et al.'s work and this thesis. The results of utilizing these consumer-grade BCI devices for detecting and classifying unaware facial recognition are outlined and explained in this chapter with an in-depth description of the tools and devices used, the participants, and a detailed explanation of the experiment and data pre-processing. The results and discussion of this experiment are covered in Chapter 4 and Chapter 5, respectively.

3.2 Experiment Description

The experiment was designed as a two-day experiment with a total of three phases (one on the first day, two on the second day). Each phase displayed a different selection of images of human faces to participants, and each phase's images had specific classes of recognition that were applied to them: No Recognition (NR), Possible Unaware Recognition (PUR), and Possible Aware Recognitions (PAR). It was designed to minimize any pre-existing facial recognitions which may have led to inaccurate data. The images shown to participants in this experiment were gathered from the FERET database [42] [41], and only featured images of human faces that were directly facing the camera and turned to gray-scale.

3.3 Participants

The EEG data recorded in this experiment is quite unique to each participant and varies widely in shape. Combined with the three-dataset method of analysis (Section 4.4) and the many machine learning classifiers used, this makes it hard to determine one or two variables to compare for the determination of a fixed number of participants required for the experiment. Along with this, in machine learning the goal is often to reduce bias and variance in the data through a large amount of sample data, so gathering as many participants as possible was beneficial. For the study presented in this thesis, a total of 41 participants were recruited via e-mail advertisement from the general population of the University of Ontario Institute of Technology (i.e. students, staff, etc.). All participants were aged in the range of 18 to 30 years old. Participation was available for anyone that could answer the following questionnaire with the correct answers as shown below:

- "How old were you as of September 1st, 2016?" [Age ranges, anything greater than 18 years]
- "Do you have normal or corrected to normal vision?" [**Yes**/No]
- "Do you have any condition or are taking any substance that may hinder your ability to view images shown to you on a computer screen?" [Yes/**No**]
- "Does your hair stick out longer than 4cm or is styled in such a way that would prevent a headset from coming into close contact with your scalp?" [Yes/**No**]
- "Are you willing to remove any head-wear to allow placement of sensors?" [**Yes**/No]
- "Are you allergic to common multipurpose contact lens solution that contains the following list of chemicals: Hydranate (Hydroxyalkylphosphonate), Boric acid, Edetate disodium, Poloxamine, Sodium borate, Sodium chloride, DYMED (polyaminopropyl biguanide)" [Yes/**No**]
- "What is your dominant hand?" [Right/Left-Handed]

These questions were asked and required specific answers generally so that the experiment could be successfully run with as few issues as possible. Questions regarding vision and substance consumption are to ensure participants could adequately view images on the screen, else the experiment may have been unsuccessful for the participant if they were to participate in it. The questions regarding hair and head-wear are specifically to ensure the BCI device could be placed on the head with as little sensor-scalp connectivity issues as possible. If a participant's hair was too long or styled in such a way that it is too thick to be parted to make room for the sensors to touch the scalp, data would not be able to be recorded cleanly or accurately. Regardless of this precaution, a small handful of participants were able to successfully

complete the questionnaire, but still had hair styled in such a way that only a minimal number of sensors made adequate contact with the scalp, leading to their data's disqualification after the experiment had concluded. For safety reasons, participants were asked about the sensor contact fluid used (question six) to ensure they were not allergic to any of the chemicals as it may put them at risk. In the case of an allergic reaction, steps were put in place to access medical services. The final question asked was about hand dominance, which was intended to be used to compare the result of left- and right-handed participants, however only three of the analyzed participants were left-handed, so results were found to be inconclusive (Section 4.4.2).

To assist with recruitment and to incentivize participants to join the experiment, each participant was given a total of $10 (CAD, $5 for each of the two sessions) for participation.

For the purposes of this thesis, participants are identified by a "P" followed by a number (e.g. "P10", "P80").

3.4 Data Recording Apparatus

The Emotiv Epoc BCI headset [16] was used to record participant EEG data. This headset was chosen for a variety of reasons. First, the goal of the experiment and this thesis is to determine whether or not unaware facial recognition can be detected and improved upon using consumer-grade BCI devices, and the Emotiv Epoc is what one would consider a consumer-grade device. While it is a bit on the pricey side for the average consumer to purchase ($800 USD for the latest model, which is the only one currently available for sale on Emotiv's online store [17]), it still falls at an acceptable price-point for consumers to purchase. The number of sensors on the device is also a factor in choosing the device. The Emotiv Epoc has a total of 14 sensors, while

other consumer-grade BCI headsets such as the Interaxon Muse [28] or the NeuroSky Mindwave [38] use five sensors and one sensor, respectively. This allows us to record a larger set of data from participants, or potentially consider the different areas of the brain when analyzing unaware facial recognitions.

The Emotiv Epoc makes use of 14 sensors, which include scalp placement locations based on the 10-20 system for scalp sensor placement [44]. In alphabetical order, the sensors are: AF3, AF4, F3, F4, F7, F8, FC5, FC6, O1, O2, P7, P8, T7, and T8.

The experimental software that the participants interacted with was written in Python. Due to the continuous flow of data from the headset, markers were sent from the software to the headset via emulated serial port to ensure start and stop points of each image was accurately collected so that analysis could be done on each individual image after the experiment.

3.5 Experiment Design & Stimuli

The experiment was broken up into three different phases, each with their own goals. Phase One took place on the first day of the experiment, and phases two and three took place on the second day. As will be explained further, it was designed this way to ensure adequate knowledge is preserved from the first day, but to not have the information still in short-term memory. In each phase, participants viewed images of faces presented to them in the experiment software (Figure 3.1 and Figure 3.2).

The human brain has a number of components and functions that are exhibited in EEG signals upon viewing faces (see Section 2.4), providing data that is unique for facial recognition tasks. This is why faces were chosen over other types of images such as places, vehicles, or some other sort of recognizable image which would not exhibit such facial-recognition-dependent features in the data. More so, facial recognition is

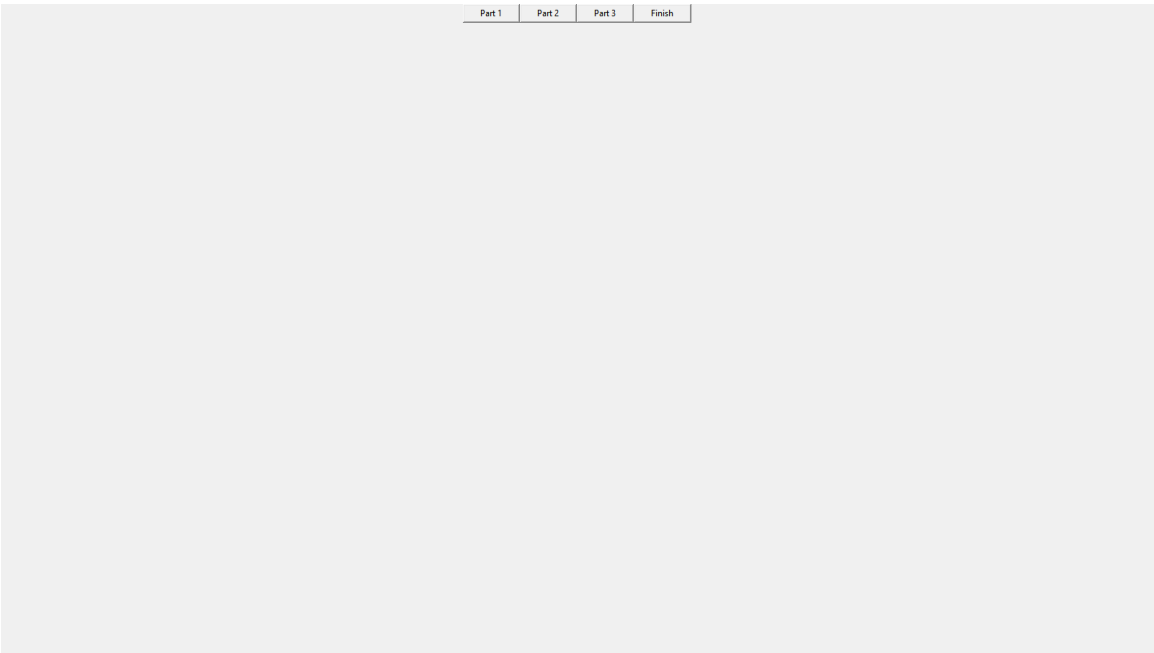

Figure 3.1: The basic interface that participants interact with before each phase featuring four buttons–one for each phase and one to finish the experiment.

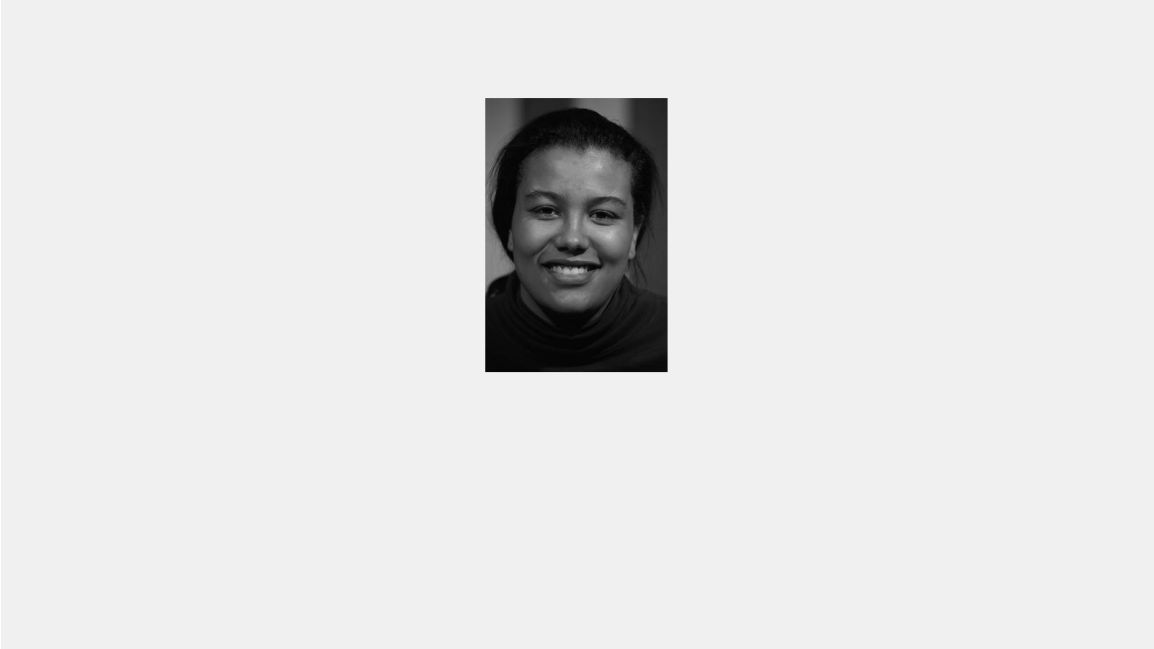

Figure 3.2: The view of what a participant sees during any of the three phases. A single image of a face appearing in the center of the screen alone. Images are from the FERET database [42] [41]

a task that we as humans do every day for every face we see. It is a very natural task that we do unconsciously, making it an interesting area of study, and hopefully one that could potentially be used to assist or improve the lives of individuals in the future. The techniques for data analysis may have use with other types of images, however this is speculation.

3.5.1 Phase One

In phase one, participants were brought into the lab and were guided through the pre-experiment activities (consent form, any participant questions, mounting of the headset). While the welcoming and discussion of the consent form lasted a few minutes, no other pre-experiment activities were done to attempt to normalize participants' mind state. After this, a computer screen placed in front of them on a desk showed the experiment software. This first phase consisted of a five-second countdown timer to ensure participants were prepared for the rest of the experiment, and then a series of images (faces) were shown to participants. Each image was shown for a total of one second, then it disappeared leaving a blank screen for one second. This was repeated 162 times for each image in the phase. Participants were asked to watch the images and were instructed specifically that they do not have to press any buttons, say anything, or make any action at all except for viewing the images. All data gathered here is what is considered as NR data, however a total of 20 images were repeated three times each (60 images) to help reinforce the implicit learning process of these faces for later phases [10] [24] [62]. Since the NR data is not the focus for classification in this experiment, the impact of the additional NR data being repeated three times in the first phase is not a concern for classification accuracy of PUR and PAR data. This phase lasted for roughly 15 minutes (from arrival to ending of the session) and consisted of a total of 162 images.

3.5.2 Phase Two

Phase two took place on the second day of the experiment where participants were asked to return to the lab. Again, the headset was mounted and participants were given a five-second countdown on the screen. After this countdown, again, a series of images of faces were shown with the same timing pattern–one second of face, one second of blank screen. A total of 102 images were shown to participants, but this time only 92 of the images were tagged as NR images. The other 10 images were taken from the set of repeated images in the first phase, which were assumed to be implicitly learned. Tagged as PUR, these images were evenly distributed through the phase. Each time a participant viewed one of these images, they were generating PUR data. As in phase one, participants were again asked to view the images without taking any action. Phase two was designed primarily to generate the first set of PUR data, but to also add additional NR image data to ensure participants were not focused on just a handful of images that may be remembered from the previous day. This phase lasted for around 5-6 minutes.

3.5.3 Phase Three

Phase three also took place on the second day of the experiment, immediately after the second phase. This time, participants were first shown a single face that they were asked to memorize. They were given as much time as they needed, and were asked to click a continue button once they felt they had the face memorized. During this memorization phase, no data was recorded. Upon clicking the button to continue, participants were again shown a five-second countdown and then another set of images (one second of image, one second of blank screen). Instead of only viewing the images, participants were asked to look specifically for the previously memorized face

Table 3.1: Image count for each recognition class.

Phase	**NR**	**PUR**	**PAR**	**Total**
Phase One	162	0	0	162
Phase Two	92	10	0	102
Phase Three	72	10	20	102
Total	**326**	**20**	**20**	**366**

within the set of images, but were asked not to do anything (click, say anything, press any buttons, etc.) again. The face that was previously memorized was evenly distributed throughout the phase much like the PUR images of phase two. Each time the participants saw the memorized face, they had an aware recognition to the face, thus PAR data was generated. Another 10 PUR images from the first phase were randomly placed within this phase to provide additional PUR data. In total, 102 images were shown to the participant in this phase with 72 being tagged as NR, 20 tagged as PAR, and 10 tagged as PUR. Adding the additional 10 PUR images allowed us to increase the number of PUR images by 100% (10 in phase two, 10 in phase three) and to also match the number of PUR images with the number of PAR images for more balanced classification later in the analysis phase. This phase also lasted for around 5-6 minutes.

A summary of the image recognition classes shown in each phase is summarized in Table 3.1.

As mentioned previously, the FERET database was used for the images shown to participants [42] [41]. This database contains images of a variety of angles of individuals' faces, but only the directly front-facing were used. A number of other databases and sources for faces were considered to be used for the experiment, however the quality or number of images of these other sources were found to be worse than the FERET database. The FERET database provides high-quality images with high consistency of lighting and backgrounds, making the images suitable for this

experiment. Combining multiple different sets of images may result in participants remembering certain backgrounds or lighting conditions, drawing focus away from the goals of the experiment. After removing a number of images that were deemed to be rather unique or obvious, 347 images were left to be used for this experiment. First, the total number of images were divided evenly for each phase. Then, images were added based on their functions. 20 images were picked to be implicitly learned (Phase Two and Phase Three each received 10 of these) and then a single image was picked to be a PAR image, repeated 20 times to match the number of PUR images. The rest of the images were then used as NR images. Initially each of the three phases was intended to have an equal number of images, but over the course of the experiment design, the numbers of each class of image were tweaked, leading to the seemingly odd numbers of NR images in each phase (e.g. 72, 92).

It is difficult to increase the number of PUR images as increasing this number results in more images being shown in repetition in Phase One (PUR images repeated three times each in Phase One), thus possibly reducing the chance that participants will learn these images implicitly. This could be countered by increasing the number of repetitions in Phase One, however this leads to the possibility that these PUR images, due to being seen many times, become learned explicitly and recognized at an aware level, thus breaking the fundamental assumption with the experiment that PUR images are learned implicitly. Testing during the design of this experiment found that using four or five repetitions of each PUR image in Phase One resulted in faces being more recognizable at an aware level, however this is from the researcher's point of view and may or may not have been seen by participants during the experiment. Previous work suggests to use between 30 and 60 samples per condition in experiments where you are measuring larger ERPs such as the P3 [32], however, due to the design of this experiment, it was difficult to achieve these numbers of PUR and PAR images.

Since the signal recorded in this experiment is the full one-second of data, this time is significantly larger than the P3 ERP, following the guidelines of Luck [32], although their suggested 30-60 range, as previously mentioned, was difficult to achieve in this experiment.

The experiment was designed as a three-phase experiment for a number of reasons. The first and primary reason for the experiment taking place over two days is to allow for the images implicitly learned on the first day to be, essentially, forgotten in short-term memory. Showing these PUR images (learned in Phase One) immediately after Phase One in Phase Two (assuming the experiment takes place all on one day) may lead to participants remembering some of the faces shown to them, thus having PUR images accidentally recognized at the fully aware level of a PAR image, providing incorrect data for analysis. The second reason the experiment is divided into three phases is based around the different goals for each type of picture. The first phase is designed specifically for participants to learn the images that are tagged as PUR. The second phase is designed as the first recall event for these PUR images. The third phase is where PAR images are introduced for the first time to participants and the instructions for the phase changes. Each of these phases has their own role in the experiment. With that said, phases two and three could be combined in the interest of experiment length. A final reason this three-phase design was chosen relates to the previous work done by Martin et al. [63]. As mentioned in Section 2.2, Martin et al.'s work makes use of the assumption that participants will recognize a certain percentage of the images shown to them without any training. Due to lifestyle and general attention to popular culture, the media, or history (generally from where their images' individuals were taken), participants may recognize more or less than expected, potentially biasing the experiment. The experiment conducted in this work eliminates any bias coming from their image choices by training participants on an

entirely new set of images that are not taken from popular culture, the media, or history, providing data without such an assumption. The training takes place in the first phase and the recall for this training takes place in phases Two and Three.

3.5.4 Post-Experiment

After phase three was completed for a participant, the headset was left on their head while a brief demo of participant brain signals was explained as well as a brief explanation of the goals of the experiment. Since the data collection was now complete, there was no worry about bias from the participants knowing the existence of unaware images.

3.6 Data Pre-Processing

Before pre-processing began, a number of participants who had generally poor recordings had their data removed from consideration. These issues with recording tended to be related to the quality of contact between the headset and the participants' scalp, generally due to hair styles or design. If, for the majority of the experiment, the sensor quality indicators in the Emotiv Testbench [20] software (for recording EEG data from the Epoc headset) indicated that the connection was poor or nonexistent for many sensors, the participants' data was not considered for analysis.

Prior to any analysis taking place, data had to be run through a number of cleaning steps. The raw data recorded from the BCI headset was first run through a bandpass filter of 0.5-12 Hz. This was chosen on the recommendation of Farquhar and Hill [21] as they found in previous work that this frequency range is "near optimal" for the filtering of EEG data for ERP classification. While the results presented here do not specifically delve into a more granular ERP-based analysis, this application in

classifying EEG data is similar to Farquhar and Hill's work. After applying this filter, an Independent Component Analysis (ICA) built into EEGLab toolbox [11]–a plug-in for MATLAB designed for analyzing EEG data–was run on the data to further clean it. Due to artifacts generated by blinking, frontal sensors were found to be nosier and required more attention. After the ICA, a general voltage threshold of ± 100 μV was used to remove any image data that was still too noisy to be used in any meaningful way.

Since we cannot control which data from the recording will record correctly or incorrectly, some images will naturally be removed via this pre-processing phase. The experiment design required there to be a rather low amount of PUR and PAR data compared to the NR data so that participants did not become aware of the repeated recognitions from the first phase of the experiment. Because of this, if an image's data was poorly recorded and had to be removed from the set, there became an imbalance between the two significant classes (PUR and PAR) which could affect classification results for that participant. By the end of data pre-processing, an average of 26.15, 2.25, and 2.38 images were removed from each participant from NR, PUR, and PAR, respectively. This is a challenge in designing an experiment that requires information to be processed at an unaware level, and a general weakness in this experiment (further explained in the Discussion, Chapter 5).

During prior work on unaware image tagging [4], it was found that the Emotiv Epoc headset did not always send data from the headset to the computer for recording in perfect intervals, thus leading to images having more or less than 128 samples in their one-second window. Some images would have, for example, 120 samples, and some may have up to 130 samples. Across all participants, the average number of samples for each image was 121 (6.57 standard deviation). In all sensors, all images were trimmed to match the lowest number of samples. For example, if

Table 3.2: A sample set of data after pre-processing. One of these will exist for each sensor (14 total)

1	5.006	3.532	0.980	−2.361	−5.938	−9.088	...	−11.226
2	13.569	15.860	18.260	20.605	22.897	25.247	...	27.743
3	−4.309	−5.912	−7.014	−7.599	−7.693	−7.374	...	−6.770
2	−6.705	−8.407	−10.056	−11.175	−11.313	−10.194	...	−7.831
3	29.988	27.482	25.684	24.491	23.618	22.655	...	21.157
⋮	⋮	⋮	⋮	⋮	⋮	⋮	...	⋮
1	6.129	5.4963	7.699	11.646	15.876	18.932	...	17.725

an image in one participant's sensor had 120 samples, all images recorded for that participant were trimmed to match the 120 samples. This helped in maintaining consistency for classification later on as all samples maintained the same number of features/dimensionality and the extra data that some images had over others did not affect the outcome for that participant.

After this pre-processing, data is now arranged in a series of rows (one row = one image sample), with each row beginning with the recognition class value (1 = NR, 2 = PUR, 3 = PAR), followed by the EEG readings at each sample. If after pre-processing there are 128 samples for a participant's sensor, the resulting dataset would consist of 129 columns and N rows, where N is the number of images a participant looked at that had valid data. Table 3.2 shows an example set of data that is the result of all pre-processing and manipulation of data. This table is just an example and is not actual data, and used to help the reader's understanding of data structure. Ellipses ("...") represent a number of data samples, removed to allow the table to fit on a page here.

3.7 Summary

The experiment outlined above makes use of three phases and two days to fulfill its goals of detecting and classifying unaware facial recognitions. In the first day, participants were given a number of images to view–some of which they were expected to learn at an unaware level by repetition. On the second day, two more phases were conducted. In the first phase (phase two), participants are shown a series of images containing many images they have never seen before, but also a number of images they had learned the previous day at an unaware level. The third and final phase again tasked participants with viewing images, but they were asked to explicitly look for an image they were shown prior to the phase beginning. It is the goal of this experiment that they would recognize some of the faces at an unaware level, and some of the faces at an aware level, and that the differences between them would later be able to be determined. After the experiment, a variety of data pre-processing and manipulation was done on all participants' data to prepare it for classification and analysis (Chapter 4).

The design of this experiment assisted in addressing some of the research questions or problems identified earlier in the thesis. This design, with its three phases and unique goals of each, allows for the capture of data representing the three classes of recognition (NR, PUR, and PAR). This confirmation of previous work provides the basis for the rest of the analysis and eventual results, and provides data to be used for classification purposes. Since the three recognition classes produced as a result of this experiment have been found to be, in many cases, unique enough to be identified by the human eye, the introduction of machine learning for automated classification may prove useful for enhancing the identification of unaware facial recognitions elicited by individuals. Using this data for classification may allow us to more rapidly and

accurately identify unaware facial recognitions. Modifications to the recorded data could provide and even greater level of accuracy and improve on existing classification accuracies of unaware facial recognitions.

Chapter 4

Analysis & Results

4.1 Outline

This chapter covers the results of the experiment as described in Chapter 3. It breaks down the results from a number of sources to explore the questions asked and objectives set in place at the beginning of this thesis (Section 1.4), and approaches these questions using multiple methods to produce the answers to said questions and objectives. To begin, a brief explanation of the methods for classification and result measurement is done to provide an understanding of the techniques used in the analysis. Next, a breakdown of the three datasets created and used in this experiment, including the individual dataset and results, and the two combined datasets with their classification results. A look into classifier bagging and boosting for one of the combined datasets is done, which looks to improve classification accuracies over the base accuracies that are achieved through the standard classifier usage. These two boosting and bagging methods are briefly compared, along with a comparison between the individual sensor and combined sensor datasets. Finally, a quick look at the application of modern neural network classification using the TensorFlow and

TFLearn Python libraries [1] [58].

4.2 Classification Methods

For unaware facial recognition classification, the methods were the same for all three types of datasets (datasets explained in more detail later). Before classification, each dataset was split with 60%/40% weightings for training and testing (respectively) datasets. This means that 60% of the total dataset was used to train a classifier, and the remaining 40% was used as test data to estimate the performance of the classifier. Each training underwent 5-fold cross validation. A total of five classifiers were used from the Scikit-Learn Python library [40]: Random Forest (RF), Gaussian Naive Bayes (GNB), SVC, Decision Tree (DT), and K-Nearest Neighbours (KNN) classifiers. Each of these classifiers were used in an out-of-the-box configuration, meaning they were used with their parameters left at the default values as set by the authors of Scikit-Learn. This was done to determine the out-of-the-box performance of these classifiers on EEG data for classifying unaware facial recognition. Future work may explore optimizing classifier parameters for better classification performance. During the 60-40 split of training/testing data, data for each sample is normalized which scales it to the unit normal (mean of 0, standard deviation of 1).

4.3 Result Measurement

To determine performance of the classifiers, the F1-score (F) (commonly referred to as the "F-score") is used. F-score is commonly used to determine the performance of binary classification. The F-score is a value within the range of 0.0 to 1.0, with 0.0 being the worst possible score, and 1.0 being the best possible score. It uses two

Table 4.1: The logic table for this experiment's classification outputs.

Event	Result
PUR classified as PUR	True Positive
PUR classified as PAR	False Negative
PAR classified as PAR	True Negative
PAR classified as PUR	False Positive

values based on the output of classification results to produce the F-score: Precision (Equation 4.2) and Recall (Equation 4.3), which are both defined by calculations of True Positives (TP), False Positives (FP), and False Negatives (FN). Table 4.1 explains which classifier output conditions create each value. Kaggle [29], a popular website that hosts data science competitions, defines F-score as

$$F = 2 \times \frac{P \times R}{P + R} \tag{4.1}$$

with Precision (P) being

$$P = \frac{TP}{TP + FP} \tag{4.2}$$

and Recall (R) being

$$R = \frac{TP}{TP + FN} \tag{4.3}$$

4.4 Datasets

For the purpose of classifying unaware facial recognition, three different datasets were created and used to test classifier performance on each. The first set tested is an individual look at each participants' sensor data, and considering each sensor as its own dataset. The next two datasets use different methods of combining participant data

into individual sets. These sets are "Combined and Averaged Dataset" (CAD), and "Combined Dataset" (CD), which are generally referred to categorically as "combined datasets" as they both make use of combined sensor data. This categorical description (versus the individual datasets) is not to be confused with the individual dataset "Combined Dataset (CD)". Each set of data is isolated within each participant so no averaging or data combination of multiple participants' data is ever done. Since EEG signals seem to be, after reviewing many participants' data, fairly unique among individuals, combining participant data may result in poor classification accuracies. Due to the relatively small size and uniqueness of collected data, determining if the data is biased in any way is difficult. Averaging over all sensors for each participant, NR and PUR data variance was 245 while PAR data was 216. These higher numbers show that, in general, recorded EEG signals in this experiment are highly variable, even when considering intra-class variation calculations. This also does not consider a combination of participants and calculating variance inter-participant. With that said, future work may explore a more general approach for unaware facial recognition that combines participant data in a more general way.

4.4.1 Individual Sensor Dataset

The first, most straight forward method of analysis is considering each participant's sensor data as an individual data set and attempting to classify unaware facial recognition within a sensor instead of an entire participant. Classification was done in the same way as all other datasets as outlined Section 4.2. After reviewing the results of the classification, it appeared that the SVC classifier struggled greatly with this dataset, producing poor results (generally either 0 or 0.609, regardless of sensor), so to provide balanced, un-corrupted results, data displayed and discussed here are not considering the SVC classifier.

Table 4.2: Participant P07's individual sensor classification results.

Sensor	Decision Tree	Gaussian Naive Bayes	K-Nearest Neighbours	Random Forest	Average
AF3	0.571	0.615	0.667	0.429	0.571
AF4	0.286	0.615	0.714	0.400	0.504
F3	0.375	0.714	0.714	0.750	0.638
F4	0.267	0.462	0.500	0.353	0.396
F7	0.429	0.667	0.714	0.571	0.595
F8	0.526	0.588	0.533	0.476	0.531
FC5	0.400	0.615	0.462	0.615	0.523
FC6	0.400	0.429	0.556	0.625	0.503
O1	0.375	0.769	0.727	0.714	0.646
O2	0.545	0.909	0.400	0.500	0.589
P7	0.429	0.400	0.556	0.471	0.464
P8	0.588	0.667	0.625	0.706	0.647
T7	0.471	0.556	0.556	0.353	0.484
T8	0.500	0.533	0.667	0.667	0.592
Average	0.440	0.610	0.599	0.545	

4.4.2 Individual Sensor Dataset Classification Results

Given that each sensor is classified separately from every other sensor, there are a total of 56 results for each participant ((# Of Sensors) × (# Of Classifiers)), which is difficult to boil down to smaller amounts of numbers for comparison. Table 4.2 provides a sample from participant P07's results. While it is inappropriate at this point to average across sensors or classifiers, this data has been provided in Table 4.2 for consideration. Taking the average data for both sensors and classifiers, we find that certain censors perform better or worse than others, and certain classifiers also show a variance in results. Comparing participant P07 (Table 4.2) and P12's data (Table 4.3), we can see that results vary by a large margin between the two, thus lending further confirmation that inter-participant classification may not provide strong results.

To get a more global sense of the best- and worst-performing sensors, individual

Table 4.3: Participant P12's individual sensor classification results.

Sensor	Decision Tree	Gaussian Naive Bayes	K-Nearest Neighbours	Random Forest	Average
AF3	0.625	0.500	0.444	0.600	0.542
AF4	0.308	0.471	0.353	0.421	0.388
F3	0.600	0.500	0.353	0.500	0.488
F4	0.182	0.333	0.333	0.706	0.389
F7	0.769	0.667	0.667	0.588	0.673
F8	0.333	0.533	0.625	0.526	0.504
FC5	0.429	0.250	0.250	0.429	0.340
FC6	0.533	0.429	0.556	0.429	0.487
O1	0.353	0.471	0.400	0.533	0.439
O2	0.533	0.400	0.500	0.526	0.490
P7	0.308	0.375	0.444	0.286	0.353
P8	0.429	0.471	0.353	0.632	0.471
T7	0.471	0.500	0.588	0.526	0.521
T8	0.526	0.588	0.625	0.571	0.578
Average	0.457	0.463	0.464	0.520	

participant data was stripped down to gather only the best- and worst-performing sensors for each classifier. Table 4.4 shows a sample of a few participants' best and worst sensors for each classifier (full participant data for this exists in Appendix 1.5). Upon determining which sensors performed the best for each participant, each sensor was counted to find which sensors tended to show up the most in best- and worst-case performances for all participants. Table 4.5 and Figure 4.1 show each sensor and how often they were the best sensor, the worst sensor, and how involved they were in both the best and worst sensor categories ("involved" being the sum of best and worst frequencies). After reviewing the sensor involvement in the best/worst frequency categories, it was found that four sensors appear an equal amount of times (F7, FC6, P7, and T8). To look into each category individually, the sensors that achieved the highest best frequencies were F7 and P7 with 10.811%, and the highest worst frequency was the FC6 sensor with 11.486%. The highest best frequency sensors are

Table 4.4: Sample best-/worst-sensor results for a few participants.

		Decision Tree	Sensor	Gaussian Naive Bayes	Sensor	K-Nearest Neighbours	Sensor	Random Forest	Sensor
P00	Best	0.769	AF3	0.615	AF3	0.615	AF3	0.714	AF3
	Worst	0.154	F4	0.143	FC5	0.125	FC5	0.133	FC5
P02	Best	0.632	AF4	0.769	F7	0.667	T8	0.667	T8
	Worst	0.143	F4	0.154	P7	0.364	AF3	0.286	P7
P03	Best	0.571	T7	0.500	F4	0.609	F4	0.700	O1
	Worst	0.267	AF4	0.154	FC5	0.286	T8	0.333	P7
P04	Best	0.625	FC5	0.615	AF4	0.632	P7	0.588	F4
	Worst	0.333	T7	0.333	F7	0.333	T7	0.308	O2
P07	Best	0.588	P8	0.909	O2	0.727	O1	0.750	F3
	Worst	0.267	F4	0.400	P7	0.400	O2	0.353	T7
P10	Best	0.857	F8	0.800	O2	0.667	FC6	0.727	F8

only a couple of percentage points higher than the next and only about 3.7% above the average (7.14%), so it does not appear that any specific sensor can be said to be the "best" sensor for classifying unaware facial recognition. The fact that many sensors are around the "random guess" or assumed average value ($100 \div 14 = 7.14$ (approx.)) shows that no sensor really stands out, allowing us to group sensors together and consider their data as a whole in further analysis (Sections 4.4.3 and 4.4.3).

Analyzing the left-/right-handedness of participants within the combined datasets becomes meaningless as all the sensor data is combined, however the individual sensor datasets can be analyzed more thoroughly for spatial results. Each participant's best and worst performing sensor data was analyzed for both left- and right-handed participants to determine if there were any trends that could be extracted from each set of participants. Table 4.6 shows the percentage of the time each sensor appeared as best and worst-performing sensors for both left- and right-handed participants for all four classifiers used. In the context of Table 4.4, "best" and "worst" mean the sensors and F-scores that were the best or worst performing overall (highest/lowest F-score) for an individual participant. In the context of Table 4.5 and Figure 4.1, the terms "best

Table 4.5: The frequency of best-/worst-performing sensors over-all and sensor involvement in best/worst frequencies. Values are in percentage of the time that a sensor showed up in each category (best/worst).

Sensor	Best Frequency	Worst Frequency	Involvement
AF3	8.108	5.405	13.514
AF4	4.054	5.405	9.459
F3	4.054	2.027	6.081
F4	4.730	4.730	9.459
F7	10.811	9.459	20.270
F8	6.757	6.757	13.514
FC5	3.378	9.459	12.838
FC6	8.784	11.486	20.270
O1	8.108	4.054	12.162
O2	8.784	6.757	15.541
P7	10.811	9.459	20.270
P8	4.730	6.757	11.486
T7	6.757	8.108	14.865
T8	10.135	10.135	20.270

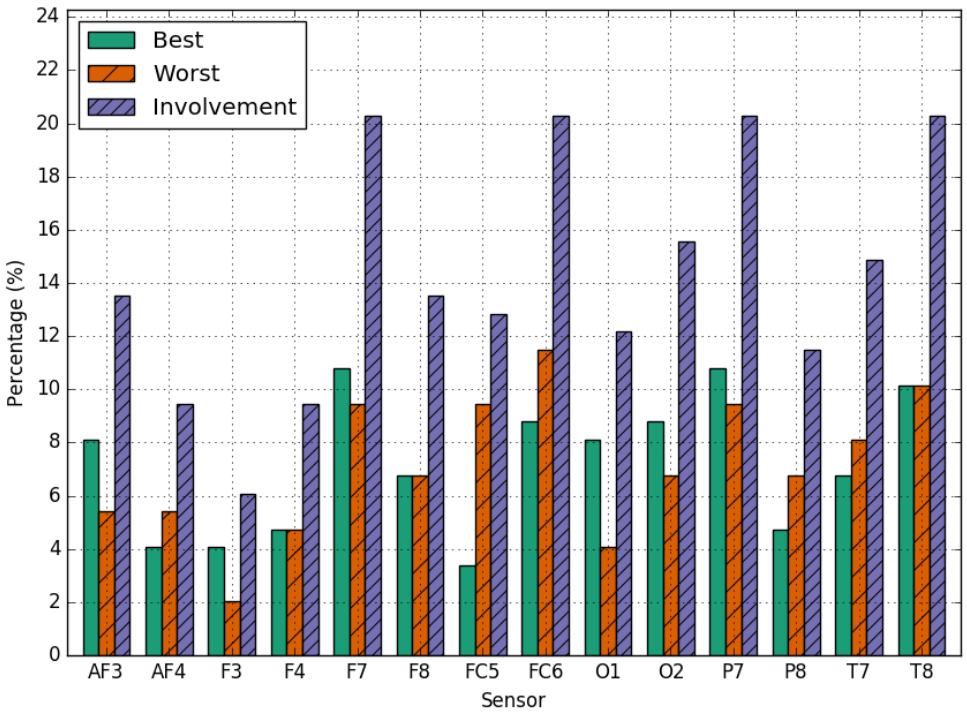

Figure 4.1: Data from Table 4.5 in visual format.

Table 4.6: Percentage of time each sensor appeared as best or worst-performing sensor for left and right-handed participants.

Sensor	Best - Left	Best - Right	Worst - Left	Worst - Right
AF3	0.00	8.82	8.33	5.15
AF4	8.33	3.68	8.33	5.15
F3	0.00	4.41	0.00	2.21
F4	0.00	5.15	8.33	4.41
F7	8.33	11.03	33.33	7.35
F8	0.00	7.35	0.00	7.35
FC5	0.00	3.68	0.00	10.29
FC6	0.00	9.56	8.33	11.76
O1	8.33	8.09	0.00	4.41
O2	0.00	9.56	16.67	5.88
P7	8.33	11.03	16.67	8.82
P8	0.00	5.15	0.00	7.35
T7	8.33	6.62	0.00	8.82
T8	58.33	5.88	0.00	11.03

frequency" and "worst frequency" simply mean the frequency at which each sensor appears in the results as the highest or lowest performing (highest/lowest F-score) sensors across participants. Only three of the participants included in this analysis were left-handed, so the sample size for this analysis is quite small, which results in extremes on both ends of the representation spectrum (some sensors never appear as best/worst, some sensors appear more frequently as best/worst) whereas the right-handed participants are more plentiful, thus giving a more accurate representation of sensor best/worst frequencies.

Since the number of left-handed participants is so small, meaningful comparison between left- and right-handed participants is not possible, but we see more obvious numbers such as the left-handed T8 sensor which is represented as the best sensor 58.33% of the time compared to the next highest of AF4, F7, O1, and P7, which appear 8.33% of the time (once each, far above the expected average representation of 7.14%). Could this suggest that the T8 sensor is more active for left-handed par-

ticipants? Possibly, but the sample size is far too low to make any suggestions of statistical significance. To compare, there is no sensor from the right-handed participants that tends to appear vastly more than the average expected value (7.14%), although a number of the sensors tend to be under-represented (e.g. AF4, F3, F4) and some over-represented (FC6)

The findings from this dataset indicate that using individual sensors performs poorly and tends to produce resulting F-scores that float around the random guess score of 0.5. When compared to the results of the combined datasets (upcoming, Sections 4.4.3 and 4.4.3), it is very weak and may not want to be considered for the purposes of classifying unaware facial recognitions. Along with the classification results, no one or two sensors stand out as being vastly more capable of classifying unaware facial recognitions than others (Table 4.5), so this method does not provide any benefits for determining specific sensors that may prove to be more accurate or useful in this application.

With this said, it is important to note that because each sensor was being considered and classified individually, the number of samples in each class is very low, thus providing results that may not be safe to make conclusions on. The reasons for this are not only because of the low image count for both the PUR and PAR data, but also because of the pre-processing phase where images are removed if they have too many artifacts or are generally poor in quality. Across all participants' sensors, the average usable image count is 17 (STD 4.68) and 17 (STD 5.73) for PUR and PAR, respectively. For many sensors across all participants, this results in entire classes being removed, or the majority of images from a class being removed.

The next two datasets consider an individual participant's data as a whole and make use of sensor combination. Since they both make use of combination instead of single-sensor analysis, these two are heavily compared.

4.4.3 Combined Datasets

Combined and Averaged (CAD)

From previous work [5], it was found that when an individual sensor's data is averaged within each class (NRs averaged, PURs averaged, and PARs averaged to form three single class representations–one of each class), each class forms an individual waveform that is easily distinguishable to the human eye (Figure 4.2). Finding this, the combination of pre-averaged data was hoped to produce high classification accuracy for an entire participant instead of just an individual sensor. In the CAD set, data from each sensor is first averaged within each recognition class. To expand on this, a single sensor's data, after pre-processing, is arranged in a number of rows (one row for each set of valid image data, meaning that a "row" is simply the data representation in a table format of a single image shown to participants), and each of the three classes (NR, PUR, and PAR) have their samples averaged row-wise (vertically across every row) to produce a single averaged waveform for each of that sensor's recognition classes (three rows in total). Each set of three rows from each sensor is then combined to form a new dataset of 42 rows (14 sensors $\times$ 3 rows from each sensor). The problem with this method is that it produces a number of rows following the formula

$$(\text{Number Of Sensors}) \times (\text{Number Of Classes}) \tag{4.4}$$

This means that the amount of data you have as a result of this method is determined by the headset you are using. With the Emotiv Epoc headset (used in this experiment), you have 14 sensors. If you were to use the Interaxon MUSE headset [28], you only have five sensors. These are small numbers of sensors, resulting in small sets of data. Increasing the number of classes for classification will help improve

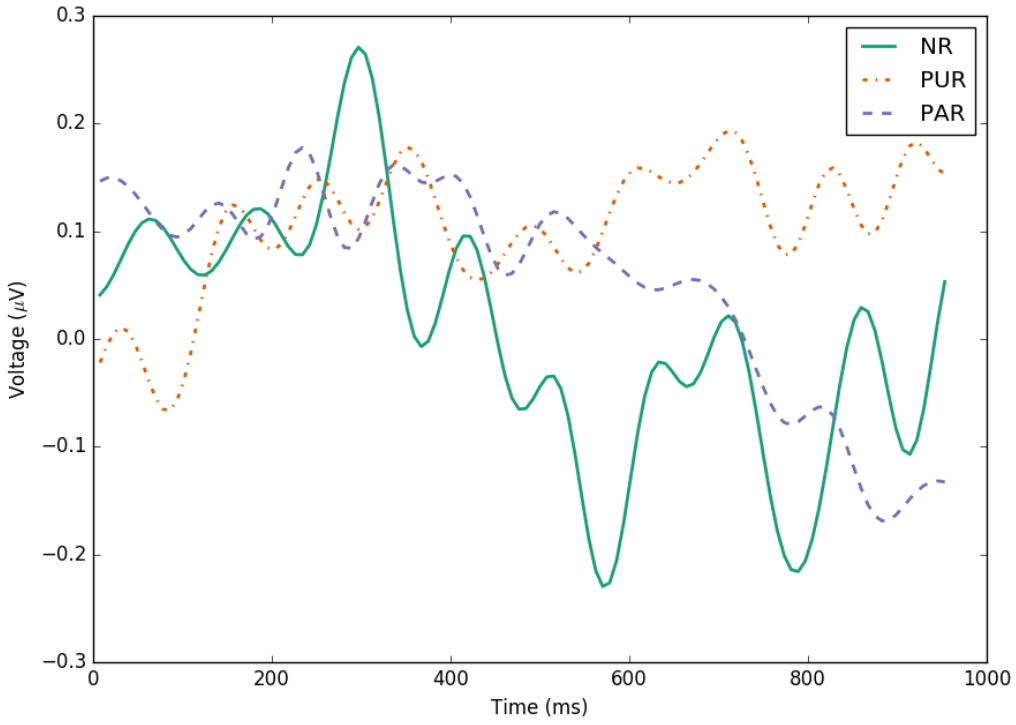

Figure 4.2: Participant P51's averaged AF3 sensor data. NR data not considered for classification purposes, but left in this figure to show the distinct separation of all three classes.

the number of samples of data, but not to the same extent as number of sensors.

Combined (CD)

The CD set is simpler to produce than the CAD set as it is a combination of each sensors' data. Each sensors' entire set of data is taken, one at a time, and added to a global table, resulting in a final set consisting of N number of samples (14 sensors $\times$ number of images in each sensor, averaging 4693 samples per participant). Unlike the CAD method, the CD method does not rely so heavily on number of sensors and classes for number of samples. It follows the formula:

$$(\text{Number Of Sensors}) \times (\text{Number Of Images}) \tag{4.5}$$

Since there are many more images shown across all sensors than sensor count, this produces a far greater sample count for classification, potentially improving statistical significance in the results.

In all datasets, all data from the NR class is removed as this work is primarily interested in detecting unaware facial recognitions. Previous work has shown that NR data is fairly obvious to find whereas unaware and aware data are far closer in terms of shape, so focus was put on classifying these two. This reduced the sample count to 28 and 495 for CAD and CD datasets, respectively.

4.4.4 Combined Dataset Classification Results

Table 4.7 shows the mean F-score across all participants for each classifier and each of the combined datasets, including the differences between the two datasets. These results are shown graphically in Figure 4.3. Shown in bold are the F-scores that were the highest between the CAD and CD dataset. The results shown here are the mean F-

Table 4.7: Average F-scores across all participants for each classifier.

Dataset	Random Forest	Gaussian Naive Bayes	SVC	Decision Tree	K-Nearest Neighbours
CAD	0.736	0.593	0.562	0.662	0.686
CD	**0.748**	**0.611**	**0.635**	**0.672**	**0.752**
Difference	0.012 (+1.63%)	0.018 (+3.04%)	0.073 (+12.99%)	0.010 (+1.51%)	0.66 (+9.62%)

scores across all participants and does not include individual participant results. The mean F-scores across participants is an acceptable value to judge and compare results so reported accuracies here are means, however full participant results are reported in Appendix 1.6. As can be seen, the CD dataset, while generally only being a little bit higher than the CAD dataset, outperformed the CAD dataset in every classifier. With that said, the KNN and SVC classifiers saw an increase of almost 10% and 13%, respectively, which is significantly larger than the 1.63%, 3.04%, and 1.51% increases from the other three. Using a statistical t-test (95% confidence interval), it was found that the difference between the CAD and CD methods for each classifier was statistically significant except for the Gaussian Naive Bayes classifier. All other classifiers (4/5) were found to be statistically significant.

4.5 Classifier Boosting

Boosting in machine learning is the act of combining weak hypotheses about a set of training data from multiple classification trials (T), and combining them to produce a single rule that defines a classification for a set of data [22]. Freund and Schapire show that if a number of predictions that exhibit accuracies of just slightly greater than the random guess (accuracy of 50%), over a number of trials, the errors in the final classification accuracy drops significantly [22].

The Decision Tree and Random Forest classifiers were used in this boosted clas-

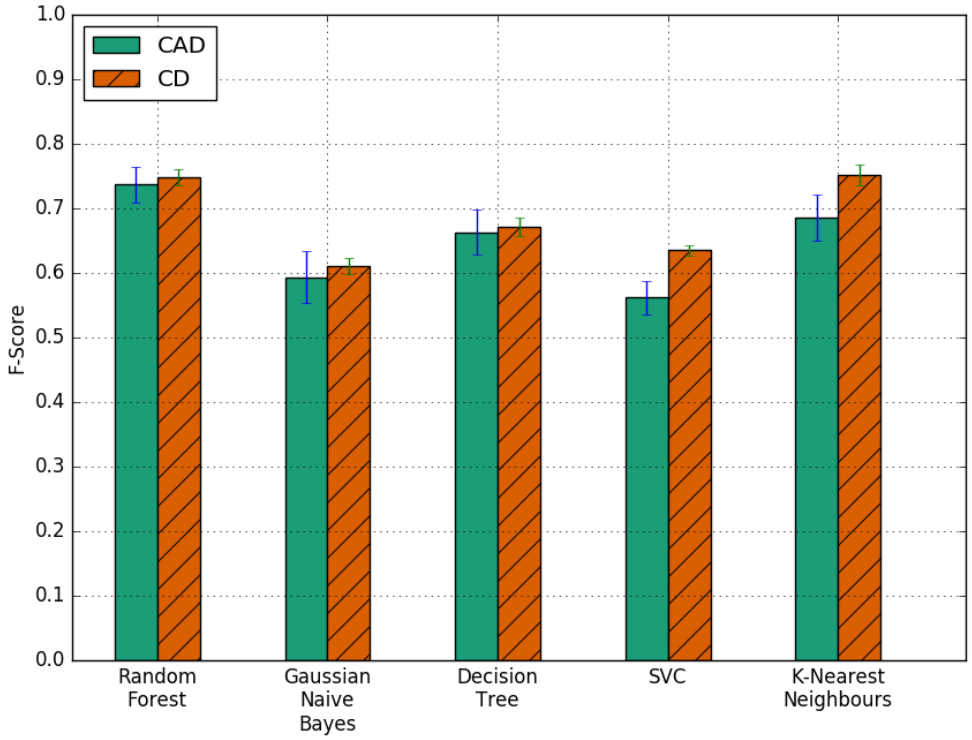

Figure 4.3: Average F-scores for all classifiers from both combined datasets, including standard error bars.

sification method. Only these classifiers of the full set used were boosted as they are considered to be "weak learners" (generally capable of producing guesses at accuracies only marginally higher than random guess levels) [49] and are generally combined (multiple trials of the same classifier having their results used together) in ensemble methods such as this to produce stronger predictions [49]. Each classifier was used in the ADABoostClassifier function (included in Scikit-Learn [47] [40]) as base classifiers, run a total of 11 times each to determine how the performance changes depending on the number of estimators used (number of estimators starting from 1 and increasing by 10, up to 100). For each of these classifications, the same method for data manipulation and pre-processing was used as the previously reported accuracies. This was done to determine if an improvement in classification results could be found using this boosting method over the single use of each of the base classifiers. Only the large CD dataset was used for this method as the set is significantly larger than the CAD set, thus producing a far larger set for the classifiers to be trained on for this boosting method. Freund and Schapire mention that the final accuracy is generally close to that of the greatest accuracy achieved throughout the boosting algorithm (e.g. if three trials produces accuracies of 0.51, 0.52, and 0.58, the final accuracy will generally be close to 0.58) [22]. If, after increasing the number of trials on this experiment's dataset, the variance of results is not large, the end results will not have benefited from the boosting algorithm.

4.5.1 Decision Tree Classifier

The decision tree classifier provided the most interesting and drastic results. Starting at an average F-score of 0.517 ($T = 1$), increasing the number of estimators by nine (up to 10) improved the F-score of the Decision Tree Classifier to 0.646–an improvement of 0.129. This was the largest change as the curve tends to flatten out,

Table 4.8: Average F-scores of boosted Decision Tree classifier.

	Trials										
	1	10	20	30	40	50	60	70	80	90	100
F-Score	0.517	0.646	0.677	0.682	0.694	0.699	0.703	0.702	**0.708**	0.704	0.706

Table 4.9: Average F-scores of boosted Random Forest classifier.

	Trials										
	1	10	20	30	40	50	60	70	80	90	100
F-Score	0.741	0.772	**0.773**	0.768	0.762	0.761	0.764	0.762	0.763	0.759	0.758

much like the shape of a logarithmic curve approaching an asymptote. Table 4.8 contains the average F-scores for this classifier, depicted as a line graph in Figure 4.4. Increasing the number of trials incrementally up to 100, we see the F-score max out at 0.708–an increase in 0.191 over the $T = 1$ value. While not as huge of a jump as the first (1-10), this shows that using the Decision Tree classifier with this combined dataset in a boosted configuration provides an increase in F-score of almost 0.2 (a 37% increase).

4.5.2 Random Forest Classifier

The Random Forest classifier did not seem to fare as well under the same conditions as the Decision Tree classifier. Using the same number of trials at each step, the Random Forest tended to produce a much more linear graph (Figure 4.5) where increases were not as profound. Average F-scores over all trials can be found in Table 4.9.

It was noted that the ADABoost function would not always run through all T trials of the Random Forest classifier. According to the Scikit-Learn documentation for the ADABoost classifier function, if the base classifier fits the data perfectly in a trial, the rest of the trials are not run and the function is stopped early [47]. Continuing with training could lead to the over-fitting of the classifier. This changed on a participant-

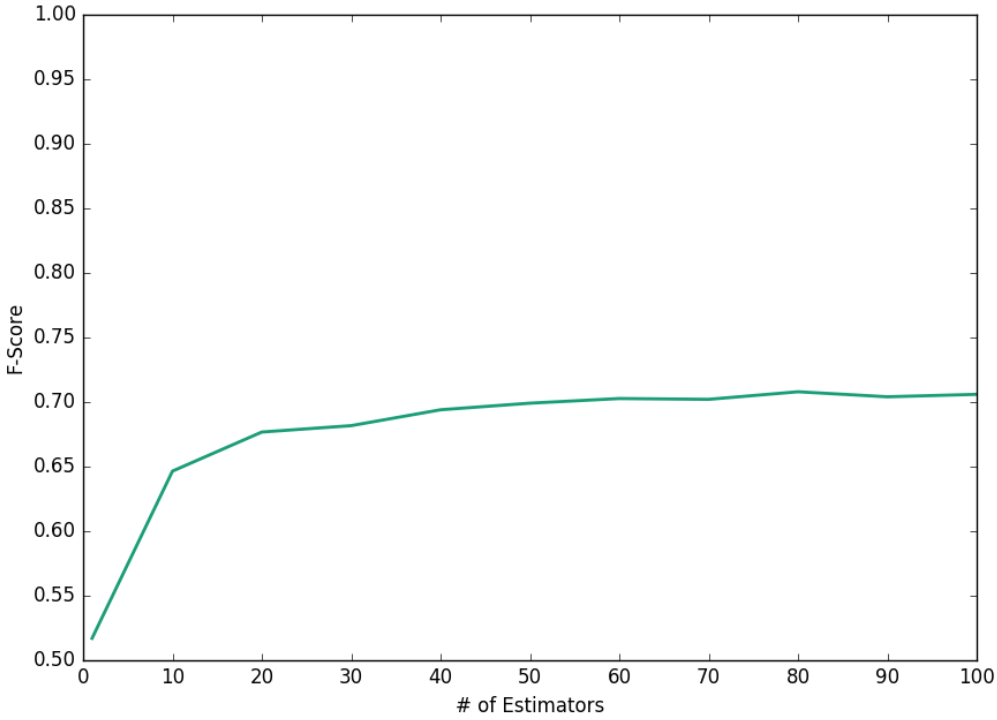

Figure 4.4: Average F-scores from the Decision Tree classifier with an increasing number of estimators used in boosting.

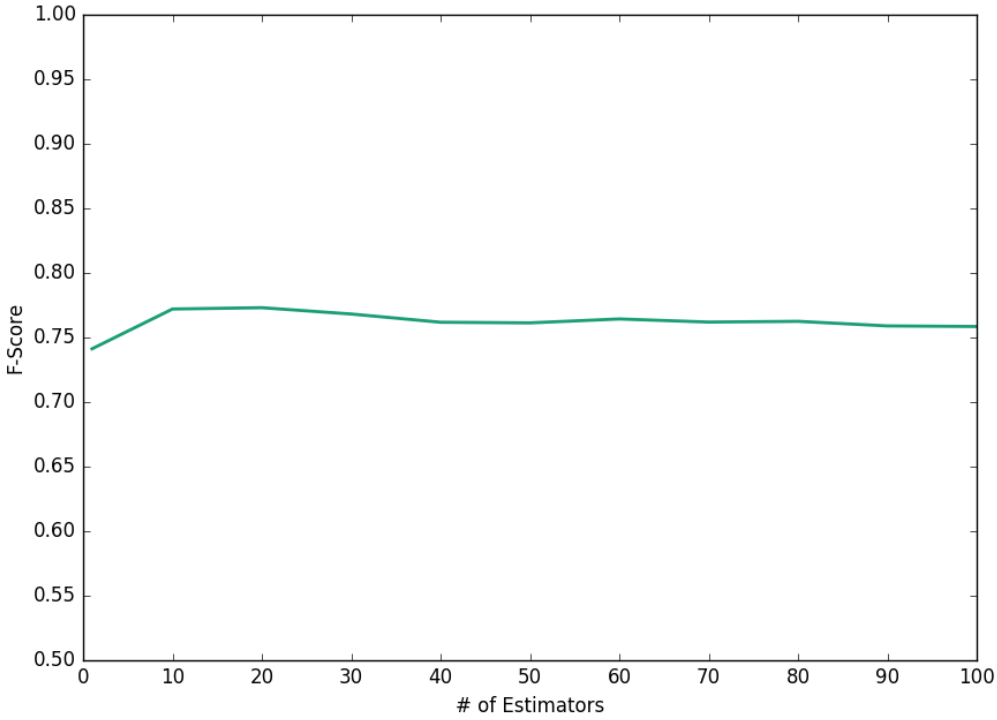

Figure 4.5: Average F-scores from the Random Forest classifier with an increasing number of estimators used in boosting.

Table 4.10: Boosted vs. Un-Boosted results of the Random Forest and Decision Tree classifiers.

	Random Forest	**Decision Tree**
Boosted (highest F-score)	0.773	0.708
Un-Boosted	0.748	0.672
Difference	+0.025	+0.036

to-participant basis depending on how well an individual participant's data could be fit by the classifier.

4.5.3 Boosted vs. Un-Boosted

For a brief comparison, the Decision Tree and Random Forest classifiers have their results compared in Table 4.10. In this table, the highest F-score achieved is shown for each classifier. As can be seen, the boosted classifiers achieve greater accuracies than the individual single-run classifiers (Table 4.7). With that said, the differences are fairly small (+0.025 and +0.036 for Random Forest and Decision Tree, respectively).

4.6 Classifier Bagging

Much like boosting, bagging uses a number of estimators in combination to attempt to provide a more accurate classification output. Each estimator uses a sub-set of the global training data to learn the data, and then, in the case of predicting a class, the estimators vote on the output [7]. Breiman et al. [7] find that if the changes in the learning dataset result in small changes in the predictions, then the results of sub-set prediction will be close to the results of the global prediction. On the other hand, great improvement in classification is found when a small change in the learning dataset results in large changes in the prediction.

The Decision Tree and Random Forest classifiers were used in this bagged clas-

Table 4.11: Average F-scores of bagged Decision Tree classifier.

	Trials										
	1	10	20	30	40	50	60	70	80	90	100
F-Score	0.648	0.734	0.760	0.757	0.757	0.756	0.760	0.760	**0.762**	**0.762**	0.761

sification method to allow for comparison to the boosting method. Each classifier was used in the BaggingClassifier function (included in Scikit-Learn [48] [40]) as base classifiers, run a total of 11 times to determine how the performance changes depending on the number of estimators used (number of estimators starting from 1 and increasing by 10, up to 100). For each of these classifications, the same method for data manipulation and pre-processing was used as the previously reported accuracies. Again, only the large combined dataset was used for this method as the set is significantly larger than the combined and averaged set, thus producing a far larger set for the classifiers to be trained on for this bagging method.

4.6.1 Decision Tree Classifier

Like with Boosting, the Decision Tree classifier again followed the shape of a rough logarithmic curve (Figure 4.6). If we consider the highest average F-score found (0.762 at 80 and 90 trials), this method resulted in an increase of almost 18% over the singe-estimator classification F-score.

4.6.2 Random Forest Classifier

The Random Forest classifier followed a more horizontally-stretched logarithmic curve shape than the Decision Tree classifier, but did not appear as flat in general when compared with the curve produced by the Boosting method. The highest F-score achieved was 0.768 at 100 trials–an increase of about 10%. While not as significant

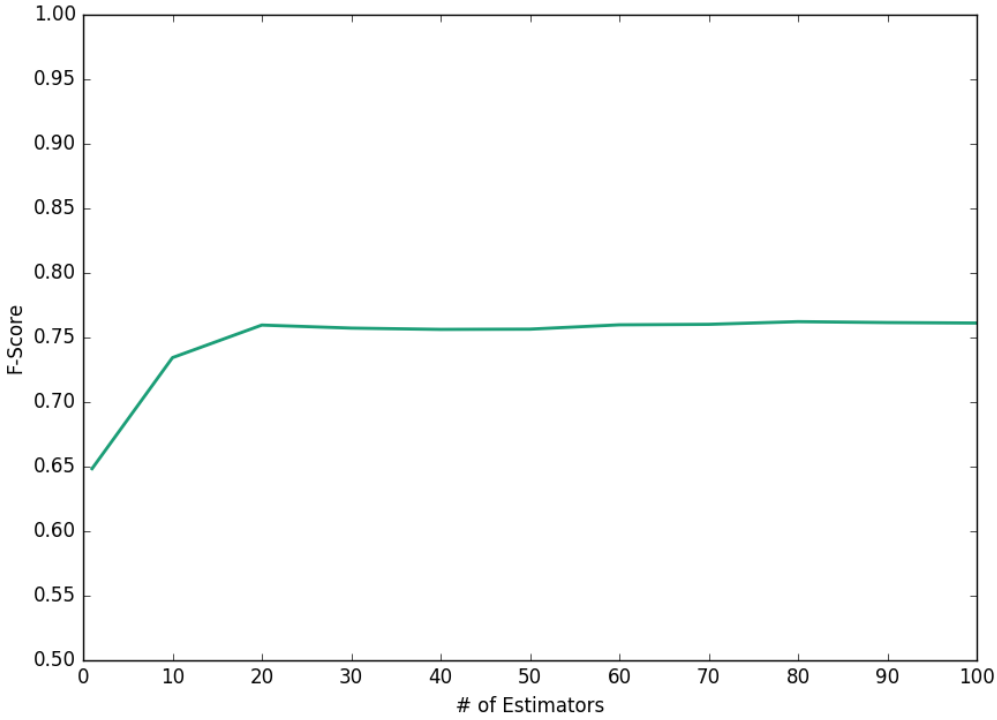

Figure 4.6: Average F-scores from the Decision Tree classifier with an increasing number of estimators used in bagging.

Table 4.12: Average F-scores of bagged Random Forest classifier.

	Trials										
	1	10	20	30	40	50	60	70	80	90	100
F-Score	0.698	0.752	0.759	0.762	0.762	0.763	0.765	0.764	0.766	0.766	**0.768**

Table 4.13: Bagged vs. Un-Bagged results of the Random Forest and Decision Tree classifiers.

	Random Forest	**Decision Tree**
Bagged (highest F-score)	0.768	0.762
Un-Bagged	0.748	0.672
Difference	+0.020	+0.090

as the Decision Tree classifier (~18% vs. ~10%), the increase is still fairly substantial and would assist in improving the general classification accuracies of unaware facial recognition.

4.6.3 Bagged vs. Un-Bagged

Table 4.13 shows a brief comparison between the bagged and un-bagged results (single-estimator vs. lone classifier output (Table 4.7). In both cases of the Random Forest classifier and Decision Tree classifier, we see improvement in average F-score. At its peak accuracy, the Random Forest classifier produces an increase in F-score of 0.02 (~2.7% - achieved at 100 trials), while the Decision Tree classifier has an increase in F-score of 0.09 (~13.4% - achieved at 80 trials). While these numbers do not seem the most significant, they are improvements to the results produced by the standard classification of the CD set.

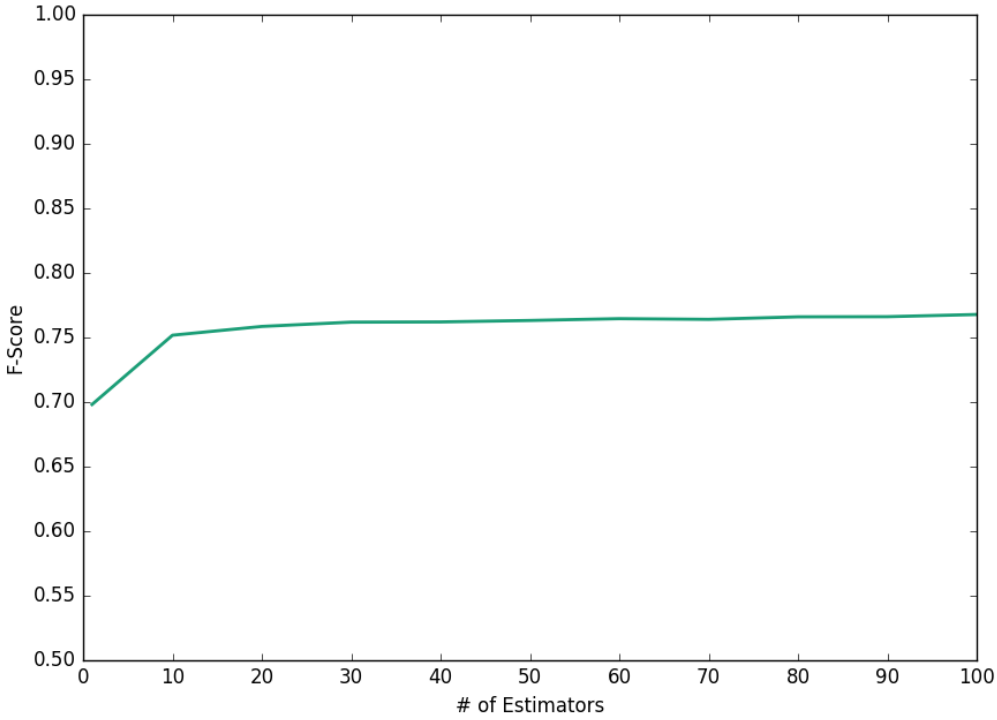

Figure 4.7: Average F-scores from the Random Forest classifier with an increasing number of estimators used in bagging.

Table 4.14: Boosting vs. Bagging results of the Random Forest and Decision Tree classifiers. Base classifier improvement values are in brackets beside each classifiers' average F-score.

	Random Forest	**Decision Tree**
Boosted	0.773 (+0.025)	0.708 (+0.036)
Bagged	0.768 (+0.020)	0.762 (+0.090)
Difference	+0.005	+0.054

4.7 Boosting vs. Bagging

Table 4.14 shows a numerical comparison between the Random Forest and Decision Tree classifiers. Judging based purely on F-score improvement over stock classifiers, the Random Forest classifier performs best on the CD set for unaware facial recognition classification when using the boosting method while the Decision Tree classifier performs better using the bagging method. Computational time to produce these accuracies is not considered in this comparison, but depending on the application, may become important.

4.8 Individual Sensor vs. Combined Sensors

It is hard to do a direct comparison between the individual sensor datasets and combined sensor datasets due to number of scores produced for each participant (Individual sensor = (# of sensors) $\times$ (# of classifiers) vs. Combined sensor = 1 $\times$ (# of classifiers)), which produces two-dimensional data for each participant in the individual sensor dataset, and one-dimensional data for each participant in the combined sensor datasets. Along with this, because the low number of samples in each of the individual sensors' dataset, there were quite often errors in classification resulting in a nullified classification. While the direct comparison here may not be appropriate for these reasons, the fact that there were so many errors (an average

of five per participant) helps highlight some of the issues that the individual sensor dataset faces when applied to classifying facial recognitions using this method.

4.9 Neural Networks

In addition to the previous methods of analysis, a neural network was built and trained on the CD dataset. The CD dataset was chosen as it has the largest number of samples for training, hopefully producing the best results and not having errors due to such few samples in the set. Following the work of Subasi and Ercelebi [57], in which they used a multi-layer neural network to classify EEG signals, the input layer was immediately followed by a second layer of 21 neurons, followed by the output layer for a fairly shallow neural network. From the TFLearn Python library [58], a Deep Neural Network function was used to classify using the previously-built neuron layers. After testing, highest output F-scores were achieved using a hyperbolic tangent activation function [60] over the default linear model [59] [61]. This design on the CD dataset resulted in an average classification accuracy of 0.668.

Table 4.15 contains a comparison between the results of the neural network with the results from the standard CD dataset classification. The neural network outperformed two of the five classifiers used on the CD dataset. Full individual participant results are shown in Appendix 1.7. It can be seen that, over all participants, accuracies climb as high as almost 90% and include six participants that had classification F-scores above 0.8.

Utilizing neural networks was not the primary focus of this thesis, but holds great promise considering the results and simple implementation they resulted from. Considering that the results found here with limited exploration are able to outperform two of the previously-used classifiers, even greater results may be found after

Table 4.15: Neural network classification F-score vs. previous classifier F-scores.

	Non-NN	NN	Difference
Random Forest vs. NN	0.748	0.668	−0.080
Gaussian Naive Bayes vs. NN	0.611	0.668	0.057
SVC vs. NN	0.635	0.668	0.030
Decision Tree vs. NN	0.672	0.668	−0.004
K-Nearest Neighbours vs. NN	0.752	0.668	−0.084

further manipulation of neural network structure and design. Future work may put more emphasis on using neural networks for classifying unaware facial recognition versus the more traditional classifiers used in this work.

Chapter 5

Discussion

5.1 Outline

The results of the experiment have been outlined previously in Chapter 4, so a discussion of the results and the experiment as a whole is conducted in this chapter. A look into the classifiers used and the implications of each on the EEG data is done to assess the benefits or drawbacks of each as well as any possible areas that a classifier may struggle or have success with. A summary of each dataset method is explored and a comparison is done between them, exploring again, benefits and drawbacks of each methods and how they may explain the results found. Further, a discussion is had about the recording apparatus in which some issues with the experiment and general recording that may have contributed to the results are highlighted. Finally, a look into some of the applications and implications of this work on the field are discussed.

5.2 Classifiers

A total of five different classifiers were used (Neural network and ensemble methods not included in this count), each on all three sets of data (minus SVC for individual classification - see Section 4.2). Each classifier was picked from the available classifiers that the Scikit-Learn Python library offered and used in an out-of-the-box configuration without modification via parameters. Depending on the classifier, they may perform better or worse on specific types or arrangements of data, which may lend some insight to the results obtained in this experiment.

5.2.1 Decision Tree

The Decision Tree classifier builds a tree by recursively splitting an attribute, basing each split on some function of that attribute, and continuing to split the attributes until some sort of ending point is reached where the final nodes ("leaf" nodes) correspond to one of the classification outcomes [26]. For testing, each sample traverses the previously-built tree until it reaches one of the leaf nodes where it assigns that sample to a class. In the case of the data here, these classification outcomes are either a PUR or PAR value. Inese et al. [26] mention that, because a slight variance in training data can produce vast differences in classifier models, the decision tree classifier is considered to be unstable and would perform well in an ensemble method, such as boosting or bagging. This may help identify why an improvement in classifier performance was found over the base classifier when used in the boosting and bagging configurations (see sections 4.5 and 4.6).

5.2.2 Random Forest

The Random Forest classifier is a combination of tree models such as the Decision Tree classifier previously mentioned, and uses samples and features chosen at random to build the decision trees [43] [3]. Since the random forest classifier is an ensemble classifier that utilizes multiple decision trees for classification, this may be why a classification accuracy improvement over the standard decision tree classifier in these datasets can be seen (0.737 vs. 0.662 for CAD and 0.748 vs. 0.672 for CD). Pu et al. [43] mention that due to the two sources of randomness in the algorithm (randomly chosen inputs and features) the Random Forest classifier tends to be able to handle large numbers of features well, which may be why some of the highest classification accuracies are seen from this classifier.

5.2.3 Gaussian Naive Bayes

The Naive Bayes algorithm is a fairly straight-forward method of classification that makes use of a comparison between the probabilities of each class given the data [50]. The Gaussian Naive Bayes classifier assumes that the data being analyzed follows a Gaussian/normal distribution for probability estimates and classification, thus being potentially more reliant on the input data than other classifiers [50]. It is known for its perspective of features within the feature set in that it assumes each feature is independent from another [50]. While the data used in this experiment is of a temporal nature (each sample can be reconstructed into a waveform) thus having each feature reliant on its surrounding features to reproduce the recorded signal, each feature within the signal can be taken out of context as the waveform itself is not important for classification, but rather, the values at specific time intervals are the important parts, so the application of the Naive Bayes algorithm on this dataset may

not have as much of a negative impact as another set might.

For each participant, a test for normality was done on each feature in the classification data to determine if the data for any given participant followed a normal distribution. In general, the majority of participant data returned p-values that were very low, thus not following normal distribution. While some features showed more normal distribution, these were generally only three or four features for an individual sensor out of the hundreds of pieces of data tested. Given the non-normal input data, it may explain why the Gaussian Naive Bayes classifier performs poorly on all three datasets (individual: ~0.50, CAD: ~0.59, CD: ~0.61).

5.2.4 SVC

The SVC classifier (Support Vector Classification) from the Scikit-Learn library [54] is a Support Vector Machine (SVM) based on LIBSVM, a popular library for the implementation of SVMs, which support a variety of tasks including classification, regression, and others [9]. The goal of an SVM is to determine a "hyperplane" (the general term for a plane or division in high-dimensional space) which separates two classes of data, and to decide on it in such a way that it maximizes the distance between each of the classes, providing the best possible classification prediction [53]. According to Mountakis et al. [35] and Noble [39], SVMs tend to perform better on datasets with few training samples. Also, they do not assume datasets follow a normal distribution, so this may help to explain why the SVC classifier out-performs the Gaussian Naive Bayes classifier. With that said, it is still found to under-perform compared to the Random Forest, Decision Tree, and K-Nearest Neighbours classifiers by a large margin (see Table 4.7).

5.2.5 K-Nearest Neighbours

The K-Nearest Neighbours (KNN) classifier is another fairly straight-forward classification algorithm. The KNN classifier stores training data for comparison when testing new data. The new data is then compared to the old data within its n-dimensional space, and the surrounding training points are used to determine which class the new data belongs to based on distance to the points in the training set [51]. The value K is used to determine how many of a new sample's nearest training points are for voting. While easy to visualize in two-dimensional or three-dimensional space, it is basically impossible for a human to visualize the 121 dimensions of the average sample in a participant's dataset. This algorithm looks for spatial groupings of points to make its decision on which class a sample belongs to, so the more closely a new point appears to a majority of training points, the more likely it will be classified the same way as the closest training points are. This means that if a number of training points are grouped incorrectly with the wrong class, then the new point, while being of the opposite class, has a good chance of being misclassified. The authors of Scikit-Learn mention that a larger K value tends to reduce the impact of noisy data on classification as it has more neighbours to compare with [51]. Given this, since the default value for K is 5 [52], increasing the K value may allow for better classification accuracies in the datasets, particularly the CD dataset where there are many more sample points to compare to.

5.3 Dataset Methods

Since the purpose of considering three datasets in the first place was to be able to compare and determine if using a modified dataset could be used to classify unaware facial recognitions well, a brief discussion of the findings is presented here.

One of the main issues with the individual sensor datasets are that not all sensors contain all data due to pre-processing and issues with recording. For example, a participant's AF3 sensor may have the full 20 + 20 samples of PUR and PAR data from the experiment, but the P7 sensor may have lost touch with the participant's scalp during recording and only managed to record 5 + 5 samples of PUR and PAR data. Even if the full 40 samples of data were recorded for all sensors for all participants, 40 samples is still a very small number of samples to be trained on, and that is not considering the 60-40 split for training and testing, further reducing the sample count for either portions. This consideration is one of the reasons why the combined datasets were considered as they not only modify the recorded data and attempt to improve classification accuracy that way, but they also, especially in the case of the CD dataset, increase the number of samples for training and testing in classification. Because of this low sample count, the results from individual sensor classification vary widely, leading to the suggestion that these results should not be relied on for statistical significance. With that said, it is an interesting insight into some of the potential issues both with the practice of recording using consumer-grade devices, and experimental design in the first place. While increasing the number of PUR and PAR images in the experiment may lead to more data for classification, it also increases the likelihood that a PUR image is noticed by the participant, inadvertently becoming a PAR image which would be incorrectly classified.

The CAD dataset falls into some of the same issues as the individual sensor datasets do. When each sensor's data is first averaged into two PUR and PAR samples and appended with the averaged samples from all other sensors, the number of samples for classification remains constant (see Section 4.4.3) because it does not rely directly on the full number of samples in each individual sensor. Regardless, the number of samples in this dataset ends up being only 28. This number is still larger

(relatively speaking) than each individual sensor's theoretical maximum of 20 samples (10 from both PUR and PAR), however it still is not very many when considering its use in classification. Another point to note is that if this method were to be used in some sort of future application, recording a single sample (e.g. viewing a single image) and attempting to classify based on it may not result in a high accuracy. Since the CAD set relies on the full experiment of recorded image samples and intra-class averaging, any future application would require a similar setup in order to be able to average the recognition classes. This may be considered a limitation of this method.

The CD dataset addresses part of the issues with the CAD and individual datasets by having a significantly larger number of samples for classification (see Section 4.4.3). While this increased count may prove to be more statistically safe for classification train/test splitting and results, it does not make use of the intra-class averaging that the CAD dataset uses. This averaging produces a single wave for each class in each sensor and is the average representation of each class, which, as seen from previous work in unaware facial recognition [5], provides easily identifiable signals from each recognition class. In the CD dataset, no averaging is done, thus feeding a large number of un-averaged recognition class data into the classifier for training and testing. Since some of the un-averaged waves may appear to be more similar to the other recognition class or other sensors' data, it may be a source of confusion for the classifiers, potentially providing weaker classification than a large set of averaged data.

Building a single set that combines the advantages of both the CAD and CD dataset would be optimal, but the very nature of the construction of each set prohibits this combination. Another source of future work may be exploring methods for combining the two datasets and providing more averaged data with greater numbers for classification in hopes of improving both the amount of data for training and the

classification results.

Despite the issues in both the CAD and CD dataset, using ensemble methods to try to improve the classification accuracy of each proved to be very useful. The boosting and bagging methods (Sections 4.5 and 4.6) were able to generally increase the classification accuracies of the CD dataset without further modifying the data within the set. The only apparent cost to this increase was the added processing time, which may be considered as not applicable in the context of trying to find methods to improve unaware facial recognition. In a real-world application, processing time may be an important factor, but in this situation it does not need to be considered. Both the boosting and bagging method improved classification accuracies for both the Decision Tree and Random Forest classifiers, showing that ensemble methods, especially when combined with a potential future optimization of classification algorithms, may provide far greater accuracies than base classifiers alone.

Using a statistical t-test (95% confidence interval), it was found that the difference between the CAD and CD methods for each classifier was statistically significant except for the Gaussian Naive Bayes classifier. All other classifiers (4/5) were found to be statistically significant. Since the individual sensor dataset was, by definition, comprised of individual sensors and covered each sensor for each participant, whereas the CAD and CD datasets were a combination of sensors, direct comparison between the individual and combined sets is not as applicable.

5.4 "The Question"

One of the big issues with studying anything on the unaware side of reactions is the lingering question "was it actually unaware?" This experiment (Chapter 3) had to perform a balancing act between the number of images that would essentially be used

as filler and discarded (NR images), and the images that were intended to be learned by participants, both at an aware level (PAR) and unaware level (PUR). In phase one (see Section 3.5.1), the images to be used as unaware recognitions were learned. The challenge here was using enough images so that adequate analysis could be done, but not too many unaware images so that participants did not recognize them too often, thus becoming aware recognitions. One of the assumptions in this experiment is that each image is explicitly pre-classified (by the author, meaning that images were shown at specific times and in specific ways to promote learning at the required levels) and does not change classification. This means that an image that was tagged as unaware had to stay as an unaware recognition for the participant throughout the experiment. If they viewed and learned the image too much and it became an aware recognition, the corresponding data would potentially look very similar to PAR data, thus being classified as PAR data, and producing a False Negative resulting in a corrupted F-score.

Based on previous work in unaware facial recognition [4], it was found that the three recognition classes (NR, PUR, PAR) after intra-class averaging produced very distinct signals that could be easily differentiated by the human eye without the assistance of machine learning. Since this experiment made use of a very similar methodology to the experiment in that work, it is reasonable to expect that the design of this experiment was adequate to safely produce the three recognition classes without too much accidental learning (PUR as PAR, PAR as PUR, etc.). With that said, a handful (unrecorded numbers) of participants mentioned after the third phase that that they noticed a few repeating images in the first phase. While this is not the intent of the experimental design, the two-day experiment design allowed for these sorts of small issues in the first day as the assumption is that participants would not remember the majority of the faces shown to them in phase one at an aware level,

but rather at an unaware level, which is the intent of the first phase.

5.5 Recording Apparatus

Something important to consider when reviewing the results of this data is the method of recording, or, more specifically, the device used in the data recording. As mentioned in Section 3.4, the Emotiv Epoc BCI headset was used for recording participant EEG data during the experiment. For the first dataset attempted (individual sensor - Section 4.4.1), each sensor is considered by itself rather than in combination with other sensors. The Epoc headset has a total of 14 sensors which are placed around the scalp, meaning that many sensors are quite close to each other. Because the sensors are sometimes very close to each other, it is unclear how similar the signal from one sensor is compared to the signal from the next closest sensor. How similar these signals are may not have too great of an impact on the individual sensor dataset, but when considered in combination (CAD and CD - Sections 4.4.3 and 4.4.3, respectively), data from two sensors might be very similar, potentially strengthening or weakening results. Another issue the "consumer-gradeness" of the headset and its comparison with more lab/medical-grade devices. Duvinage et al. [12] ran an experiment to analyze the performance differences for P300-based applications between the Emotiv Epoc BCI device and the ANT acquisition system using the Waveguard cap consisting of 128 sensors [37]. They note that the Epoc under-performs compared to the ANT system with Waveguard cap (which is to be expected), and find that the Epoc should be used only for non-critical applications such as games or other miscellaneous functions and to use more lab/medical-grade devices for serious medical purposes. This suggests that more modern or professional devices may perform better for detecting unaware facial recognitions due to its higher quality recording

and additional sensor capabilities. It would be interesting to use a lab/medical-grade EEG cap in future work to see what the results look like when applied to unaware facial recognition detection and classification compared to the current results obtained using the consumer-grade Epoc.

Another area of problem that arose with the Emotiv Epoc headset has to do with the participants themselves. Before the experiment, all participants were asked to try to remain as still as possible during the experiment. While many participants were able to stay fairly still, a number of participants still moved around, doing thing such as repositioning their legs, scratching an itch, or changing their posture, and almost all participants moved around at least a little bit. These sorts of movements contributed to the artifacts and general poor signal quality recorded by the headset, which is tough to avoid and seems to be a general issue with recording BCI data (mentioned in previous work [4]). If the experiment were to be done again, adding some sort of chin/head rest for participants to rest on during the experiment may help to promote still sitting, however this solution may be uncomfortable for the participants. This is speculation, however.

Similar to the previous point, but not entirely to do with participants, it was found that, for a handful of participants and through previous experiments, the connection from the headset to the computer would drop, recording no data for the duration of the disconnection. It would generally reconnect itself back to the computer after a few seconds, but during this time no data was recorded, thus missing important image data. Since it was important that participants had not seen the images on the screen before, the experiment could not be stopped and restarted once the headset had reconnected. Unfortunately, the Epoc has only wireless connectivity, so a direct cable to ensure 100% connectivity is not possible. This is an advantage that lab/medical grade devices have over consumer-grade devices as they are not designed to be as ca-

sual or movement friendly whereas the consumer-grade devices play more on modern connectivity trends to reduce clutter or bulk from wires or larger attachments.

5.6 Implications

This thesis has shown that yes, unaware facial recognitions can be recorded and extracted from subject EEG data with fairly high accuracies using consumer-grade BCI devices, even using out-of-the-box machine learning classifiers, which makes for rapid classification at near on-demand levels. The fact that this could be implemented in a cost-effective way and with low amounts of time, the techniques explored in this thesis could be implemented into more application-based scenarios.

One potential application for something like this is in law enforcement. Police line-up procedures have been shown to be somewhat biased [33] [8] [34], which could potentially lead to false or inadvertently incorrect selection of criminal perpetrators. Using unaware facial recognitions, even if the participant only saw the perpetrator for a brief amount of time during the incident, they may be able to determine if an unaware facial recognition takes place while viewing the line-up. It is unclear as of yet if bias is witnessed within subject EEG signals, so this application may or may not work depending on that finding.

While not yet perfect, these techniques could pave the way for future researchers to further explore unaware facial recognition and EEG processing to improve existing results.

Chapter 6

Conclusions and Future Work

Through the experimentation completed in this thesis, the results of determining whether or not unaware facial recognitions can be detected and classified with greater accuracies than previously found using consumer-grade BCI devices and out-of-the-box machine learning tools are described, and whether or not previous findings can be improved upon. From the work done here and based on the research objectives and hypotheses provided in Chapter 1, the following conclusions have been drawn from this work:

- First, consumer-grade devices, and more specifically the Emotiv Epoc, can be used to accurately record EEG data from participants with enough accuracy to detect each of the three classes of recognition outlined previously (no recognition, unaware recognition, and aware recognition). This outcome was expected from this experiment as previous work conducted [5] showed, using various methods of analysis, that each class could be accurately found.

- Second, using out-of-the-box machine learning tools was able to enhance and increase speed of large-scale automated classification of participant EEG data

recorded from the experiment, however at lower levels of accuracy than expected. Previous work [5] suggested that using intra-class averaging could produce very distinct signals for each recognition class, potentially giving very high classification accuracies, but it was found after analysis that the dataset which had not been averaged tended to produce the highest accuracies.

- Third, the combination of multiple sensors tends to outperform single-sensor classification, both in terms of classification accuracy and classification quality (far less errors and issues with the combined datasets).

6.1 Findings

Each participant's recorded data was modified in three different ways to determine which method performs the best in the detection and classification of unaware facial recognitions. The first method, dubbed the "individual dataset", uses each sensor's recorded data individually and attempts to classify unaware facial recognitions using only a single sensor's data. This results in 14 F-scores for each sensor, for each classifier used (four classifiers). The next two methods make use of a combination of sensors. The first of the two–the Combined and Averaged dataset (CAD)–first averaged each recognition class within a sensor, then combined each sensor's averaged data together to form a single dataset. The second combined dataset was simply dubbed the "Combined Dataset" (CD) as it took all sensors and combined their data without any averaging, resulting in a fairly large dataset with many samples. The individual dataset consisting of single-sensor classifications was found, in general, to perform worse than the combined datasets. The combined datasets (CAD and CD sets) tended to perform better than the individual datasets in general, however the CD dataset was found to perform better than the other two over-all with greater

F-scores for each classifier tested. Some potential reasoning for these results may have to do with the size of the datasets. In both the individual and CAD datasets, very few samples were in each set, giving the classifiers very little data to train on. The CD dataset had a much larger sample count, which gave the classifiers far more samples to train on and may have improved the classification accuracies by being able to more accurately determine the differences between each of the recognition classes.

Upon finding base-classifier levels of accuracies, a number of these classifiers were used in ensemble methods, namely Boosting and Bagging, to improve single-trial classification performance. Using a Boosting method, improvements of approximately 3.34% and 5.35% for the Random Forest and Decision Tree classifiers, respectively. For the Bagging method, improvements of approximately 2.67% and 13.39%, respectively. This shows that using the base out-of-the-box classifiers in these ensemble methods allows for a further increase in output F-scores for unaware facial recognition classification, providing additional methods for this detection with minimal additional modification to datasets or classification methods.

6.2 Contributions

The work discussed throughout this thesis has shown the possibility of using consumer-grade BCI devices and modern machine learning techniques to classify unaware facial recognitions. Through this, three different methods have been compared–each with their own pros and cons–to determine the best method of classifying these unaware facial recognitions with the highest accuracies. To summarize, the following have been contributed through this thesis:

- Three methods of data manipulation: Three methods for data manipulation have been developed that produces three different sets of data for classification.

Each dataset takes a unique form and provides its own benefits and drawbacks in terms of ease and accuracy in classification. Through these three methods, an in-depth comparison between their structure, their performance, and the implications of each to gain the best classification accuracies have been provided.

- Unaware facial recognition classification improvements: Using the previously-mentioned datasets, previous efforts in classifying unaware facial recognitions (see Section 2) have been improved, contributing these methods as a new stepping stone for future research in unaware facial recognition.

- Applications for consumer-grade equipment: Consumer-grade hardware has been used in this research and applied to the field of facial recognition to provide a more cost-effective and accessible approach to BCI research.

6.3 Future Work

Throughout this thesis there are a number of areas that are explicitly not explored due to scope and scale challenges, but would make great extensions to this work. The following sections describe areas that were generally not included in the work presented here, but would make interesting subjects for future work to be conducted upon.

6.3.1 Generality

One of the first areas that could be explored is the idea of generality across participants. All analysis done in this thesis is done within each participant and never combines participant data for classification purposes–only output results to get a more broad idea of how classifiers are performing. This leaves the opportunity to

potentially combine all participant data and attempt classification on this far larger set of data. Being the most naive approach of simply combining the data and running classification, this may or may not provide useful results, but approaching the problem with new methods of pre-processing or analysis may provide evidence that, given a large enough pool of participant data, a general hypothesis could be made covering inter-participant classification.

6.3.2 Classifier Optimization

Another area that is touched on, but not very deeply, is the classifiers used. This work focuses on using out-of-the-box classifiers to determine performance on the three datasets demonstrated in the above sections, but very little/no optimization was done on classifier parameters to try to improve classification from a classifier optimization point-of-view. Even fairly basic and naive methods for testing and improving classification accuracies such as a brute-force grid search of all parameters may reveal improvements for these datasets. Additional literature review may help provide further insight into the workings of each classifier, providing a greater understanding of how to improve achieved accuracies.

6.3.3 Boosting/Bagging Optimal Estimators

The two sections discussing boosting and bagging methods for improving classification rate had a look at determining the number of estimators involved that give the greatest results. It is not in the scope of this work to try to determine a method for producing the most optimal number of estimators, but future work could be done to identify the factors that go into this and to see the best performance per estimator count for each method and each classifier.

6.3.4 Input Datasets

As mentioned in Section 5.3, each dataset used in this thesis has its own positives and negatives with regards to producing more accurate classifications. The CAD dataset provides averaged data for each sensor to improve combined accuracy, but has a very low number of samples for classification whereas the CD dataset has the large sample count that the CAD dataset lacks, but the data is messier than the CAD dataset and may be a source of confusion in classification. It would be optimal to have a dataset that has the benefit of large sample size as well as the cleaner data, but without any of the drawbacks. This may not be possible, but future work could explore methods of manipulating data to produce the clean, averaged data from the CAD set, but with the large number of samples from the CD dataset.

6.3.5 Neural Networks

A brief analysis, comparison to previously used classification methods, and discussion of the application of using neural networks for classifying unaware facial recognition data was had in Section 4.9. This thesis does not explore using neural networks very deeply and was used only as a brief comparison. Given the optimistic results, future work could further the use of neural networks in classifying unaware facial recognition, potentially surpassing the classification accuracies shown from the classification algorithms discussed in the analysis in Chapter 4.

6.4 Conclusion

Consumer-grade BCI devices are becoming more and more ubiquitous, and the applications of these devices need not be only for specific groups such as casual consumers, low-budget researchers, hobbyists, or other users who may be looking for BCI

equipment at consumer-grade levels. As time goes on, new BCI devices will become available with less invasiveness, higher quality of recordings, and a more application-focused design, leading to new applications and uses for a broad spectrum of users. Consumer-grade devices may also exist to fill a gap in the BCI field where expensive lab-/medical-grade equipment can't be used or where they don't make sense to be used such as in a mobile lab environment, field research, or the more modern application of utilizing wireless EEG devices for day-to-day tasks that expensive and complicated recording systems would not lend themselves very well to.

To close, a statement which encapsulates and describes the work completed here in this thesis:

> The utilization and combination of consumer-grade BCI devices, modern machine learning techniques, and new methods for dataset manipulation provide fairly accurate detection and classification of unaware facial recognitions, and can improve on previous efforts to detect unaware facial recognition.

Though experimentation and data analysis, it has been found that consumer-grade BCI devices can be utilized to detect and record unaware facial recognitions with acceptable accuracy, and that modern machine learning tools can be applied to the task of classifying these states of recognition. While not as accurate as originally hypothesized, this work has provided acceptable accuracies to at least act as a proof-of-concept, and may provide a strong stepping stone for future researchers to build on and improve techniques for classifying both aware and unaware facial recognition.

Bibliography

[1] Abadi, M., Agarwal, A., Barham, P., Brevdo, E., Chen, Z., Citro, C., Corrado, G. S., Davis, A., Dean, J., Devin, M., Ghemawat, S., Goodfellow, I., Harp, A., Irving, G., Isard, M., Jia, Y., Jozefowicz, R., Kaiser, L., Kudlur, M., Levenberg, J., Mané, D., Monga, R., Moore, S., Murray, D., Olah, C., Schuster, M., Shlens, J., Steiner, B., Sutskever, I., Talwar, K., Tucker, P., Vanhoucke, V., Vasudevan, V., Viégas, F., Vinyals, O., Warden, P., Wattenberg, M., Wicke, M., Yu, Y., and Zheng, X. Tensorflow: Large-scale machine learning on heterogeneous systems, 2015. Software available from tensorflow.org.

[2] Anderson, C. W., and Sijerercic, Z. Classification of EEG Signals from Four Subjects During Five Mental Tasks. Tech. rep., Colorado State University, Fort Collins, 1996.

[3] Baumann, F., Li, F., Ehlers, A., and Rosenhahn, B. Thresholding a Random Forest Classifier. *Advances in Visual Computing 8888* (2014), 95–106.

[4] Bellman, C., AlOmari, R., Fung, A., Vargas Martin, M., and Liscano, R. *Challenges in the Effectiveness of Image Tagging Using Consumer-Grade Brain-Computer Interfaces.* Springer International Publishing, Cham, 2016, pp. 55–64.

[5] Bellman, C., Vargas Martin, M., Macdonald, S., Alomari, R., and Liscano, R. Excuse Me, Do I Know You From Somewhere? Unaware Facial Recognition Using Brain-Computer Interfaces. *Proceedings of the 50th Hawaii International Conference on System Sciences* (2017), 1217–1225.

[6] Bobes, M. a., Lopera, F., Comas, L. D., Galan, L., Carbonell, F., Bringas, M. L., and Valdes-Sosa, M. Brain potentials reflect residual face processing in a case of prosopagnosia. *Cognitive Neuropsychology 21*, 7 (2004), 691–718.

[7] Breiman, L. Bagging Predictors. *Machine Learning 24*, 421 (1996), 123–140.

[8] Brigham, J. C., and Ready, D. J. Own-race bias in lineup construction. *Law and Human Behavior 9*, 4 (1985), 415–424.

[9] Chang, C.-C., and Lin, C.-J. LIBSVM: A Library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology 2*, 3 (2011), 1–27.

[10] Chun, M. M., and Jiang, Y. Contextual Cueing: Implicit Learning and Memory of Visual Context Guides Spatial Attention. *Cognitive Psychology 36*, 1 (1998), 28–71.

[11] Delorme, A., and Makeig, S. EEGLAB : An Open Source Toolbox for Analysis of Single-Trial EEG Dynamics Including Independent Component Analysis. *Journal of Neuroscience Methods 134* (2004), 9–21.

[12] Duvinage, M., Castermans, T., Petieau, M., Hoellinger, T., Cheron, G., and Dutoit, T. Performance of the Emotiv Epoc Headset For P300-Based Applications. *Biomedical Engineering Online 12*, 1 (2013), 56.

[13] Ehrlichman, H., and Wiener, M. Ehrlichman Wiener - EEG Asymmetry During Covert Mental Activity.pdf. *Psychophysiology 17*, 3 (1980), 228–235.

[14] Eimer, M., Gosling, A., and Duchaine, B. Electrophysiological markers of covert face recognition in developmental prosopagnosia. *Brain 135*, 2 (2012), 542–554.

[15] Emotiv. Cortex arcade. `https://www.emotiv.com/product/cortex-arcade/`. (Accessed on 01/17/2017).

[16] Emotiv. Emotiv epoc - 14 channel wireless eeg headset. `https://www.emotiv.com/epoc/`. (Accessed on 01/17/2017).

[17] Emotiv. Emotiv epoc+ 14 channel mobile eeg. `https://www.emotiv.com/product/emotiv-epoc-14-channel-mobile-eeg/`. (Accessed on 02/11/2017).

[18] Emotiv. Emotiv sdk and apps for developers and proprietary research. `https://www.emotiv.com/developer/`. (Accessed on 01/17/2017).

[19] Emotiv. Minddrone. `https://www.emotiv.com/product/minddrone/`. (Accessed on 01/17/2017).

[20] Emotiv. What does emotiv pure eeg software include? `https://emotiv.zendesk.com/hc/en-us/articles/201226865-What-does-Testbench-software-include-`. (Accessed on 05/17/2017).

[21] Farquhar, J., and Hill, N. J. Interactions Between Pre-Processing and Classification Methods for Event-Related-Potential Classification Best-Practice Guidelines for Brain-Computer Interfacing. *Neuroinform 11*, 2 (2013), 175–192.

[22] Freund, Y., and Schapire, R. E. A Decision-Theoretic Generalization of On-Line Learning and an Application to Boosting. *Journal of Computer and System Sciences 55*, 1 (1997), 119–139.

[23] Fukunaga, K. Holdout, Leave-One-Out, and Resubstitution Methods. In *Statistical Pattern Recognition*, W. Rheinboldt, Ed., 2 ed. Academic Press, West Lafayette, 1990, ch. 5.3, p. 592.

[24] Goujon, A., Didierjean, A., and Poulet, S. The Emergence of Explicit Knowledge From Implicit Learning. *Memory & Cognition 42*, 2 (2014), 225–236.

[25] Hajcak, G., MacNamara, A., and Olvet, D. M. Event-Related Potentials, Emotion, and Emotion Regulation: An Integrative Review. *Developmental Neuropsychology 35*, 2 (2010), 129–155.

[26] Inese, P., Igor, T., and Arkadijs, B. Decision Tree Classifiers in Bioinformatics. *Rigas Tehniskas Universitates Zinatniskie Raksti 44* (2010), 118.

[27] Interaxon. Developer kit. `https://www.choosemuse.com/developer-kit/`. (Accessed on 01/17/2017).

[28] Interaxon. Muse ™ — meditation made easy. `https://www.choosemuse.com/`. (Accessed on 01/17/2017).

[29] Kaggle. Mean f score. `https://www.kaggle.com/wiki/MeanFScore`, November 2016. (Accessed on 02/09/2017).

[30] Keirn, Z. A., and Aunon, J. I. A New Mode of Communication Between Man and His Surroundings. *IEEE Transactions on Biomedical Engineering 37*, 12 (1990), 1209–1214.

[31] Lee, J. C., and Tan, D. S. Using a Low-Cost Electroencephalograph for Task Classification In HCI Research. In *Proceedings of the 19th ACM Symposium on User Interface Software and Technology* (Montreux, Switzerland, 2006), pp. 81–90.

[32] Luck, S. J. Ten Simple Rules for Designing ERP Experiments. In *Event-Related Potentials: A Methods Handbook*, T. C. Handy, Ed. MIT Press, Cambridge MA, 2005, ch. 2, p. 408.

[33] Malpass, R. S., and Devine, P. G. Eyewitness identification: Lineup instructions and the absence of the offender. *Journal of Applied Psychology 66*, 4 (1981), 482–489.

[34] Marcon, J. L., Meissner, C. A., Jackiw, L. B., Arbuthnott, K. D., and Pfeifer, J. E. Examining the cross-race effect in lineup identification using caucasian and first nations samples. *Canadian Journal of Behavioural Science 40*, 1 (2008), 52–57.

[35] Mountrakis, G., Im, J., and Ogole, C. Support vector machines in remote sensing: A review. *ISPRS Journal of Photogrammetry and Remote Sensing 66*, 3 (2011), 247–259.

[36] Németh, K., Zimmer, M., Schweinberger, S. R., Vakli, P., and Kovács, G. The background of reduced face specificity of N170 in congenital prosopagnosia. *PLoS ONE 9*, 7 (2014).

[37] Neuro, A. Waveguard original. `https://www.ant-neuro.com/products/waveguard_original`. (Accessed on 02/22/2017).

[38] NeuroSky. Mindwave. `http://store.neurosky.com/pages/mindwave`. (Accessed on 02/11/2017).

[39] Noble, W. S. What Is a Support Vector Machine? *Nature Biotechnology 24*, 12 (2006), 1565–1567.

[40] Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., and Duchesnay, E. Scikit-learn: Machine learning in python. *Journal of Machine Learning Research 12* (2012), 2825–2830.

[41] Phillips, P., Moon, H., Rizvi, S., and Rauss, P. The feret evaluation methodology for face recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence 22* (2000), 1090–1104.

[42] Phillips, P., Wechsler, H., Huang, J., and Rauss, P. The feret database and evaluation procedure for face recognition algorithms. *IEEE Trans. Pattern Analysis and Machine Intelligence 16* (1998), 295–306.

[43] Pu, X., Fan, K., Chen, X., Ji, L., and Zhou, Z. Facial expression recognition from image sequences using twofold random forest classifier. *Neurocomputing 168* (2015), 1173–1180.

[44] Rodriguez, A., Rey, B., and Alcaniz, M. Validation of a low-cost EEG device for mood induction studies. *Studies in Health Technology and Informatics 191*, June 2013 (2013), 43–47.

[45] Schacter, D. Implicit Memory : History and Current Status. *Journal of Experimental Psychol 13*, 3 (1987), 501–518.

[46] Schmidhuber, J. Deep Learning in neural networks: An overview. *Neural Networks 61* (2015), 85–117.

[47] Scikit-Learn. Ada boost classifier. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.AdaBoostClassifier.html. (Accessed on 03/09/2017).

[48] Scikit-Learn. Bagging classifier. http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.BaggingClassifier.html. (Accessed on 03/09/2017).

[49] Scikit-Learn. Ensemble methods. http://scikit-learn.org/stable/modules/ensemble.html#adaboost. (Accessed on 03/03/2017).

[50] Scikit-Learn. Naive bayes. http://scikit-learn.org/stable/modules/naive_bayes.html. (Accessed on 02/23/2017).

[51] Scikit-Learn. Nearest neighbors classification. http://scikit-learn.org/stable/modules/neighbors.html#classification. (Accessed on 02/27/2017).

[52] Scikit-Learn. sklearn.neighbors.kneighborsclassifier. http://scikit-learn.org/stable/modules/generated/sklearn.neighbors.KNeighborsClassifier.html#sklearn.neighbors.KNeighborsClassifier. (Accessed on 02/27/2017).

[53] Scikit-Learn. Support vector machines - mathematical formulation. http://scikit-learn.org/stable/modules/svm.html#svm-mathematical-formulation. (Accessed on 04/14/2017).

[54] Scikit-Learn. Svc. http://scikit-learn.org/stable/modules/generated/sklearn.svm.SVC.html#sklearn.svm.SVC. (Accessed on 03/06/2017).

[55] Shalgi, S., and Deouell, L. Y. Is Any Awareness Necessary for an Ne? *Frontiers in Human Neuroscience 6*, May (2012), 1–15.

[56] Shashibala, T., and Gawali, B. W. Brain Computer Interface applications and classification techniques. *International Journal of Engineering and Computer Science 5*, 7 (2016), 17260–17267.

[57] Subasi, A., and Erçelebi, E. Classification of EEG signals using neural network and logistic regression. *Computer Methods and Programs in Biomedicine 78*, 2 (2005), 87–99.

[58] Tang, Y. TF.Learn: TensorFlow's High-level Module for Distributed Machine Learning. 1–7.

[59] TFLearn. Activations - linear. `http://tflearn.org/activations/#linear`. (Accessed on 04/07/2017).

[60] TFLearn. Activations - tanh. `http://tflearn.org/activations/#tanh`. (Accessed on 04/07/2017).

[61] TFLearn. Core layers - fully-connected. `http://tflearn.org/layers/core/#fully-connected`. (Accessed on 04/07/2017).

[62] Tseng, Y.-C., and Li, C.-S. R. Oculomotor Correlates of Context-Guided Learning in Visual Search. *Perception & Psychophysics 66*, 8 (2004), 1363–1378.

[63] Vargas Martin, M., Cho, V., and Aversano, G. Detection of Subconscious Face Recognition Using Consumer-Grade Brain-Computer Interfaces. *ACM Transactions on Applied Perception 14*, 1 (2016), 1–20.

Appendix

1.5 Individual Sensor Results

Tables 1.1 and 1.2 contain the full best-/worst-sensor dataset for all participants. Due to certain participants having an entire class of data removed during pre-processing, certain participants had issues with their classification and were removed from the table. The results have been split into two tables due to their size.

1.6 Combined Sensor Dataset Results

1.6.1 CAD Dataset Participant Results

Table 1.3 contains individual participant results for all sensors used in the CAD dataset classifications. Due to classes missing in sensors from pre-processing, the CAD dataset participant count is slightly lower than the CD dataset (31 vs. 37).

1.6.2 CD Dataset Participant Results

Table 1.4 contains individual participant results for all sensors used in the CD dataset classifications.

Table 1.1: Each participants best and worst performing sensors for each classifier.

		Decision Tree	Sensor	Gaussian Naive Bayes	Sensor	K-Nearest Neighbours	Sensor	Random Forest	Sensor
P00	Best	0.769	AF3	0.615	AF3	0.615	AF3	0.714	AF3
	Worst	0.154	F4	0.143	FC5	0.125	FC5	0.133	FC5
P02	Best	0.632	AF4	0.769	F7	0.667	T8	0.667	T8
	Worst	0.143	F4	0.154	P7	0.364	AF3	0.286	P7
P03	Best	0.571	T7	0.500	F4	0.609	F4	0.700	O1
	Worst	0.267	AF4	0.154	FC5	0.286	T8	0.333	P7
P04	Best	0.625	FC5	0.615	AF4	0.632	P7	0.588	F4
	Worst	0.333	T7	0.333	F7	0.333	T7	0.308	O2
P07	Best	0.588	P8	0.909	O2	0.727	O1	0.750	F3
	Worst	0.267	F4	0.400	P7	0.400	O2	0.353	T7
P10	Best	0.857	F8	0.800	O2	0.667	FC6	0.727	F8
	Worst	0.333	P7	0.286	T7	0.364	O2	0.143	FC6
P12	Best	0.769	F7	0.667	F7	0.667	F7	0.706	F4
	Worst	0.182	F4	0.250	FC5	0.250	FC5	0.286	P7
P14	Best	0.769	P7	0.800	O2	0.769	FC6	0.706	F3
	Worst	0.364	AF4	0.375	F8	0.500	AF4	0.526	P8
P15	Best	0.706	P8	0.667	T8	0.706	F3	0.833	F7
	Worst	0.267	O1	0.200	FC6	0.286	AF4	0.267	T7
P17	Best	0.737	T7	0.750	T8	0.857	O1	0.778	P7
	Worst	0.375	AF4	0.444	O2	0.364	F7	0.333	FC6
P18	Best	0.667	FC6	0.588	FC6	0.800	O2	0.636	FC6
	Worst	0.143	F7	0.182	FC5	0.154	T7	0.267	AF3
P19	Best	0.625	F7	0.625	O1	0.800	P7	0.700	FC6
	Worst	0.286	AF3	0.421	F8	0.471	FC6	0.444	F3
P20	Best	0.857	F7	0.706	F7	0.778	F7	0.636	AF3
	Worst	0.286	F4	0.286	T7	0.235	AF4	0.286	T7
P25	Best	0.667	F8	0.824	T7	0.600	O2	0.750	T7
	Worst	0.154	T8	0.250	T8	0.267	T8	0.353	AF3
P26	Best	0.750	AF4	0.714	P7	0.667	P8	0.588	AF4
	Worst	0.286	O2	0.286	FC5	0.286	FC5	0.182	O2
P28	Best	0.750	O2	0.667	FC5	0.625	FC6	0.737	FC5
	Worst	0.267	F8	0.286	T8	0.267	T8	0.308	P8
P31	Best	0.667	T7	0.625	T8	0.706	T7	0.750	P7
	Worst	0.250	FC6	0.421	P8	0.400	FC6	0.429	F7
P32	Best	0.778	FC6	0.714	AF3	0.769	AF3	0.706	AF4
	Worst	0.154	F8	0.286	F8	0.400	F8	0.182	P7
P33	Best	0.714	FC6	0.824	O1	0.706	O1	0.714	P8
	Worst	0.375	AF4	0.154	F8	0.375	F8	0.364	T7
P36	Best	0.667	T8	0.706	T8	0.750	T8	0.632	P7
	Worst	0.250	P7	0.250	P8	0.429	FC6	0.353	O2

Table 1.2: A continuation of Table 1.1

		Decision Tree	Sensor	Gaussian Naive Bayes	Sensor	K-Nearest Neighbours	Sensor	Random Forest	Sensor
P41	Best	0.667	F3	0.667	T7	0.667	O2	0.700	P7
	Worst	0.286	P8	0.364	T8	0.154	T8	0.333	T8
P45	Best	0.632	O2	0.632	P7	0.706	O2	0.778	F4
	Worst	0.167	T8	0.133	T8	0.267	F7	0.375	T8
P46	Best	0.706	F3	0.875	F8	0.933	AF4	0.737	AF3
	Worst	0.333	AF3	0.500	FC6	0.286	FC6	0.400	FC6
P47	Best	0.800	F8	0.667	F8	0.667	AF3	0.778	O1
	Worst	0.222	P8	0.222	FC6	0.333	F7	0.400	FC6
P48	Best	0.706	F8	0.625	FC6	0.625	O2	0.750	F8
	Worst	0.167	F4	0.182	FC5	0.154	O1	0.375	F7
P49	Best	0.667	P7	0.533	P7	0.706	F8	0.750	F7
	Worst	0.182	FC6	0.143	FC6	0.154	O2	0.167	FC6
P51	Best	0.714	P7	0.700	O1	0.667	O2	0.667	F4
	Worst	0.182	AF3	0.182	FC5	0.267	FC6	0.267	AF3
P52	Best	0.571	P7	0.706	FC5	0.632	FC5	0.667	F8
	Worst	0.133	T8	0.167	T8	0.333	P8	0.143	T8
P53	Best	0.737	P8	0.625	P8	0.700	T7	0.778	F4
	Worst	0.167	F3	0.133	AF3	0.267	O1	0.250	F8
P54	Best	0.750	T8	0.769	AF3	0.706	O1	0.700	FC6
	Worst	0.182	T7	0.200	FC6	0.400	P8	0.353	AF4
P55	Best	0.714	P7	0.750	P7	0.667	P8	0.667	T8
	Worst	0.133	F3	0.286	O1	0.267	FC5	0.143	O1
P59	Best	0.615	T8	0.667	T8	0.714	T8	0.571	T8
	Worst	0.200	O2	0.143	F7	0.154	F7	0.250	F7
P71	Best	0.667	O2	0.800	FC6	0.750	FC6	0.667	F3
	Worst	0.333	F7	0.333	O2	0.308	FC5	0.400	F7
P72	Best	0.800	AF3	0.737	O1	0.706	T7	0.824	T7
	Worst	0.167	P7	0.471	P7	0.526	P7	0.375	F7
P73	Best	0.750	P7	0.667	F7	0.625	F7	0.824	AF3
	Worst	0.167	F7	0.308	O1	0.353	F8	0.154	P8
P77	Best	0.625	T8	0.875	O2	0.667	O1	0.875	O1
	Worst	0.167	FC5	0.200	T7	0.353	T7	0.235	P7
P80	Best	0.857	F7	0.857	F7	0.857	F7	0.857	F7
	Worst	0.286	F4	0.308	P7	0.556	P7	0.267	P8

Table 1.3: Individual participant results for the CAD dataset.

User	Random Forest	ROC AUC	Gaussian Naive Bayes	ROC AUC	Decision Tree	ROC AUC	SVC	ROC AUC	K-Nearest Neighbours	ROC AUC
P00	0.714	0.667	0.800	0.750	0.667	0.583	0.000	0.500	0.800	0.750
P02	0.500	0.500	0.571	0.500	0.545	0.583	0.625	0.500	0.500	0.667
P07	0.800	0.750	0.615	0.583	0.000	0.333	0.000	0.500	0.800	0.750
P10	0.571	0.500	0.545	0.583	0.769	0.750	0.462	0.417	0.625	0.500
P12	0.923	0.917	0.769	0.750	1.000	1.000	0.600	0.667	0.727	0.750
P15	0.714	0.667	0.615	0.583	0.769	0.750	0.545	0.583	0.714	0.667
P17	0.615	0.583	0.667	0.667	0.714	0.667	0.667	0.667	0.615	0.583
P18	0.714	0.667	0.625	0.500	0.615	0.583	0.500	0.500	0.625	0.500
P19	0.769	0.750	0.909	0.917	0.500	0.500	0.667	0.750	0.909	0.917
P20	0.667	0.667	0.000	0.250	0.667	0.667	0.714	0.667	0.667	0.667
P26	0.444	0.583	0.000	0.417	0.667	0.667	0.500	0.500	0.769	0.750
P28	0.769	0.750	0.923	0.917	0.769	0.750	0.857	0.833	1.000	1.000
P31	0.923	0.917	0.923	0.917	0.545	0.583	0.727	0.750	0.800	0.750
P32	0.857	0.833	0.800	0.750	0.857	0.833	0.545	0.583	0.857	0.833
P33	1.000	1.000	0.800	0.750	0.769	0.750	0.615	0.583	0.000	0.500
P36	0.800	0.750	0.400	0.500	0.800	0.750	0.429	0.333	0.667	0.500
P41	0.667	0.750	0.667	0.750	0.600	0.667	0.667	0.750	0.500	0.667
P45	0.727	0.750	0.400	0.500	0.600	0.667	0.625	0.500	0.500	0.667
P46	0.909	0.917	0.571	0.500	0.800	0.833	0.667	0.667	0.750	0.667
P47	0.909	0.917	0.833	0.833	0.667	0.750	0.500	0.667	0.714	0.667
P48	0.250	0.500	0.250	0.500	0.600	0.667	0.250	0.500	0.286	0.583
P49	0.667	0.667	0.545	0.583	0.500	0.500	0.286	0.583	0.923	0.917
P51	0.667	0.500	0.625	0.500	0.500	0.667	0.706	0.583	0.706	0.583
P52	0.800	0.833	0.429	0.333	0.923	0.917	0.667	0.750	0.800	0.750
P53	0.857	0.833	0.444	0.583	0.706	0.583	0.533	0.417	0.615	0.583
P54	0.706	0.583	0.571	0.500	0.667	0.667	0.500	0.500	0.750	0.667
P55	0.923	0.917	0.571	0.500	0.923	0.917	0.667	0.750	0.857	0.833
P59	0.857	0.833	0.714	0.667	0.909	0.917	0.600	0.667	1.000	1.000
P71	0.727	0.750	0.727	0.750	0.727	0.750	0.286	0.583	0.667	0.667
P73	0.667	0.750	0.500	0.500	0.400	0.500	0.500	0.667	0.500	0.500
P77	0.727	0.750	0.571	0.500	0.364	0.417	0.400	0.500	0.625	0.500

Table 1.4: Individual participant results for the CD dataset.

User	Random Forest	ROC AUC	Gaussian Naive Bayes	ROC AUC	Decision Tree	ROC AUC	SVC	ROC AUC	K-Nearest Neighbours	ROC AUC
P00	0.848	0.860	0.636	0.648	0.819	0.836	0.613	0.500	0.833	0.849
P02	0.657	0.675	0.492	0.523	0.607	0.650	0.000	0.500	0.713	0.741
P03	0.822	0.821	0.650	0.646	0.798	0.800	0.664	0.500	0.890	0.892
P04	0.862	0.868	0.554	0.603	0.753	0.765	0.000	0.500	0.853	0.866
P07	0.594	0.605	0.486	0.564	0.564	0.599	0.608	0.500	0.617	0.616
P10	0.696	0.648	0.573	0.548	0.540	0.557	0.000	0.500	0.658	0.633
P12	0.803	0.779	0.656	0.621	0.682	0.682	0.000	0.500	0.811	0.804
P14	0.733	0.704	0.598	0.560	0.687	0.675	0.000	0.500	0.709	0.714
P15	0.640	0.616	0.573	0.587	0.645	0.643	0.000	0.500	0.633	0.625
P17	0.721	0.690	0.647	0.632	0.626	0.614	0.000	0.500	0.667	0.682
P18	0.620	0.594	0.637	0.636	0.626	0.631	0.649	0.500	0.681	0.699
P19	0.711	0.692	0.565	0.567	0.628	0.656	0.667	0.500	0.634	0.629
P20	0.798	0.788	0.735	0.729	0.705	0.715	0.641	0.500	0.795	0.783
P25	0.543	0.562	0.526	0.591	0.457	0.517	0.561	0.500	0.494	0.565
P26	0.631	0.606	0.531	0.538	0.549	0.539	0.658	0.500	0.624	0.633
P28	0.881	0.872	0.715	0.715	0.867	0.864	0.000	0.500	0.936	0.931
P31	0.827	0.826	0.713	0.723	0.775	0.773	0.659	0.500	0.860	0.866
P32	0.711	0.710	0.687	0.679	0.603	0.590	0.652	0.569	0.701	0.714
P33	0.745	0.725	0.597	0.615	0.644	0.664	0.000	0.500	0.725	0.690
P36	0.775	0.740	0.621	0.602	0.711	0.673	0.605	0.580	0.834	0.816
P41	0.755	0.725	0.622	0.591	0.717	0.699	0.000	0.500	0.850	0.843
P45	0.703	0.683	0.524	0.504	0.655	0.656	0.536	0.504	0.691	0.665
P46	0.837	0.813	0.726	0.709	0.717	0.695	0.000	0.500	0.735	0.734
P47	0.701	0.675	0.446	0.449	0.611	0.605	0.627	0.500	0.725	0.733
P48	0.706	0.664	0.595	0.580	0.613	0.593	0.000	0.500	0.691	0.664
P49	0.834	0.826	0.690	0.679	0.727	0.732	0.663	0.500	0.870	0.866
P51	0.703	0.709	0.589	0.612	0.634	0.636	0.000	0.500	0.752	0.741
P52	0.763	0.758	0.658	0.664	0.714	0.727	0.000	0.500	0.854	0.862
P53	0.740	0.728	0.495	0.535	0.689	0.670	0.663	0.500	0.732	0.719
P54	0.755	0.738	0.640	0.639	0.578	0.602	0.659	0.500	0.707	0.718
P55	0.833	0.821	0.710	0.718	0.783	0.782	0.662	0.500	0.832	0.839
P59	0.850	0.855	0.566	0.597	0.776	0.776	0.000	0.500	0.872	0.870
P71	0.788	0.769	0.625	0.556	0.670	0.653	0.000	0.500	0.735	0.727
P72	0.796	0.799	0.562	0.579	0.591	0.611	0.617	0.500	0.713	0.704
P73	0.765	0.749	0.627	0.634	0.710	0.720	0.000	0.500	0.769	0.763
P77	0.686	0.667	0.604	0.624	0.594	0.603	0.642	0.500	0.715	0.720
P80	0.832	0.828	0.721	0.717	0.815	0.823	0.663	0.486	0.910	0.901

1.7 Neural Network

Table 1.5 contains all participants' classification accuracies using the neural network as described in Section 4.9.

Table 1.5: Full individual classification F-scores using the neural net as described in Section 4.9

Participant	F-Score
P00	0.750000
P02	0.516556
P03	0.792899
P04	0.645161
P07	0.491228
P10	0.559140
P12	0.747573
P14	0.617801
P15	0.619883
P17	0.649746
P18	0.626506
P19	0.550265
P20	0.726257
P25	0.517647
P26	0.544379
P28	0.893617
P31	0.820809
P32	0.627907
P33	0.579545
P36	0.767677
P41	0.662857
P45	0.620000
P46	0.751323
P47	0.596273
P48	0.610837
P49	0.729282
P51	0.523256
P52	0.839378
P53	0.500000
P54	0.684783
P55	0.836957
P59	0.836957
P71	0.618421
P72	0.623377
P73	0.747475
P77	0.601093
P80	0.879310