A Study of Password Recall, Perceived Memorability, and Strength Using BCIs

by

Ruba Alomari

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

Doctor of Philosophy

in

Computer Science

University of Ontario Institute of Technology

Supervisor: Dr. Miguel Vargas Martin

April 2018

THESIS EXAMINATION INFORMATION

Submitted by: Ruba Al Omari

Doctor of Philosophy in **Computer Science**

Thesis title: Gibberish doesn't help with remembering: A Study of Password Recall, Perceived Memorability, and Strength Using BCIs.

An oral defense of this thesis took place on April 3, 2018 in front of the following examining committee:

Examining Committee:

Chair of Examining Committee	Dr. Khalil El-Khatib
Research Supervisor	Dr. Miguel Vargas Martin
Examining Committee Member	Dr. Julie Thorpe
Examining Committee Member	Dr. Ramiro Liscano
External Examiner	Dr. Robert Biddle
University Examiner	Dr. Faisal Qureshi

The above committee determined that the thesis is acceptable in form and content and that a satisfactory knowledge of the field covered by the thesis was demonstrated by the candidate during an oral examination. A signed copy of the Certificate of Approval is available from the School of Graduate and Postdoctoral Studies.

Acknowledgments

I am immensely grateful for the support I have received from my supervisor Dr. Miguel Vargas Martin. He was generous in his time, funding, and most importantly mentorship. I learned much from Dr. Martin and will be forever indebted to him.

I would like to thank the amazing research team in ERC2100B and SIRC4120, Shane MacDonald, Christopher Bellman, Amit Maraj, Spencer Lamash, and Dr. Ramiro Liscano, for their time, insightful discussions, and feedback.

I would also like to thank my thesis defense Examining Committee. It was a great honour to have Dr. Robert Biddle, Dr. Faisal Qureshi, Dr. Julie Thorpe, and Dr. Ramiro Liscano on the committee.

Furthermore, I would like to thank the Natural Sciences and Engineering Research Council of Canada (NSERC) for its generous financial support during my studies.

Finally, a huge thank-you to my family for their continued and unconditional love and support. I would not have completed this work without your encouragement.

Ruba Alomari

Abstract

Passwords are considered the most common method of authentication and studies are frequently conducted to understand users' password habits. In this thesis, we run two empirical studies that provide information to further our understanding of the trade-off between security and usability in passwords, using off-the-shelf brain-computer interfaces (BCIs). Initially, we conducted an experiment with 19 participants, where password recall was studied. We followed this with a second experiment with 77 participants, where perceived password memorability and recall were studied. In both experiments, the effect of password strength on user's behaviour was investigated.

Password memorability and strength were studied by collecting electroencephalogram (EEG) potentials upon presentation of different passwords to participants. After the presentation of passwords, participants were asked to perform either password recall or password memorability ranking based on the experiment. Features from the EEG signals were extracted in three domains: power spectrum from the frequency domain, statistics from the time domain, and wavelet coefficients from the time-frequency domain. Feature selection methods were used, and the selected parameters and feature subsets were submitted for classification based on the different tasks performed by participants.

Password recall, being the most established metric of password memorability, was investigated thoroughly in both experiments. An average accuracy of 85% was obtained when predicting password recall from short-term memory. Prediction of password recall from long-term memory was performed over 8-10 days period. On the first day, an accuracy of 81% was achieved, whereas a near-to-random guess results were found on the second and eighth days. Prediction of users' judgment of password memorability was performed with an 82% accuracy.

Password strength effect on password recall and perceived memorability was investigated, and a strong influence was found with an effect size of 6.8 on password recall from short-term memory,

and 3.8 on memorability perception. The results present empirical data that may help explain the common practice of users selecting weak and memorable passwords, also suggesting users are able to sense password strength and make usability decisions based on that.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Summary

Passwords have been criticized since they were first used, and research for replacements has been extensively done in the hope of ending the password era, yet it still dominates authentication systems. Different methods of authentication have been developed, since computers were invented, to replace passwords. Evaluation of these methods is mostly based on two main areas, security and usability [18, 82, 90, 15]

Bonneau et al. [18] evaluated two decades of proposals to replace text passwords. They surveyed a wide range of proposed systems including biometrics, tokens, graphical passwords, federated login protocols, cognitive schemes, and one-time passwords. They proposed a framework for comparison based on 25 security, usability, and deployability benefits. Bonneau et al. found that none of the proposed methods come near to the benefits provided by text passwords, and suggested that "we are likely to live considerably longer before seeing the funeral procession for passwords arrive at the cemetery."

Wiedenbeck et al. [122] describe the "password problem" arising from expecting users to comply with two conflicting requirements. First, passwords should be easy-to-remember. Second, passwords should be secure, i.e., they should look random, and be hard-to-guess; they should be frequently changed, and be different for different accounts of the same user; they should not be written down or stored in plain text. These conflicting requirements make it almost impossible for humans to meet, resulting in users compensating by creating weak passwords and handling them insecurely.

In this thesis, we capitalize on recent advances in brain-computer interfaces (BCIs), which have enabled them as affordable consumer-grade devices for non-medical purposes such as academic research, marketing, and entertainment. Using BCIs, we determine additional factors that contribute to understanding users' behaviour. We study password features by recording electroencephalogram (EEG) potentials upon presentation of different types of passwords to the participants while asking them to perform different tasks related to password usability and security. The EEG data is then used to classify and consequentially predict password usability. Empirical data was collected to investigate the following questions:

(1) Is there a correlation between EEG signals elicited upon presentation of a password and its:

   (a) perceived memorability

   (b) strength

   (c) recall

   If there is a correlation, can the EEG signals help predict these metrics?

(2) Is there a correlation between password strength and its perceived memorability or recall? In other words, does the user recognize password strength and perceive stronger passwords as less memorable, hence less usable?

Our methodology found that EEG signals are different when users are presented with passwords that are perceived as the most memorable, compared to those perceived as the least memorable. A difference is also detected in EEG signals elicited when users are presented with passwords that were successfully recalled ten seconds later, compared to those that were not. Password strength is found to have a major influence on recall and user judgment of password memorability.

The results suggest the possibility of predicting users' perspective on password usability -such as memorability perception- using EEG signals collected with off-the-shelf BCIs. The results also advert to the users' ability to sense password strength, and to make decisions on memorability based on the strength of passwords.

Additionally, different feature extraction and feature selection methods are used, and the results are compared in terms of classifier performance. Results show wavelet coefficients extracted in the time-frequency domain have the highest performance.

1.2 Thesis Statement and Scope of Research

Thesis Statement: Password memorability and strength attributes are studied using consumer-grade BCIs and machine learning algorithms. EEG data was collected in two experiments, in the first one, password recall and strength were investigated. In the second experiment, perceived password memorability, recall, and strength were investigated. Different feature extraction and selection methods were used, and machine learning was utilized to analyze the data. The findings suggest the possibility of predicting perceived password memorability based solely on EEG signals. Furthermore, it was found that password strength heavily influenced its recall from short-term memory and how users perceived its memorability. The results present empirical data that may help explain the common practice of users selecting weak and memorable passwords.

This research has been primarily concerned with studying participant's brain waves upon presentation of passwords and how they can guide us in understanding users' behaviour in terms of password memorability. Despite other aspects that can have deciding factors on memorability, such as age, it is important to mention that these aspects are beyond the scope of this dissertation.

Terms and Acronyms

Brain-Computer Interfaces (BCIs): "A brain–computer interface is a communication system that does not depend on the brain's normal output pathways of peripheral nerves and muscles. ... Current BCI's record electrophysiological signals using noninvasive or invasive methods [123]."

Electroencephalogram and *Electroencephalography:* Although both terms are referred to as EEG, there is a difference in the meaning. An Electroencephalography is a method of recording the electrical potentials that are generated from the brain cells, using flat, small metal discs called electrodes attached to the scalp. Whereas, an electroencephalogram refers to the recorded electrical potential or brain signal, which was first discovered by Hans Berger in 1929 [14]. In this work, a noninvasive BCI was used to record brain signals through specific tasks; however, when EEG signals or EEG data are mentioned, they refer to the brain recorded signals, rather than the method of recording.

Guessability: How many guesses a particular cracking algorithm with particular training data would take to guess a password. It has become a common metric of password security, and aims to model real-world attackers and to provide per-password strength estimates [114]. In this thesis,

password guessability is used when referring to specific passwords guessability that are estimated and reported. Otherwise, password strength is used when referring to passwords in general forms, such as weak passwords or strong passwords as defined within the context.

Support-Vector Machine (SVM): is a powerful machine learning algorithm [28], and is commonly used in classifying EEG signals [99, 111, 73]. It gained traction for its accuracy in many fields, especially in the biomedical domain [22, 45].

1.3 Motivation

This work is motivated by the desire to understand users' password behaviours and habits, because of the widespread of passwords as a primary method of authentication, especially in online systems. While difficulties of passwords are well-known, their continuous use is accepted and justified for a variety of practical reasons [18, 52, 103]. A central problem is that passwords memorability aligns with their guessability; passwords that users prefer tend to be guessable by attackers and hacking tools. As a result, a better understanding of this phenomena is of great potential to help in increasing the security of authentication, and therefore protecting privacy and access to resources.

It is often reported in the literature that users choose poor passwords, and that they trade-off password security for memorability [103, 17, 126, 116, 33, 125]. These reports are based either on analyzing password leaks, or empirical user studies analyzing users' behaviour. To the best of my knowledge, no empirical work has been done studying the effect of factors such as password strength on password memorability and how users perceive passwords of different strength. For example, Taneski et al. [103] recently surveyed research done on password use and security between 1979 and 2014. They addressed the question concerning what are the major problems with creating and managing textual passwords, and report "higher entropy makes a password more difficult for a user to memorize" as one of the answers. Taneski et al. based this conclusion on Cisar and Cisar's [25] work. Cisar and Cisar included the "Entropy and memorability" section in their research from a paper published by the Gartner Group [5]. The paper published by the Gartner Group does not include any data or references to support this claim. In fact, in the same section, when discussing the percentage of users who write their passwords down as an indicator of password entropy, the

Gartner paper mentions that "This is speculative, is based on anecdotal evidence and assumes a normal (Gaussian) distribution."

The lack of empirical data studying password recall and perceived memorability in correlation to strength, is one of the motives to conduct this research. It is not possible to explore that correlation through leaked password datasets, because recall information, such as authentication at first login, number of attempts, as well as if users write their passwords down, for the leaked passwords are not available. We decided to run experiments to explore these correlations.

Another motive is the ability to utilize EEG techniques to study the way the users sense password usability and how they make decisions based on that. The results may help the security community understand more of the human factors when dealing with passwords.

1.4 Contribution

The novel contribution of this dissertation is understanding users' password behaviour in terms of memorability and strength. The findings indicate that the human brain may have the ability to sense password memorability and strength beyond the obvious. Prediction of perceived password memorability was achieved with acceptable accuracy based on collected EEG signals. This work also provides insight on the effect password strength has on password recall, and how users perceive password memorability. A secondary contribution is the comparison of the performance of different feature extraction and feature selection methods.

1.5 Organization of Thesis

The remaining parts of this thesis are structured as follows. In Chapter 2, the concepts utilized in the design of the user studies are introduced. First, we discuss the memory model used in designing the recall testing. Next, password strength and the methods used to estimate password guessability are described. Then, the algorithm used in generating a password based on a user-provided seed is introduced. Last, a brief overview of brain-computer interfaces technology and EEG research on the relationship between cognitive tasks and EEG activity is presented.

Related work is discussed in Chapter 3. The methodology and user studies design are described in Chapter 4. The results and analysis are presented in Chapter 5, and then discussed in Chapter 6. Finally, conclusions and future work are discussed in Chapter 7.

Chapter 2

Background

2.1 Introduction

This chapter introduces the background material required to better understand this dissertation, starting with an explanation of the memory model that is used in designing the two user studies in Section 2.2, followed by a brief history of password entropy in Section 2.3, along with an explanation of the password guessability estimation tool that is used in the studies design. The variants used in the Persuasive Text Passwords (PTP) system are introduced in Section 2.4, which we rely on to design the algorithm used in generating study passwords. We end with a review of brain-computer interfaces technology in Section 2.5.

2.2 Memory Model

Text passwords are purely memory-based. However, human memory has limited temporal capacity. In 1956, Miller [77] published his research "The Magical Number Seven, Plus or Minus Two" which became to be known as Miller's Law. Miller argued that the human brain has a short-term memory span of seven –plus or minus two– items. Memory span is defined as the longest list of items that a person can repeat back in correct order immediately after presentation on 50% of all trials. Later research challenged the magical number seven, and claimed that memory span is not constant, other factors were found that impact the memory span, such as the category and the characteristics of the items. For example, Baddeley et al. [11] found memory span to be inversely related to word length.

Moreover, Gregg [48] found words with higher frequency have a higher memory span than those of lower frequency; which helps explain why users choose to create shorter passwords with familiar words.

Atkinson and Shiffrin proposed a general theoretical framework to view human memory in 1968, which became to be known as the multi-store memory model [9]. Their framework differentiates between the structural memory and the control processes. In this model, memory is divided into three structural components: the sensory register, the short-term store, and the long-term store. Structure of the memory as proposed by Atkinson and Shiffrin is shown in Figure 2.1 and is described in more details below.

Sensory Register is composed of multiple registers, one register for each sense (sight, hearing, taste, smell, and touch). Once information is presented to the subject through a sensory register, it is kept there for a very brief period of time. Atkinson and Shiffrin used visual information as a prime example of this register. When attention is given to the sensory registers, the information is copied from them to the short-term store. Otherwise, when there is no attention, the information decays and is forgotten. The period of time sensory registers can keep information depends on which sense received the information. Most research is done in the auditory sense and visual sense fields. It is believed that the visual sensor register keeps the information for up to 1 second [95], and the auditory sensor register keeps the information for 2 seconds, which can reach 5 seconds depending on context [31] [110].

Short-Term Store, also referred to as short-term memory, is where information is transferred when an individual pays attention to the information received through sensory registers. The information decays and is forgotten in a similar way to what happens to the information in the sensory register. The short-term memory store has a longer retention period for information than that of the sensory register. Atkinson and Shiffrin suggested that information is lost within a period of about 30 seconds. Short-term memory retention period is also based on the type of information and can be slightly longer depending on the items being processed. For example, Peterson and Peterson [86] found that short-term memory retains verbal information for a period of about 20 seconds. Atkinson and Shiffrin designate short-term store as a rehearsal buffer, where a person can keep a limited amount of information as long as they rehearse it. The information is then moved to the long-term store.

Long-Term Store, also referred to as long-term memory, is believed to be a relatively permanent repository. The amount of information transferred from the short-term store to the long-term store

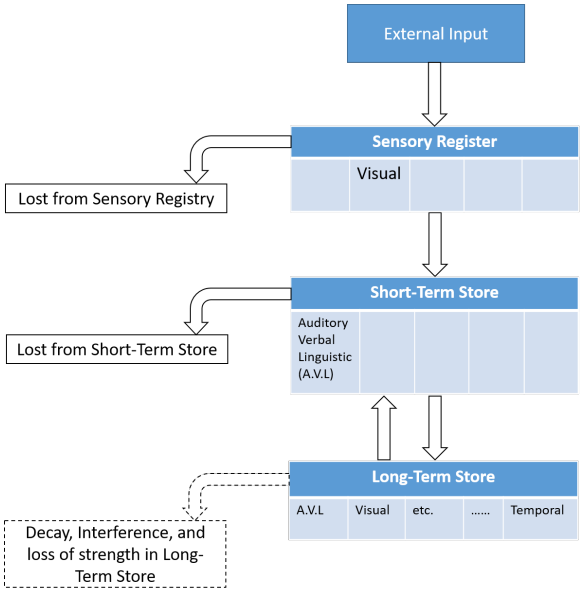


Figure 2.1: Structure of the Memory as proposed by Atkinson and Shiffrin.[1]

is a function of control processes. Atkinson and Shiffrin note that the transfer of information from short-term store to long-term store does not mean the removal of the information from the short-term store. Instead, the information is copied rather than moved. The more attention paid to the item in short-term memory, the stronger it exists in long-term memory. Long-term memory is assumed to be a fairly permanent repository for information in this model. However, information decay in this store due to reasons such as decay and fading (when an item was not encoded specifically enough), interference (when an item in memory interferes with another item), and retrieval failure. Items are retrieved from memory by a process called recall, which requires the person to reproduce the item from memory. Retrieval is facilitated by means of established connections with previous structures in long-term memory. The more connections that are established, the easier retrieval will be later [90]. An uncued recall is the most fallible type of retrieval mechanism, and becomes more difficult as people age [84]. Both types of capacity and duration limits in long-term memory are controversial [29].

The multi-store memory model proposed by Atkinson-Shiffrin is criticized for a number of reasons [88], and later memory models (e.g., [10, 89, 55]) expanded on the multi-store model; however they all agree that short-term memory (also called working memory) has limited capacity, and

[1]Reprinted from Psychology of learning and motivation, Vol 2, Atkinson, R. C. and Shiffrin, R. M, "Human memory: A proposed system and its control processes.", p 89-195, Copyright (1968), with permission from Elsevier.

that rehearsing items while in the short-term memory strengthen their association in the long-term memory.

In the design of the user studies in this dissertation, we keep Atkinson and Shiffrin memory model in mind when testing password recall from short-term and long-term memory, also when providing a rehearsal tool to participants to practice typing their password.

2.3 Password Strength

Claude Shannon [93] defined the term *entropy* in information theory as "a statistical parameter which measures in a certain sense, how much information is produced on the average for each letter of text in the language. If the language is translated into binary digits (0 or 1) most efficiently, the entropy H is the average number of binary digits required per letter of the original language." While Shannon was alluding to English text strings, the term entropy became widely used in cryptography as a measure of the difficulty in guessing or determining a password or a key [23].

Estimating the entropy of a password depends on the number of options available for each character. If a password was binary and composed only of zeros and ones chosen randomly, there will be 2^n possible values of that password, where n is the number of bits in the password. The password, in this case, has n bits of entropy. As a general rule, the entropy of a randomly chosen password is calculated as n^l where n is the number of available characters, and l is the length of the password. For example, if a password was based on the standard keyboard, where there are 95 printable characters, the space available for each password character is 95. The entropy of 8 characters long password that is randomly chosen based on a standard keyboard will be $95^8 \approx 6.6 \times 10^{15}$, which is almost equivalent to 2^{52}. In this case, the password is said to have 52 bits of entropy. Entropy H is given by:

$$H = log_2(n^l) \tag{2.1}$$

where n is the number of possible characters space in the password, and l is the length of the password.

Unfortunately, users tend to choose easier-to-remember passwords that include names, short words, dates, and patterns resulting in easier to guess passwords [103, 17, 126, 116, 33, 125]. This

makes estimating entropy for users' passwords a difficult task because they are not random and do not follow a specific frequency distribution.

Other approaches based on Shannon's entropy have been developed to estimate password strength, such as *guessing entropy* and *minimum entropy*. Guessing entropy is claimed to be the most important measure of password strength because it evaluates the password resistance to guessing attacks. Guessing entropy is an estimate of the average amount of work required to guess the password of a selected user [23], and is based on the location of each character in that password. For example, a guessing entropy based on Shannon's method will give the first letter in a password an entropy value of 4 bits, whereas the entropy of the next character will be 2 bits. Different characters' combinations create different values of the password guessing entropy. An 8-character password from a standard 95-character English keyboard has a 52-bits of entropy (when randomly chosen) compared to 18 bits of entropy if created by a user with no password rules implemented.

In recent years, using guessing entropy derived from Shannon's theory to determine a password strength has become increasingly impractical. Bonneau and Shutova [19] cautioned against optimistic security estimates using Shannon's estimates of entropy. Other algorithms have been developed based essentially on password length. A modern-day password cracker was proposed by Wheeler [120], who developed an algorithm to estimate password strength called *zxcvbn*. The algorithm consists of three phases: match, estimate, and search. The matching phase finds the following patterns: token, reserved, sequence, repeat, keyboard, date, and brute force. The principle behind zxcvbn is to catch common patterns, and not penalize sufficiently complex passphrases. Using different password databases, Wheeler carried out four attacks: PCFG, Markov, HashCat, and John The Ripper, to test the password strength estimator accuracy for zxcvbn, NIST entropy, and KeePass Password Safe [66] (an open source password manager). The results showed that both NIST and KeePass substantially overestimate password strength.

In this thesis, we use the password strength estimator designed by Wheeler to estimate the strength of passwords presented to participants; as it provides accurate results at a low cost, more details in Chapter 4. Note that the strength estimation method used in this dissertation is based on heuristics, and that password strength estimation is not a fixed value across time as cracking tools and techniques are constantly improved and updated.

2.4 Preload, Replace, and Insert

Forget et al. [43] proposed a lightweight system they called Persuasive Text Passwords (PTP), to influence users in creating more secure passwords, without reducing usability. The study was conducted on 83 participants. The proposed system is based on a password created by the user; then the PTP system will guide the user in generating a stronger version of their chosen password. This is done by placing randomly-selected characters at randomly-determined positions in the users' initial password. They used three variants in the PTP system to generate the stronger version of the password: Preload, Replace, and Insert, as described in [43]:

- "Preload. Users are given the system-assigned characters before creating their password. The characters are positioned randomly within the first eight character slots. Users create their password around the system-placed characters.
- Replace. Users first choose an initial password as they would for a typical password system. The system replaces characters in the users' passwords at random positions with randomly-chosen characters.
- Insert. After users select an initial password, as usual, the system inserts randomly-selected characters at random positions between user-chosen password characters, lengthening the password."

Note that the Replace variant they used in the study was limited to replacing two characters only. The Insert condition had three variants of Insert-2, Insert-3, and Insert-4. The results of the study showed that under the three variants, PTP improved the security of users' passwords. However, the system had limitations, as to how much a password can be improved security-wise before users start employing coping mechanisms to deal with memorizing the stronger passwords. They found that inserting three random characters in the password, is the most users can remember, without posing a load on the user that results in the need for longer periods of time to remember the new password.

In this thesis, we created an algorithm based on the PTP system variants with some modification to generate passwords from seed words provided by the participants (more on this in Section 4.3.4).

2.5 Brain-Computer Interfaces and EEG

Advances in cognitive neuroscience and brain imaging technologies have provided us with the ability to interface directly with the human brain; through the use of sensors that can monitor some of the physical processes that occur within the brain which correspond with certain forms of thought [102]. These technologies were used to build brain-computer interfaces, which are communication systems that do not depend on the brain's normal output pathways of peripheral nerves and muscles. Research into BCI systems has mainly involved the recording of electroencephalographic (EEG) signals using surface electrodes [30].

Consumer-grade BCIs have come a long way in the past decade. They became more available to the regular consumer for entertainment purposes at reasonable prices, which also facilitates research interested in investigating and utilizing brain signals in non-medical applications. BCIs are now considered mature enough that Human-Computer Interaction (HCI) researchers must add them to their tool belt when designing novel input techniques [102]. BCIs had a variety of applications in the past decades, such as:

- BCIs for Assistive Technology: Applications aimed at providing assistive technologies for people with motor disabilities, e.g., communication (Yes/No communications and spellers), environmental control (thermostat and television), mobility (wheelchair control and robotics).

- BCIs for Recreation: BCI control for mainstream applications such as games, virtual reality, and creative expression (music and visual arts). A number of simple mainstream game controllers primarily based on BCIs have become available on the market.

- BCIs for Cognitive Diagnostics and Augmented Cognition: BCIs have been developed to aid in diagnosing, influencing, and augmenting cognitive function; e.g., comma detection meditation training, visual image classification, and attention monitoring [102].

EEG research on the relationship between cognitive tasks and EEG activity has produced extensive literature. There is a number of reasons for questioning the reliability of the correlations between individual cognitive processes and accompanying changes in EEG signals [30]. One of these reasons is the difficulty in finding which cognitive processes are taking place during the performance of a particular cognitive task and where in the brain the activity is located; another reason is whether factors outside the immediate cognitive processes may be causing the changes in signals [30]. For

example, factors such as emotional state [12], gender [118], and the difficulty of the task [47]. Being aware of such additional factors and how they affect the signals does not need to be a disadvantage; instead, it could provide further ways of training subjects to control parts of their EEG activity [30].

In this dissertation, the location in the brain where the activity causing the change in EEG signals is, lays outside the scope of our work, and is likely of more interest to the cognitive scientists. We are also not researching the effect additional factors, such as emotions and age, may have on the change in EEG signals. Instead, we try to minimize these effects by analyzing the data within subjects for some of the tasks.

Chapter 3

Related Work

3.1 Introduction

In this chapter, a summary of password benefits and why passwords persist to exist is presented in Section 3.2. The current literature around users password habits and behaviour is discussed in Section 3.3. Work done in studying human memory recall and recognition of different types of words is looked at in Section 3.4. Finally, we survey the use of EEG in studying words in Section 3.5.

3.2 Password Benefits

Passwords offer a large number of benefits, Bonneau et al. [18] summarized those benefits and compared them to other authentication methods. Figure 3.1 shows the results of this comparison. It is clear that no scheme examined is perfect; the researchers note that no scheme achieves all usability benefits. Not a single scheme is dominant over passwords, i.e., does better on one or more benefits and does at least as well on all others. Almost all schemes do better than passwords in some criteria, but all are worse in others. Most proposed methods performed better in terms of security than passwords, which suggests that developers of these methods are coming from a security background. All proposed methods performed worse in terms of deployability. Despite many of the proposed methods being memory-wise effortless, addressing a major issue with the current use of passwords, yet these methods failed to compare to other important benefits provided by text passwords.

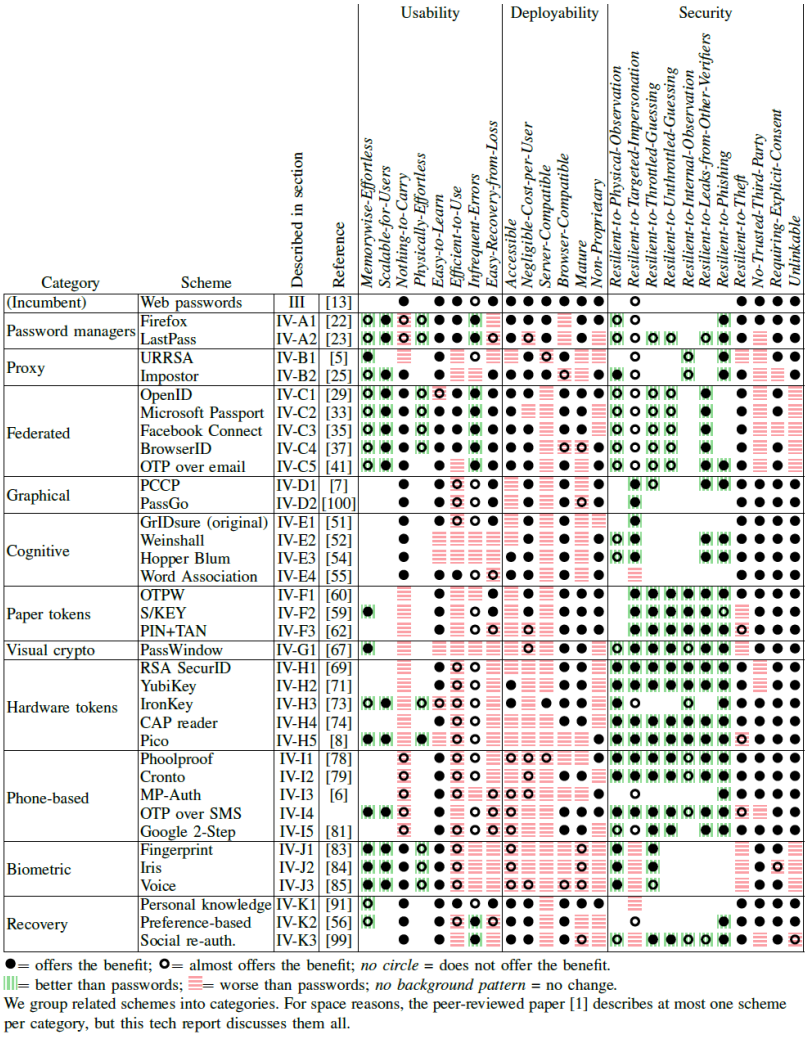


Figure 3.1: Comparative evaluation of the various schemes examined by Bonneau et al. [18].[1]

Herley and van Oorschot [52] discuss the reasons why password replacement attempts have failed every time, and why no progress has been done in the past two decades. They state that usability has degraded significantly, while security has not improved. In their research agenda, they discuss the resilience of passwords, the confusion as to the properties needed, costs, and benefits of replacing passwords; and how seeking a best-fit solution is better than seeking a silver bullet solution. The researchers conclude with two major findings, first that passwords are actually the best fit for many of the scenarios they are used in nowadays, in which case research should be done on how to

[1] ©[2012] IEEE. Reprinted, with permission, from [Bonneau et al., The Quest to Replace Passwords, IEEE Symposium on Security and Privacy, 2012]

support the use of passwords in these scenarios rather than replacing them. Second, in the cases where passwords are not the best fit, there needs to be work to clearly identify the requirements expected from the replacement, as well as an evaluation method to score and compare the proposed alternatives against these requirements.

3.3 Users' Password Habits

Studies are frequently conducted to understand users' password habits, some are done by the industry through large-scale studies, and others are done by academia through laboratory and other methods to observe users' password behaviour, such as password reuse habits and password complexity.

Password Memorability

Password memorability is often reported on when discussing the cognitive load password requirements put on users. Florêncio et al. [41] reported that Internet users maintain, on average, 25 accounts that require passwords, whereas, Adams and Sasse [3] reported that –if multiple passwords cannot be avoided– four or five is the maximum number of regularly used passwords that users can be expected to cope with. The number is lower if passwords are infrequently used. Due to this cognitive load, users tend to choose easier-to-remember passwords that include names, short words, dates, and patterns resulting in easier to guess passwords [103, 17, 126, 116, 33, 125]. Brown et al. [21] found that two-thirds of passwords are created around the user's personal characteristics, and that names and birthdays are the primary information used in creating passwords. In addition, users have a large number of passwords which are re-used and easily forgotten [41, 32, 51].

Stanton and Greene [97] studied password recall using complex password-like stimuli; they investigated how memorable complex character strings of different lengths –that might be used as higher-entropy passwords– are. Participants were given one password at a time from Table 3.1, and there was a set of three task phases, practice, verification, and entry. During the entry phase, participants had to type the memorized password ten times, as fast and as accurately as they could.

Their results indicated that the longer a character string is, the longer it takes a person to memorize, recall, and enter it. Longer strings also increase the probability of errors.

Hun et al. [57] conducted a similar experiment studying the effect of the length of system-generated PINs on memorability. They used a 48-hour interval to study long-term memorability

Table 3.1: Character string order, string, and length used in Stanton and Greene [97] experiment.

Order	String	Length
1	5c2'Qe	6
2	m#o)fpˆ2aRf207	14
3	m3)61fHw	8
4	d51)u4;X3wrf	12
5	p4d46*3TxY	10
6	q80¡U/C2mv	10
7	6n04%Ei'Hm3V	12
8	4i_55fQ$2Mnh30	14
9	3.bH1o	6
10	ua7t?C2#	8

and found that system-generated 4-digit PINs outperform 6-, 7-, and 8-digit PINs in long-term memorability. However, they noticed that there was no statistically significant difference in memorability between 6-, 7-, and 8-digit PINs.

Quite often, user text password behaviour is studied in comparison to other types of proposed passwords. Although these user studies objective is to explore the proposed password systems, it allows us at the same time to understand more of the users' text password habits through these studies.

Yan et al. [126] studied passwords based on mnemonic phrases and compared them to conventional user-generated passwords, and to random user-generated passwords. 288 participants were divided into three groups. The first group was the control group; participants were asked to create a conventional password of seven characters, one of which is non-letter. The second group was the random password group, participants were given a sheet of paper with the letters A–Z, and the numbers 1–9 printed on it; and were told to select a password by closing their eyes and randomly picking eight characters. This group (the random password group) were advised to keep a written record of the password with them until they had memorized it. The third group was the pass phrase group, participants were asked to choose a password based on a mnemonic phrase. The results showed that passwords based on mnemonic phrases were as easy to remember as conventional user-created passwords, and are harder to guess than conventional user-created passwords.

Many password studies recommend different methods to help improve the security of passwords, for example proposing methods to influence users to create more secure passwords [43], user security training [3], and implementing password policies that prevent users from using common words and

repeated patterns [16]. Some of the approaches burden the user who resorts to different coping techniques.

Password Policies

Password policies are used to encourage users to create strong passwords and to ensure users are following the minimum requirements for creating passwords. However, users do not always follow these policies. Adams and Sasse [3] studied 139 users' security behaviour and perception toward passwords in 1999. The initial data was collected through a web questionnaire. Out of the 139 participants, 30 participants were also invited for in-depth interviews. The results showed low-security motivation among users can be caused by security mechanisms and policies that don't take into account the users' work practices and usability. They recommend designers of security procedures, must realize that users are the key to successful security system. The study warned that mechanisms that look secure on paper will fail in practice, as long as security departments do not understand how the mechanisms they design are used in practice.

Eleven years after Adams and Sasse's study, Inglesant and Sasse [60] raised the questions: has anything changed? Are recommendations made to improve password policies being implemented? The researchers weigh on the cost of unusable password policies. They studied a password diary of one week, for 32 staff members, from two organizations. The diaries were to capture "what happens during the day," and were followed by interviews focused on the context around the passwords identified in the diary. Their results showed that a number of factors in password policies lead to frustration among users, who are not able to comply with the policy. An example of an excessively strict password policy which caused frustration to users, was a policy that combined the requirements for the password to be excessively strong, and for it to be frequently changed, and for the new password to differ significantly from previous passwords. This caused users' reduction of productivity. It also caused them to adopt coping techniques, such as writing the passwords down and recycling them. This is counterproductive to the password policies objectives, yet is caused by the password policies themselves. The study argues against the popular view that "if only [users] understood the dangers, they would behave differently," and instead claims that "if only security managers understood the true costs for users and the organization, they would set policies differently."

User Persuasion

A small number of studies try to use a persuading approach. They investigated the use of persuasion to convince users into creating strong passwords. Weirich and Sasse [119] published one of the early studies addressing this method. The researchers discuss the use of fear (security policy threatening punishment when users do not comply), and why it works for some users and not for others. They conducted interviews with 17 participants. Results showed that many users do not expect to suffer personal consequences from using weak passwords. The authors discuss steps to persuade users such as: presenting danger as a threat to the organization reputation, punishing non-compliant actions if they were a result of carelessness rather than lack of knowledge, and making the punishments be known by other users. They conclude that although users cannot be forced to behave in the desired secured fashion, there should be an effort to persuade them.

Horcher and Tejay argued that users' selection of stronger passwords start in convincing the users of the importance of their selection [54]. They discussed how implementing mandatory controls will provide a simple level of information security, because there is a difference between knowing the correct security behaviour and doing it. They claim that users make choices based on their assessment of the importance of rules, not simply because the rule exists. And that when there is a conflict between convenience and functionality, the user must have a motivation to follow the rule, in addition to knowledge of the rule. User willingness to follow the rules determines how they will act, so information security programs must be designed to persuade users of the importance of secure behavior, not just inform them. The effect of user security training and education on generating a strong password has been heavily researched. However, this is outside the scope of this literature review.

Thorpe et al. [105] discuss a different kind of user persuasion. The researchers showed how change to a graphical password interface could modify the distribution of user-chosen passwords, and thus possibly the security it provides. They conducted a user study with 34 participants divided into two groups. The participants were presented with a background image for the users to create a graphical password on. The difference between the two groups was the way the background image was presented. In one group the images were presented as a curtain pulled from left to right, and in the second group, the same background image was pulled from right to left. The user study results showed that image presentations significantly modified the distribution of user's first click-points.

3.4 Studying Words' Memorability

What is the difference between a password and a word? Can the results obtained from studies that investigate the human memory of words be generalized to passwords, and what difference does this thesis offer? These are some of the questions we asked ourselves over the course of four years while preparing this dissertation.

Passwords can be thought of as a special type of words. Earlier, when authentication by passwords was invented, passwords could be any word. For example, a six-characters dictionary word like "hello" qualified as a password. As technology advanced, it allowed for a password that consists of a simple word to be guessed within fractions of a second using readily available dictionary attack tools. This made password requirements change over time and become more complex, and now a single dictionary word does not qualify as a password.

Passwords differ in two main aspects from dictionary words. First, passwords are more complicated in composition; users are required to have symbols and numbers in addition to characters in their passwords. So although a password can have a dictionary word, it is expected to have other strings with it, and in some cases as in the passphrases, it is expected to be more than one word. Second, users are expected to remember passwords for extended periods of time, having said that, research done on memorability of words, can be used as a guide to performing research on memorability of passwords.

Research studies investigated memorability of common or high-frequency (HF) words versus uncommon or low-frequency (LF) words. DeLlosh and McDaniel [35] suggested that there is a better recall and serial order memory for HF words than LF words in pure lists. Hulme et al. [58] found that memory span was lower for non-words versus words the participants listened to. Worthen and Roark [124] showed better recall for bizarre words versus common words depending on the context and representation type. However, the subject of common HF words versus LF words memorability is complex and depends on many factors such as recall versus recognition, mixed lists and pure lists, familiarity effect on recognition, task nature, and subsequent memory [24, 127, 107, 108, 76].

Studies also looked into users judgment of word memorability of HF versus LF words. Guttentag et al. [49] found evidence that participants are not aware of the higher recognition memorability of LF words, and that their judgment changes based on the time the judgment occurs in the study. Benjamin Aaron [13] replicated Guttentag et al.'s experiment and found supporting evidence to Gut-

tentag findings, in addition to finding out that users rely on different cues when making judgments during the study period than they do when making judgments during the recognition test. Based on Guttentag et al. finding that users are not aware of the effect word frequency has on them when making memorability judgment; we investigate the same idea to find out if users are not aware of factors such as password strength when making a judgment on password memorability.

3.5 Use of EEG to Study Word Recognition

The use of EEG to predict recognition of studied words and pictures is investigated in the literature. A number of studies showed a difference in elicited event-related potentials (ERP) for objects that are recognized versus objects that are not recognized.

Sanquist et al. [92] used EEG to study ERPs in brainwaves for two tasks: a judgment task and a recognition task. In the judgment task, participants were presented with a pair of words and were asked to judge if the paired words were the same or different based on criteria provided to them. In the recognition task, participants were presented with a number of words, a part of which they have already studied in the judgment task. They were asked to report if they recognize the words presented. All words studied were high-frequency words. Their findings suggest that the ERP components have specific characteristics based on the stimuli's subsequent recognizability. Items that were recognized showed more positive slow waves and larger late positive ERP components than the items that were not recognized.

Friedman and Trott [44] used ERPs to study memory encoding in young and old adults. They asked subjects to study two lists of sentences. Each sentence contained two nouns and was preceded by the phrase List 1 or List 2. After the study of the two lists, pairs of nouns were presented, and subjects were asked to make judgments regarding their old/new status; whether they "know" or "remember" the pair of words; and if the item was from List 1 or List 2. They observed larger positive ERPs for the subsequently remembered items than those for both subsequently known or subsequently missed items.

Paller et al. [83] recorded EEG for 10 subjects while they were studying 10-word lists. Subjects were asked to judge words as interesting/uninteresting or edible/inedible (e.g., apple is edible) based on the task assigned to them. Next, subjects were assigned a distraction task, after which, they were asked to perform a recall of the studied words. After a second distraction task, subjects were

asked to perform a recognition test of the studied words. The study found ERPs elicited by words recalled during the recall test had higher voltages than ERPs elicited by words not recalled.

Nie et al. [80] used ERPs to compare memory recognition of faces versus words and found that words are similar to faces, and that any detected difference is affected by whether there is a previous representation in long-term memory and not whether the stimulus involves letters or faces.

Other studies found evidence that brain activity preceding the presentation of a stimulus can contribute to subsequent memory encoding. Noh et al. [81] studied EEG recordings collected from subjects studying the pictures of cars and birds. In the recognition task, participants were asked to judge if they studied the presented picture or not. The researchers observed that the per-stimulus and the during-stimulus information distinguished between recollection and familiarity. The during-stimulus distinguishing information occurred in the alpha band (8–12 Hz).

Alomari et al. [6] compared EEG data collected from 19 participants while studying two lists of words, one list consisted of high-frequency words and the other list of randomly generated characters and symbols (non-words). There was an observable difference in brain waves associated with the two different lists. Next, participants were asked to recall the studied words from both lists. Words that had higher voltages EEG signals had a higher chance of being recalled. Some studies measured Event-Related Potentials (ERPs) elicited during the presentation of a stimulus using functional Magnetic Resonance Imaging (fMRI), to investigate the effect of word frequency [127, 24]. Bridger et al. [20] used fMRI to analyze ERPs to explore recognition processes based on word frequency mirror effect. Zubicaray et al. [34] used fMRI to investigate word frequency and strength effects; their findings supported models that interpret higher recognition (not recall) rate for LF words compared to HF words.

Chapter 4

Methodology and User Study Design

4.1 Introduction

In this chapter, we discuss the design of the two user studies carried out in this thesis, as well as data collection, data pre-processing, and feature extraction. Both user studies were conducted over three sessions and spanned a period of 8-10 days.

The first study was conducted with 19 participants and aimed at studying password recall from short-term memory and long-term memory, as well as the effect of password strength on recall; it will be referred to as *Password Recall* user study. In this study, we collected EEG data using an Emotiv Epoc headset, which is a low-cost alternative to medical-grade EEG recording devices for researchers who want to consider out-of-the-lab applications [36]. Duvinage et al. [36] compared a medical-grade system, the ANT device, and the Emotiv Epoc headset by determining their respective performances in a P300 BCI using the same electrodes; they also reviewed previous Emotiv studies. Their findings showed, that in terms of performance, contrary to some BCI leader criticisms but coherently with previous Emotiv studies, the Emotiv Epoc headset performance is above random and not due to muscular or ocular artifacts [36]. This was supported by far above chance classification rates. However, they emphasize that when comparing the Emtoiv Epoc headset to the medical-grade headset, a large under performance of the Emotiv device was noticed, and the headset is

recommended to be only chosen for non critical applications such as games and communication systems.

The second user study was conducted with 75 participants, with the objective of investigating users' judgment of password memorability, password recall from long-term memory, and the effect of password strength on users' judgment; this study will be referred to as the *Password Perceived Memorability* study. EEG data in this study was collected using a Muse headband, which is a wearable wireless headband; it is relatively new in the market, and its primary use is marketed as a meditation training device. Its commercial availability, moderate cost, portability, and dry sensors, make it increasingly popular. A number of studies were conducted to evaluate the performance of the Muse headband. Krigolson et al. [69] demonstrated the possibility of conducting ERP research without being reliant on event markers, using a portable Muse headband and a single computer. Abujelala et al. [2] used the Muse headband to evaluate user engaged enjoyment in two tasks (playing two different video games on a tablet), their findings supported previous literature on enjoyment; which they were able to measure reliably using the headband. Wiechert et al. [121] used the Muse to analyze data collected during five activities that represent day-to-day functions as a proof of concept. The activities were reading, doing nothing, watching a video, playing a game, and listening to music. Using a number of classification techniques, they showed that it was possible to identify both the persons and the activities with a reasonable degree of precision. Surangsrirat and Intarapanich [100] analyzed EEG data collected by the Muse during meditation and other activities. Upon analyzing the recordings, they demonstrated that the collected data has enough features to enable the headband to be used as a tool for research as well.

As discussed in Section 2.3, zxcvbn is used to estimate the strength of passwords presented to participants, it reports the number of guesses needed crack a password, and the password's strength on a scale of 0-4 as shown in Table 4.1. All the password lists used in both experiments in this dissertation were run through zxcvbn to maintain consistency in the results and during analysis.

The *Password Recall* user study is explained in Section 4.2, and the *Password Perceived Memorability* user study in Section 4.3. Feature extraction and selection methods are discussed in Section 4.4.

Table 4.1: Description of password strength, the number of guesses needed to crack the password, and the score.

Description	# of Guesses	Score
Too guessable, risky password	$< 10^3$	0
Very guessable, protection from throttled online attacks	$< 10^6$	1
Somewhat guessable, protection from unthrottled online attacks	$< 10^8$	2
Safely unguessable, moderate protection from offline slow-hash scenario	$< 10^{10}$	3
Very unguessable, strong protection from offline slow-hash scenario	$\geq 10^{10}$	4

4.2 Password Recall User Study

In this user study, we run an experiment to investigate password recall from short-term memory and long-term memory using EEG data collected upon presentation of a number of passwords. To study recall from short-term memory, participants were presented with a number of passwords, after 10 seconds of presentation, they were asked to perform a recall of the passwords they saw. This task was performed for 68 passwords over a period of two days.

The long-term memory recall is studied by asking the participants to choose one of two passwords generated based on a seed word provided by them. They were given a rehearsal tool to help memorize the password, and asked to return to the lab on the second and eighth days for a recall of the password. This task was performed for one password over a period of 8-10 days.

Details of the experiment hypotheses, design, and data collection are described below.

4.2.1 Hypotheses

The research questions under investigation are:

(1) Is it possible to use machine learning and EEG data collected upon presentation of passwords to predict short-term memorability of passwords?

Hypothesis 1 (H1) *It is possible to predict passwords' recall from* ***short-term memory*** *based on EEG data collected from participants upon presentation of the passwords.*

(2) Is it possible to use machine learning and EEG data collected upon presentation of passwords to predict long-term memorability for passwords?

Hypothesis 2 (H2) *It is possible to predict passwords' recall from* ***long-term memory*** *based on EEG data collected from participants upon presentation of the passwords.*

(3) Are stronger passwords always harder to recall?

Hypothesis 3 (H3) *There is a correlation between password strength and its recall.*

4.2.2 Study Design

We designed an experiment to measure and collect data to test the three hypotheses listed above. The memory model proposed by Atkinson and Shiffrin and discussed in Section 2.2 is used [9] in the study design. The study was designed to test hypotheses H1, H2, and H3 in the same experiment, so it will be clarified where each hypothesis is tested during the experiment explanation. The study design was as follows:

EEG recordings were collected from participants while they were presented with two categories of passwords, common and random (see Section 4.2.3). Participants were asked to perform a number of tasks, and classification was then conducted based on recall data collected from these tasks (H1 and H2), or based on entropy data calculated by the password strength estimator(H3).

Password recall was tested from both short-term and long-term memory. First, to test recall from short-term memorability (H1), participants were presented with passwords from the two categories (mentioned above) alternatively and were asked to try to memorize them. They were then asked to recall the passwords they saw by typing them in a text box on the next screen. Each password was presented for a period of 10 seconds to make sure the item is moved from the sensory registry to short-term memory, and that the recall is done within the time span the item stays in the short-term memory (which is up to 30 seconds).

Second, to test recall from long-term memory (H2), participants were asked to memorize one password on Day 1, and were asked to come back to the lab for a recall of that password on the second and eighth day of the experiment. In summary, H1 and H2 were tested as follows:

Test recall from short-term memory: Passwords were presented to participants for a period of 10 seconds, then participants were asked to perform an immediate recall of the password after presentation.

Test long-term memorability: Participants were asked to memorize one password over a period of 8-10 days, they were asked to come back to the lab to do a password recall on the second day of the experiment, and again on the eighth day. A window of 24–48 hours was given for participants to accommodate their schedule. This period is common and has been used by password user studies in the field [57, 39, 113].

To explore the effect password strength has on recall (H3), no EEG data was used. Recall data collected from the first task, where short-term memory was tested, was combined with password guessability scores and explored through different statistical tests.

4.2.3 Password Lists

Two password lists were generated. The first list consisted of the most commonly used passwords published by SplashData [96], an organization specializing in password management applications. The published annual list "Worst Passwords of (year)" consists of the most common 25 passwords compiled from millions of passwords that were compromised and leaked during that year, and is referred to as the worst 25 passwords of that year. We used the lists published during the period 2011–2015 to compose an initial list of 125 passwords.

The list each year shows a large number of common passwords that are reused from the year before. We removed the recurring passwords in the initial list and ended up with 48 unique passwords out of the 125 passwords. The list of those 48 passwords will be referred to as the *common* list. These passwords are considered weak because when a password database is leaked, the passwords are used to train password guessing tools, which makes these passwords easy and fast to crack, thus the name "Worst passwords." Zxcvbn's password strength estimator was used to test how guessable each of the 48 passwords in the *common* list is.

Table 4.2 shows the *common* passwords list that was consolidated, along with its strength score and the number of guesses it took to compromise. Note that the majority of the listed passwords score as too guessable (Score = 0), and only five passwords are very guessable (Score = 1). These scores are expected, as explained above, these passwords are considered very weak.

The second list was created using the "Secure Password Generator" online tool [1]. To match the same number of passwords in the *common* list, 48 random passwords were generated; each was eight characters long and included lowercase characters, uppercase characters, symbols, and numbers. These rules were chosen to create the random passwords as they are common and widely used. Characters and digits that are easily mistaken for each other were excluded, such as '1' (the digit one) and 'l' (lowercase L). Since participants will use the computer in the lab and not their own computers, characters that may have different locations on a keyboard based on the keyboard layout were excluded. This is based on a pilot study we did, where participants found it difficult to locate some of these characters (e.g., { } [ ] ( ) /\' " ‘ ~, ; : . <>). This password list will be

Table 4.2: The *common* passwords list.

Password	Score	Guesses log_{10}	Password	Score	Guesses log_{10}
123456	0	0.30	password	0	0.48
12345678	0	0.60	qwerty	0	0.70
12345	0	0.85	123456789	0	0.78
football	0	1.18	1234	0	0.90
1234567	0	1.00	baseball	0	1.11
welcome	0	2.05	1234567890	0	1.40
abc123	0	1.15	111111	0	0.95
1qaz2wsx	0	1.46	dragon	0	1.04
master	0	1.28	monkey	0	1.20
letmein	0	1.23	login	1	4.27
princess	0	1.78	qwertyuiop	1	5.93
solo	0	3.02	passw0rd	0	0.70
starwars	0	1.72	mustang	0	1.32
access	0	1.90	shadow	0	1.25
michael	0	0.70	superman	0	1.43
696969	0	1.30	123123	0	1.98
batman	0	1.63	trustno1	0	1.57
iloveyou	0	1.68	adobe123	1	5.81
admin	0	2.85	photoshop	1	4.25
sunshine	0	1.67	password1	0	2.28
azerty	0	2.46	000000	0	1.54
ashley	0	1.81	jesus	0	2.10
ninja	1	3.12	bailey	0	1.77
654321	0	0.48	qazwsx	0	1.50

Table 4.3: The *random* passwords list.

Password	Score	Guesses log_{10}	Password	Score	Guesses log_{10}
dz9?W8jp	2	8	Ss9%y-S!	2	8
K8V#eMDE	2	8	23sC#hqz	2	8
_t3_KjbH	2	8	eDJ%2rkN	2	8
WVzG&h7e	2	8	d9Himi^F	2	8
IQp%794+	2	8	v8J!7#R%	2	8
2%bFCg6a	2	8	*&@2V_v7	2	8
TqAb2!TU	2	8	8n+hY9kc	2	8
XQ5qta-_	2	8	Bf@=J6z+	2	8
36rmp&N7	2	8	VPu8#f-7	2	8
z?Hm_^9X	2	8	d2x$S*Z5	2	8
95–DrhM!	2	8	9Hwdj9*K	2	8
3#!dhXxP	2	8	2!Hp2pQj	2	8
C3&b^EPp	2	8	N+ay9#cv	2	8
Mp8n%S5x	2	8	INHm^4it	2	8
6x7wTb-z	2	8	aYQ_9k7M	2	8
^hF36efG	2	8	Y2sVNa*=	2	8
36I!f$rD	2	8	Vu2fp=84	2	8
stVU&4sT	2	8	UG5f&2i=	2	8
6Wy9fX+R	2	8	fY7-d@pT	2	8
M^5He%e$	2	8	Cd2_KV6r	2	8
d#E9S5Lf	2	8	&8nzyFRQ	2	8
bP+2ay^B	2	8	sBZ-2=QU	2	8
D36EZ-qc	2	8	4xE@JwA9	2	8
56D**MjD	2	8	Jf-2Z*9I	2	8

referred to as the *random* list, and the passwords are shown in Table 4.3. Note that all of the listed passwords score as somewhat guessable (Score = 2).

These two opposing lists were chosen in order to simulate very easy-to-remember passwords (common English words) and very difficult-to-remember passwords (random characters and symbols). We anticipated that the difference in recall of the two lists would be noticeable, so we wanted to study the brain signals associated with these two categories as a validation of the hypotheses being investigated as well as a baseline. The two lists also simulate different password strengths, too guessable and somewhat guessable passwords.

4.2.4 Participants

We recruited 19 volunteers to participate through a university-wide email advertisement: 8 females and 11 males. Volunteers were first asked to fill an eligibility questionnaire (questionnaire

Table 4.4: Distribution of programs among participants.

Program	Count
Engineering and Applied Science	10
Health Sciences	3
Sciences	2
Social Science And Humanities	2
Business and Information Technology	1
Commerce	1
Total	19

in Appendix B), and if deemed eligible were contacted to schedule the experiment's first session. Upon arrival for the experiment, participants signed a consent form (Consent form replicated in Appendix C.1) in accordance with procedures approved by the University's Research Ethics Board, and were asked to fill a short pre-study questionnaire to collect demographical information. Volunteers were undergraduate students and ranged in age from 18 to 30 years, with an average age of 21.5 and standard deviation of 2.0. All participants had a normal or corrected-to-normal vision and declared not having a history of neurological or mental disorders. When asked on a scale of 1 (novice) to 5 (expert), "how would you rate yourself with respect to your computer skills?" 5 participants rated their computer skills as 5, whereas 10 answered 4, and 4 rated their computer skills as 3. Participants came from diverse fields of study shown in Table 4.4. Participants were compensated with $10 for their participation, distributed over the three sessions.

4.2.5 Experiment Procedure

The experiment was held in a lab environment, and the experiment protocol was explained to participants in detail. The BCI headset was then fitted to the participants, and they were asked to look at a 17-inch monitor about 50 centimeters in front of them. The experiment consisted of three sessions distributed over three days.

Day 1: On the first day (Session 1) participants were asked to provide a word that is a minimum of 4 characters and a maximum of 8 characters, which will be used as a seed word to create a password for the study (H2). This seed word is referred to as *PWDSEED*.

Next, participants were asked to study and recall a number of passwords that appeared on the screen as per the instructions in Figure 4.1. A total of 50 passwords appeared on the screen, 48 passwords of which were from the two password lists *common* and *random* (H1), whereas two

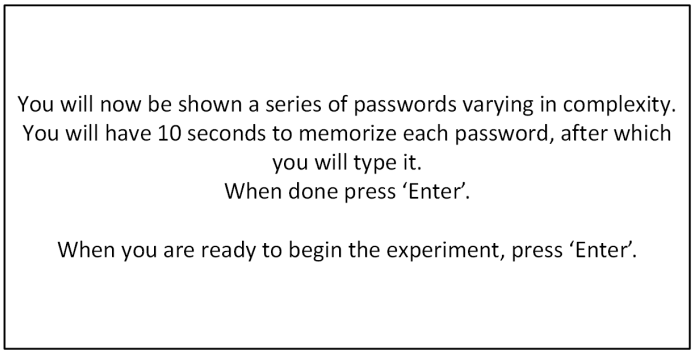


Figure 4.1: Instructions to study and recall passwords appearing on the screen.

passwords were based on the seed word provided earlier by the user (H2). The 48 passwords presented to participants were divided equally and alternatively between the *common* list and the *random* list, with each password appearing on the screen for a period of 10 seconds. All passwords appeared in black font on a white background and were centered in the middle of the screen. The password would then disappear and a screen with an empty text box would be presented where participants were given unlimited time to recall and type the password they just saw. Once the recall was done, they were asked to hit the Enter key, and a password from the *random* list would appear following the same procedure. EEG data was recorded during the whole experiment.

During the presentation of the passwords to the participant, two passwords were generated based on the seed word provided by the participant. These two password candidates are referred to as Password Candidate 1 (*PWC1*), and Password Candidate 2 (*PWC2*). The rules used to derive these two passwords resemble the method created by Forget et al. [43] and discussed in Section 2.4. Our derivation rules are not identical to the rules used in PTP, but rely on the same principle. The derivation rules used were as follows:

Capitalize: Every letter in the seed word has a one of three chance of being capitalized. This rule is applied to all seed words.

Replace: If the seed word has any of the following characters {a, e, s, t, c, h, o, i} then each of these characters has a 30% chance of being substituted with one of the following characters {@, 3, $, +, (, #, 0, 1} respectively. These replacements were chosen as a type of leetspeak, where symbols that look like Latin letters are used in their place.

Insert: If the seed word length is between 4 and 6 characters, then insert 2 more random numbers to bring the password length to 6-8 characters. These 2 random numbers can be inserted both at

Table 4.5: Examples of *PWC1* and *PWC2* along with their guessability scores.

PWDSEED	**Guesses** (log_{10})	***PWC1***	**Guesses** (log_{10})	***PWC2***	**Guesses** (log_{10})
test	1.9	teS+72	5.1	t3$T78	4.9
march	2.1	m@r(H53	5.3	1M@rcH6	7.0
spider	2.1	37$PIdEr	5.8	sP1D3r20	6.0

the beginning or at the end of the seed word. They can also be added one number at the beginning and one number at the end of the seed word. If the seed word length is 7 characters, insert 1 more random number following the same pattern. All possible insertions have equal chances and are performed randomly.

Table 4.5 shows examples of *PWC1* and *PWC2* created from given seed words, along with their estimated strength. Note that the seed words shown in Table 4.5 are created by the researcher for illustration purposes only, since most of the seed words provided by participants contained identifiable information.

PWC1 and *PWC2* were inserted back into the password lists. *PWC1* was inserted in the 8^{th} position, and *PWC2* was inserted in the 18^{th} position. The 8^{th} location was chosen to introduce the first password candidate, to ensure the participant got comfortable with the task at hand. The second password candidate was introduced in the 18^{th} position to avoid presenting it toward the end of the list, as the participant may start experiencing fatigue going through the 50 passwords.

EEG signals were recorded for *PWC1* and *PWC2* the same way they were recorded for the other 48 passwords. By the end of this session, participants were shown the two passwords (*PWC1* and *PWC2*) and were asked to choose a password that will be used during the study as shown in Figure 4.2. The password chosen by the participant is referred to as *PWDCH*. Once the participant chose a password, they were redirected to a screen where they were given the option to rehearse the password until they felt they memorized it. Participants had the option to hide the password from the screen and practice entering the password. The rehearsal tool gave feedback as *Correct* or *Incorrect* when the password was entered as shown in Figure 4.3. The experiment protocol on Day 1 is summarized in Figure 4.4.

Day 2: On the second day (Session 2), participants were asked to return to the lab. At the beginning of the session, they were prompted to recall the password they chose on Day 1. If they failed to recall the password three times, they were given the option to see the password and practice

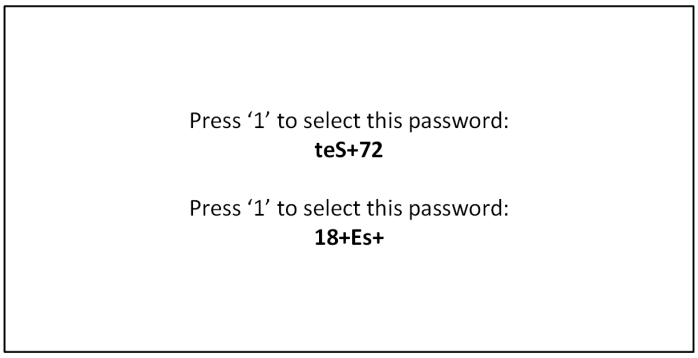


Figure 4.2: Choice is given to participants to choose one of the two passwords. The seed word in the above example is *test*.

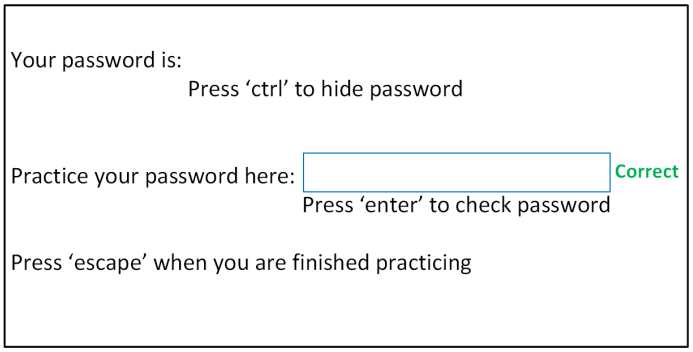


Figure 4.3: Rehearsal tool interface to practice typing the password and testing it until the participant feels comfortable remembering the password.

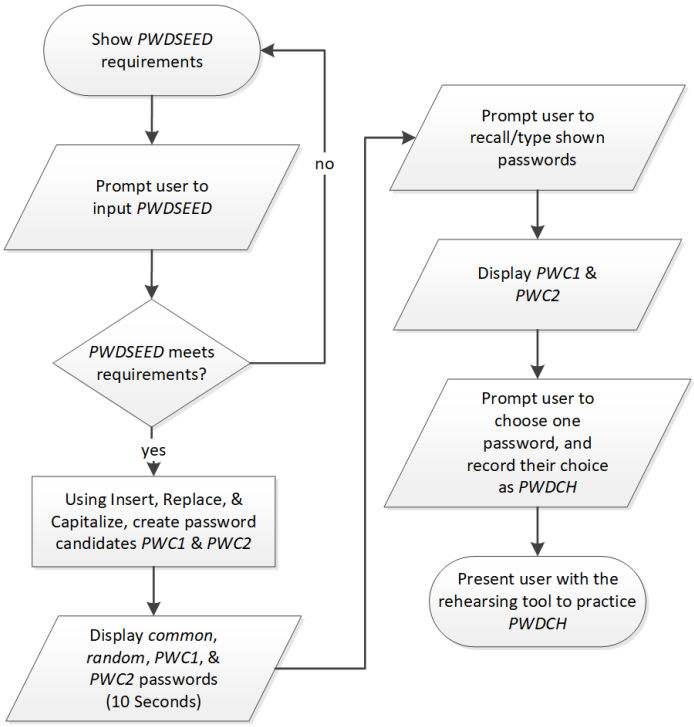


Figure 4.4: Summary of password recall user study procedure on Day 1.

typing it. Next, participants were shown another 20 passwords from the *common* and *random* lists for 10 seconds each and following the same protocol on Day 1.

Day 8: On the eighth day (Session 3), participants returned to the lab one week after the second session, and were asked to recall the password they chose on Day 1. Next, they filled a post-study questionnaire, where participants were explicitly asked if they wrote the password during the study, and encouraged to answer truthfully: "Did you, at any time during the study, write down or record your password in any way? Please be honest in your answer - it's OK if you did." The answers were used to exclude participants who answered "Yes" to this questions from recall data analysis.

A summary of the passwords shown on each day is depicted in Figure 4.5.

4.2.6 Data Acquisition - Emotiv Epoc

The Emotiv Epoc headset, shown in Figure 4.6 [38], was used for data capture, which records data at 128 Hz. The Emotiv Epoc headset is a consumer-grade BCI and has 14 electrodes (AF3, F7, F3, FC5, T7, P7, O1, O2, P8, T8, FC6, F4, F8, and AF4). Electrodes are located according to the Modified Combinatorial Nomenclature (MCN) naming system, which is based on the international

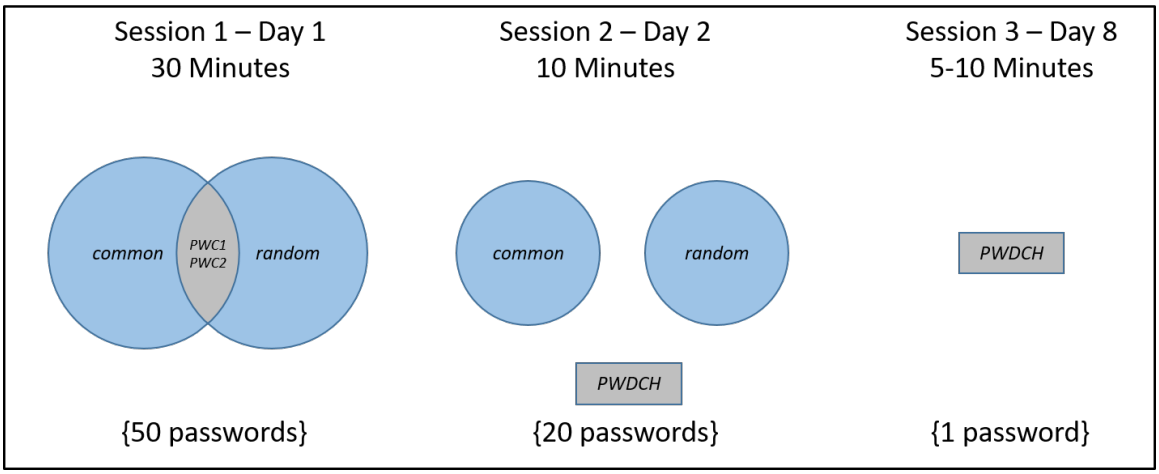


Figure 4.5: Summary of the three sessions in the *Password Recall* user study. Passwords with the blue background are used to test recall from short-term memory (H1) and password strength (H3), whereas passwords with the gray background are used to test recall from long-term memory (H2).

Figure 4.6: Emotiv Epoc headset used to collect EEG data in the *Password Recall* study [38].

10-20 system for EEG electrode positioning [62], but fills in intermediate sites halfway between those of the existing 10-20 system. Electrode locations are displayed in Figure 4.7.

A custom application was written using C# and Visual Studio to accommodate the needs in this experiment. Markers were sent from the custom application to the headset via an emulated serial port to ensure clear beginning and ending of EEG data for each password that was presented to the participants. The markers are used for segmentation during the analysis.

4.2.7 Data Pre-Processing

EEG data pre-processing and analysis was performed using EEGLAB, a MATLAB toolbox, because of its rich libraries and a wide range of functions. The raw data consist of voltage signals collected from participants using the Emotiv Epoc headset. A marker was placed in the data at the exact time the stimulus (the password) was presented to the participant, using the custom application. The markers have unique values to differentiate between *common* passwords and *random* passwords.

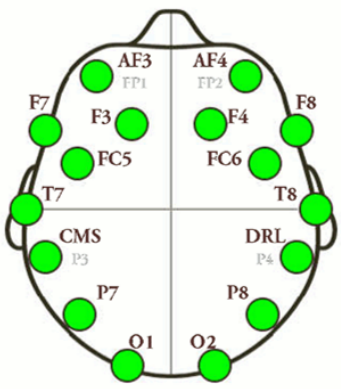


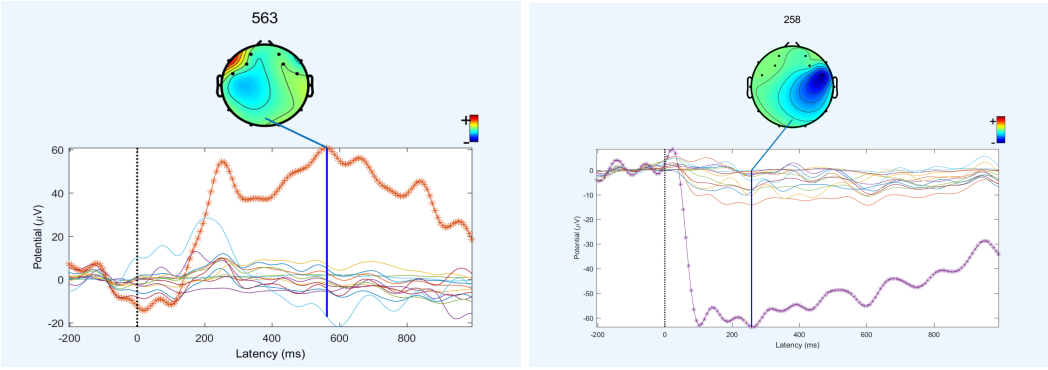
Figure 4.7: Emotiv Epoc electrodes scalp locations according to 10-20 positioning system [37].

During data preprocessing for classification, we used the EEG data collected from the 19 participants on Days 1 and 2, and for Day 2 we ended up with EEG data from 18 participants due to poor data quality for one of the participants. No classification tasks were done on Day 3. The number of passwords used for classification was 912 passwords from Day 1 (48 passwords per participant for 19 participants), and 360 passwords from Day 2 (20 passwords per participant for 18 participants). This resulted in EEG data for a total number of $1,272$ passwords to be used in the classification of *common* vs. *random* passwords.

Prior to analysis, raw data was pre-processed. First, a dual pass Butterworth filter with a passband of 0.1 Hz – 40 Hz was applied. This filter includes the bands that are of interest to us to be used during analysis. A notch filter of 59–61 Hz was also applied to the acquired data to remove noise that could have been caused by the power line frequency.

Data was then segmented into 68 segments per subject, each segment was for a duration of one second after the onset of the stimulus, and was sorted –using the markers– by stimulus type, so *common* passwords and *random* passwords had separate files.

Next, we conducted artifact detection which comes from two main sources. The first source is physiological artifacts, which include muscle activity, eye movement, and blinks. The second source is external artifacts, such as movement of an electrode due to headset movement, line noise, and head swaying and swinging. To minimize artifacts at recording time, participants were asked to avoid movement as much as possible. Each segment was inspected manually, to exclude signals that showed eye movement or muscle artifacts. EEGLAB provides a tool which allows the rejection of data based on visual inspection. The data of a total number of 13 channels were removed. An



(a) Participant ID#08 - *common* (b) Participant ID#09 - *random*

Figure 4.8: An example of channels removed by visual inspection.

example of two channels which were removed is shown in Figure 4.8, the channels' data marked with asterisks were removed. The same was done for data collected on Day 2, and data from both sessions were combined per participant for further processing.

Artifacts were further removed through two steps: an Independent Component Analysis (ICA), and an amplitude filter. ICA is a linear decomposition method and is commonly used to identify and cancel eye blinking and movement artifacts. When implemented, it decomposes the EEG raw data into signals filtered from the channel data with the time of each activity. ICA produced 14 signal components (equal to the number of channels fed in) for each segment, which were visually inspected to remove the ones that indicated artifacts. Of these possibilities, components that are concentrated in sensors towards the front of the head are selected to be removed first, as these are most likely blinks or eye movements. Signals that have high activity concentration in a single location (typically spiking in two or three sensors) are also flagged for removal, as these are most likely muscle movements (Pion-Tonachini et al. [87] provide a helpful tutorial on how to tell components apart on their EEG crowd labeling project [104]). An average of 6 signal components per participant's entire recording was removed.

The amplitude filter was applied with a voltage cut-off value of $\pm 100\ \mu$V, which resulted in the removal of 266 passwords from the $1,272$ passwords considered for classification: 188 passwords were removed from Day 1 data, and 78 passwords were removed from Day 2 data. The number of remaining passwords used for classification after preprocessing was $1,006$.

Next, feature scaling was used. In this process, the value of each feature in the data is set to have zero-mean. This is done by subtracting the mean from each feature, then dividing the feature value by its standard deviation.

Finally, the data was averaged over multiple trials for the same stimulus category for the same subject, where classification will be done individually per participant.

For analysis of password recall over short-term and long-term memory, the number of passwords in the analysis was not related to the EEG recordings. The number of passwords used was as follows:

- Short-term memorability: The immediate recall of *common* and *random* passwords within 10 seconds on the first session, $1,292$ passwords (68 passwords per participant for 19 participants). Note that recall data was Boolean data (1-Correct Recall, 0-Incorrect Recall) recorded by the application, and was usable data regardless of the EEG signals recording.

- Long-term memorability over the period 24–48 hours: The recall of $PWDCH$ on the second session, 15 passwords (19 participants minus four participants who reported writing down their password).

- Long-term memorability over the period 8–10 days: The recall of $PWDCH$ on the third session, 10 passwords (14 returning participants minus four participants who reported writing down their password).

Table 4.6 summarizes the number of passwords used, and the number of participants whose data are used in each analysis task.

4.3 Perceived Password Memorability User Study

In this user study, we run an experiment to investigate password recall, users' judgment of passwords memorability, and the effect of password strength on users' perception of password memorability. To investigate users judgment of passwords memorability, participants are presented with a number of passwords, during which their EEG data are recorded. After the presentation of passwords, participants are asked to rank the passwords from most-memorable to least-memorable based on their personal opinion. They were asked explicitly not to consider the password strength in their ranking.

Table 4.6: Number of passwords used in each analysis task.

Analysis	Day 1	Day 2	Day 8	Total # of Passwords
[H1, H2] Usable EEG data for password classification (EEG)	724 Passwords (19 Participants)	282 Passwords (18 Participants)	- -	1,006
Usable data for Recall (Boolean):				
[H1] Short-Term Recall	912 Passwords (19 Participant)	380 Passwords (19 Participants)	- -	1,292
[H2] Long-Term Recall (24-48 Hours)	- -	15 Passwords (15 Participants)	- -	All
[H2] Long-Term Recall (8-10 Days)	- -	- -	10 Passwords (10 Participants)	All

To investigate password recall from long-term memory, participants were given two passwords to memorize, and a rehearsal tool to practice the two passwords; these two passwords are generated based on seed words provided by the participants. Participants were then asked to recall the passwords at three different times, the first time is on the same day by the end of the session, then on the second day, and finally on the eighth day of the experiment.

Details of the experiment hypotheses, design, and data collection are described below.

4.3.1 Hypotheses

(1) Is it possible to predict a password's perceived memorability as judged by the users, based on EEG signals elicited upon presenting that password?

Hypothesis 4 (H4) *It is possible to predict how a user* ***perceives a password memorability*** *based on EEG data collected using a BCI.*

(2) Does password strength affect the way users judge a password memorability? In other words, does the human brain recognize password strength and perceive stronger passwords as less memorable, hence less usable?

Hypothesis 5 (H5) *There is a correlation between password strength and its perceived memorability.*

(3) Is it possible to use machine learning and EEG data collected upon presentation of passwords to predict long-term memorability for passwords?

Hypothesis 2A (H2A) *It is possible to predict passwords recall from* ***long-term memory*** *based on EEG data collected from participants upon presentation of the passwords.*

Note that hypothesis **H2A** is a retesting of **H2** with a larger sample and some difference in study design.

4.3.2 Study Design

We designed an experiment to measure and collect data to test the hypotheses from Section 4.3.1. There are two tasks in this experiment, and we will clarify what each task tests.

In the first task, password recall over long-term memory is tested (H2A). At the beginning of the session, participants were asked to memorize two passwords, generated using seed words provided by the participants. Participants were then assigned a second task which served as a distraction task, and at the same time was designed to test other parts of the hypotheses. By the end of the second task, participants were asked to recall the two passwords from earlier in the session; they were also asked to recall the same two passwords on the second day and eighth day of the experiment using an online system.

In the second task, predicting perceived password memorability (H4), and password strength effect (H5) are tested. Participants are presented with a number of blocks, each containing five passwords, while their EEG data is recorded during presentation. Once the presentation of one block is completed, participants are presented with the five passwords in one grid and asked to rank them based on how they judge the passwords memorability.

To explore the effect password strength has on perceived memorability (H5), no EEG data was used. Memory judgment data collected from the participants while they ranked the passwords, was combined with password guessability scores and explored through different statistical tests.

All the passwords used in this task were extracted from the 2012 LinkedIn password leak where 6.5 million passwords were leaked [65]. A random subset of $22,000$ passwords was taken out of the password leak and brute-forced to recover the true unhashed passwords. The subset of passwords was then streamed through the password guessability estimator and assigned a corresponding strength score between 0 and 4. Based on these scored passwords, we created five bins of passwords, labeled

Table 4.7: Names of the five bins created based on passwords strength.

Bin Name	Description	Score
weakest	too guessable	0
weak	very guessable	1
somewhat weak	somewhat guessable	2
strong	safely unguessable	3
strongest	very unguessable	4

Table 4.8: The *weakest* password list.

construction	monkeymonkey	relationship	qwertyqwerty	championship
123qweasdzxc	abc123abc123	abcdefghijkl	1qaz2wsx3edc	professional
lovelovelove	efrainefrain	sn00pysn00py	xirtamxirtam	wenwenwenwen

as *weakest*, *weak*, *somewhat weak*, *strong*, and *strongest*. The names and corresponding scores of each bin are shown in Table 4.7. The labels are chosen to indicate the strength of the passwords in each bin, where *weakest* bin has passwords of score 0, and *strongest* bin has passwords of score 4.

Each bin consisted of passwords selected from the processed password set mentioned earlier. In order to eliminate a known bias that users strongly prefer simple noun bi-grams which are common in natural language when choosing passwords [19], we only used passwords with the same length of 12 characters for the five bins. A total number of 15 passwords were randomly selected from each bin, and each participant was shown the same passwords in random order.

Passwords were presented to participants in 15 blocks; each block contained five passwords, one password from each of the five bins. Each password appeared on the screen for three seconds, and one second of a blank screen in between passwords. Once the five passwords appeared to the participant individually, they were asked to rank them in the next screen. Participants were asked to rank the five passwords from 1 being the most memorable to 5 being the least memorable, and to press enter once done. The same procedure is repeated for the 15 blocks.

Passwords appeared in black font on a white background and were centered in the middle of the screen. The perceived password memorability ranking task is summarized in Figure 4.9.

In this task, the focus is on the two bins *weakest* and *strongest*, since these are the two extreme bins of the ranking scale of the passwords, hence the difference –if any– should be more visible than the difference between the middle bins (*weak*, *somewhat weak*, and *strong*). Table 4.8 and Table 4.9 list the passwords in the *weakest* and *strongest* bins respectively.

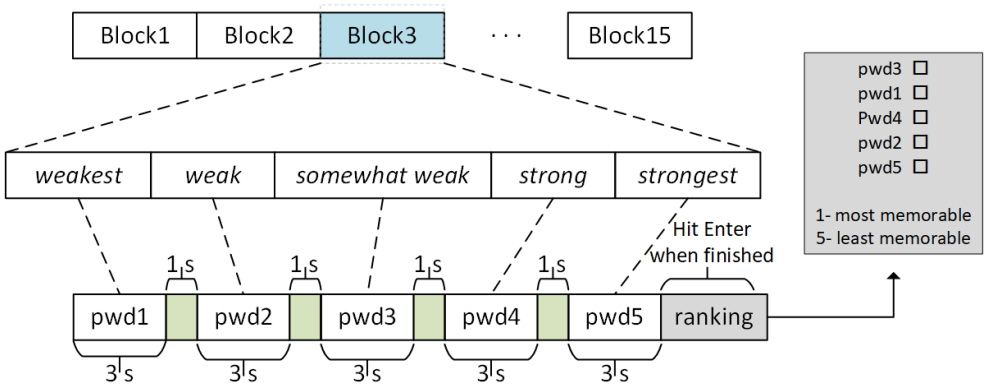


Figure 4.9: Summary of the memory judgment task in the *Perceived Memorability* user study.

Table 4.9: The *strongest* password list.

280275s10810	VEERANJANEYA	Brygadzistka	irq7pozitron	INDRAPRASTHA
nishiyashiki	mammaigan777	Vistemboire9	blomsterbarn	draawsneklav
chemiometria	pizzakoerier	Rom828atwtfg	imdonkarnash	Maerchenmond

4.3.3 Participants

We recruited 77 volunteers to participate through a university-wide email advertisement, 24 females and 53 males. Participants were first asked to fill an eligibility questionnaire (questionnaire in Appendix B), and if deemed eligible were contacted to schedule the experiment's first session. Upon arrival for the experiment, participants signed a consent form (Consent form replicated in Appendix C.2) in accordance with procedures approved by the University's Research Ethics Board, and were asked to fill a short pre-study questionnaire to collect demographical information. Volunteers were mostly undergraduate students with an average age of 21.2 and standard deviation of 2.9. All participants had a normal or corrected-to-normal vision and declared to not have a history of neurological or mental disorders. When asked on a scale of 1 (novice) to 5 (expert), "how would you rate yourself with respect to your computer skills?" 11 rated their computer skills as 5, and 40 participants answered 4, whereas 25 rated their computer skills as 3, and only 1 participant rated their computer skills as 2. Participants came from diverse fields of study shown in Table 4.10. Participants were compensated with $10 for their participation, distributed over the three sessions.

Table 4.10: Distribution of programs among participants.

Program	Count
Engineering and Applied Science	27
Health Sciences	17
Sciences	11
Social Science And Humanities	8
Business and Information Technology	7
Commerce	5
Education	2
Total	77

4.3.4 Experiment Procedure

The experiment consisted of one in-lab session and two on-line sessions. On the first day in the lab, the experiment protocol was explained to participants in detail. The Muse headband was then fitted to the participant, and they were asked to look at a 17-inch monitor about 50 centimeters in front of them. The experiment consisted of three sessions distributed over three days as explained below.

Day 1: On the first day (Session 1), participants were asked to provide two words, each is a minimum of 4 characters and a maximum of 8 characters, which will be used as seed words to generate two passwords for the study. These two seed words will be referred to as *Seed1* and *Seed2*. Once *Seed1* is entered and meets the requirements, the same algorithm used in the password recall user study (See Section 4.2.5) is used to generate a password, i.e., Capitalize, Replace, and Insert. This password will be referred to as *Password1*. The participant is prompted to enter the second seed word, and the same procedure is followed to generate *Password2*. Next, the generated *Password1* is presented to the user for a period of three seconds while the EEG signals are being recorded using the BCI headband. Finally, a rehearsal tool is provided to the participants, to practice typing the password for as long as they need to, once done, they hit the Enter key to proceed to the second password. The same procedure is followed for *Password2*.

After the two passwords are generated, and the participant finished practicing them, the second task starts. In this task, participants are asked to rank passwords based on their judgment on how memorable the passwords measure as shown in Figure 4.10. They are presented with 15 blocks; each block contains five passwords. Each of the five passwords appears individually in the center of the screen for a period of three seconds. Once the first block is presented, a screen appears prompting the participants to rank the passwords as shown in Figure 4.11.

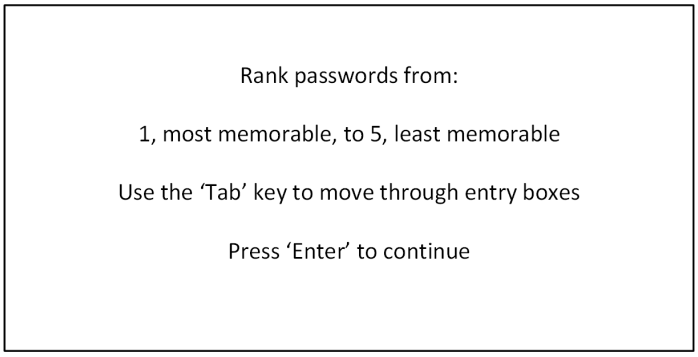


Figure 4.10: Directions to rank passwords as shown to the participants.

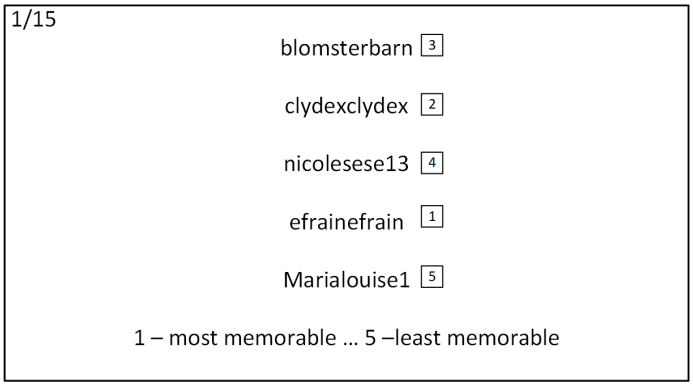


Figure 4.11: A screen prompts the participants to rank the passwords previously presented to them.

Once the 15 blocks are completed, participants are asked to reproduce or recall *Password1* and *Password1* from the first task. This recall is considered a recall from long-term memorability since the distraction task took around 25 minutes. If they failed to recall any of the two passwords three times, they were given the option to see that password and practice typing it. At this point, Session 1 is completed, and participants are advised that they will receive an email with a link to the online system to recall their password on the second day, and another email one week later. A summary of the experiment procedure is shown in Figure 4.12.

Day 2: On the second day (Session 2), participants were sent an email containing a link to the online system, reminding them to login for the second session. The participants were asked to reproduce the two passwords they were assigned on Day 1, and after three failed attempts had the option to see the forgotten password, and practice it.

Day 8: On the eighth day (Session 3), the same procedure from session 2 was followed, participants received an email containing a link to the online system, reminding them to login for the last session. Participants were asked to recall *Password1* and *Password2*, and they were redirected to fill a post-study questionnaire, where they were explicitly asked if they wrote down the password during the study. Again, participants were encouraged to answer truthfully, and the answers were used to exclude participants who answered "Yes" to this questions from recall data analysis.

For days 2 and 8, a 24-hour flexibility window was allowed to accommodate participants' schedules, so participants were asked to login within a 24–48 hour window for days 2 and 8, after which the session was no longer available.

4.3.5 Data Acquisition - Muse

EEG data was recorded using Muse headband by InteraXon [61] shown in Figure 4.13. The Muse default sampling rate for data recording of 220 Hz was used (see [50] for technical specifications of the headband). The Muse headband has four channels: TP9, AF7, AF8, and TP10. Channel locations are shown in Figure 4.14. MuseIO was used to connect the computer to the Muse headband over Bluetooth; MuseIO is a tool that connects to and streams data from Muse headband, it provides access to raw EEG, power bands (alpha, beta, delta, gamma, and theta), and blink + jaw clench detection. It sends a stream of Open Sound Control (OSC) messages containing Muse data that other programs can receive. The recording was done via the Open Sound Control protocol (OSC), using a combination of Python, and Muse's SDK. Using a custom application, a marker was placed

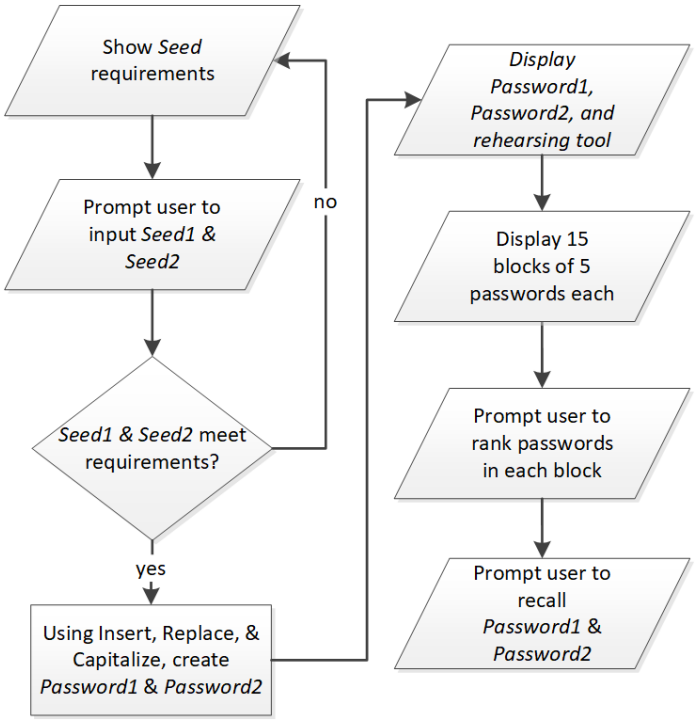


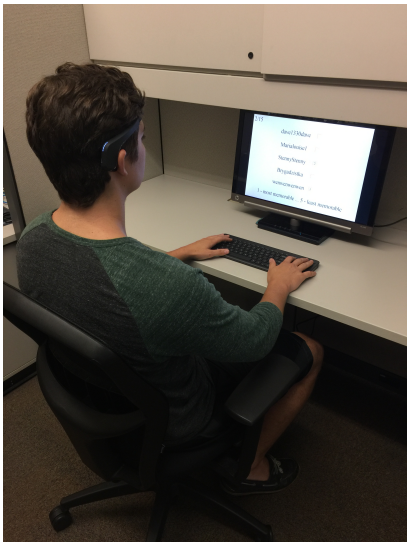
Figure 4.12: Summary of password *Perceived Memorability* user study procedure on Day 1.

upon the onset of the stimulus (the password). The markers have five unique values to differentiate between passwords bins.

4.3.6 Data Pre-Processing

EEG data was recorded during the complete period of the experiment. Data recorded using the Muse headband has a DC offset of around 800 μV. Hence the first step of pre-processing the raw data was to demean each of the four channels. Muse headband comes with a number of preset configuration; researchers can choose which preset to be used. The preset we used in the experiment applied a notch filter of 60 Hz on the acquired data at the time of data recording; this notch filter helps remove noise that could be caused by the power line frequency. Next, a dual pass Butterworth filter with a passband of 0.1–40 Hz was applied. This filter includes the bands that are of interest to us during analysis. Data was then segmented into 75 segments per participant; each segment was for a total duration of 1.3 seconds, which spanned 300 ms prior to and 1000 ms after the onset of the stimulus. Segments were baseline corrected using a 300 ms window preceding stimulus onset. Next, artifact detection was performed and artifacts were removed through two steps. First, remove

(a) Muse headband.



(b) A participant is taking part in the experiment.

Figure 4.13: Muse headband used in the *Perceived Memorability* user study experiment.

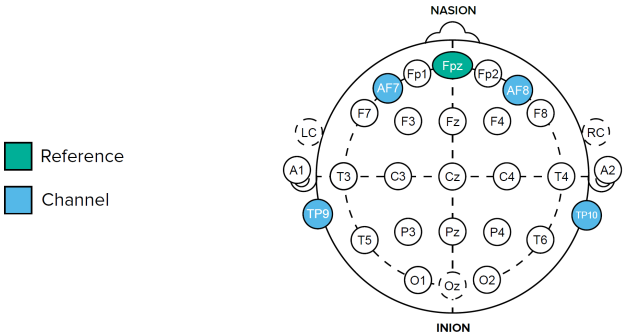


Figure 4.14: Muse electrode locations by 10-20 International Standards [50].

segments where the variance in the segment exceeds a threshold value of 100 μV, and second, remove segments where the maximum voltage reading of the signal exceeds 100 μV. This process resulted in the removal of an average of five segments per participant. In addition, the EEG data for two of the participants were removed, due to noisy low-quality data, ending up with usable data from 75 participants. Feature scaling was also implemented on the dataset.

After EEG data preprocessing, we ended up with the EEG data collected from 75 participants on Days 1. This resulted in the EEG data for $5,625$ passwords (75 passwords $\times$ 75 participants). Finally, the EEG data of $1,125$ passwords in the *weakest* bin were averaged across participants for each channel, and the same was done for the $1,125$ passwords in the *strongest* bin; yielding 120 rows of data: 4 channels $\times$ 15 passwords (averaged over 75 participants) per bin.

For analysis of password recall over long-term memory, recall data for two passwords of 75 participants were used, resulting in 150 passwords for the recall analysis. The EEG data for *Password1* and *Password2* underwent the same pre-processing of demeaning, filtering, baseline correction, and artifact detection, however, no segments were averaged in the password recall data. Recall data were studied per participant based on whether that participant successfully recalled *Password1* and *Password2*.

4.4 Feature Extraction and Selection

EEG signals feature extraction is a subject of broad and current interest in the BCI community. Studies comparing methods of feature extraction [4, 112, 115], channel selection [7], and classifiers' performance [71, 59, 8] have been done extensively, yet feature extraction and selection remain the most challenging task when working with EEG data, and there is no agreed upon optimum feature set to be used. We extracted features from three domains: frequency domain, time domain, and time-frequency domain.

The frequency domain feature set was composed by estimating the power of each frequency band. The frequency bands used are delta (0–4 Hz), theta (4–8 Hz), alpha (8–12 Hz), and beta (12–40 Hz). The power estimate was done for the four bands per signal per channel. The feature set containing the power estimates of the frequency bands will be referred to as Power Spectrum FS.

The time domain feature set was created by extracting commonly used statistical features of the EEG time sequence data. Seven features were extracted: Mean, Maximum, Minimum, Standard

Deviation, Variance, Skewness, and Kurtosis. The statistical features will be referred to as Statistics FS.

The time-frequency domain feature set was generated using Discrete Wavelet Transformation (DWT) [26]. Wavelet representation lies between the spatial and Fourier domains [72]. When using wavelets to perform signal processing, the first step is selecting a suitable mother wavelet. The most important wavelet families are Haar, Daubechies, Symlets, Coiflets, and biorthogonal. Among the various wavelet families, the Daubechies family of wavelets is known for its orthogonality property and efficient filter implementation [56]. Subasi [98] compared the effect of wavelet family on EEG signals classification accuracy, tests were carried out using Daubechies of order 4 (db4), Symmlet of order 10 (sym10), Coiflet of order 4 (coif4), and Daubechies of order 2 (db2). It was noticed that the Daubechies wavelet gives better accuracy than the others, and db4 is slightly better than db2. Therefore, we employed Daubechies as the mother wavelet in this dissertation, with an order 4 decomposition of EEG signals using filtering, and decimation to obtain the approximation and detailed coefficients. Decomposition of EEG signals was done into five levels for the EEG data collected in the *Password Recall* user study, and into six levels for the EEG data collected in the *Perceived Memorability* user study. The number of levels the signal is decomposed into depends on the sampling frequency. Emotiv Epoc headset sampled the data at 128 Hz, whereas the preset used in the Muse headset sampled the data at 220 Hz.

The 5-level wavelet decomposition tree for *Password Recall* user study is shown in Figure 4.15. The feature set consisted of the detail coefficients ($D_3..D_5$) and the last approximation coefficient (A_5). *Perceived Memorability* user study 6-level wavelet tree is shown in Figure 4.16. The feature set consisted of the detail coefficients ($D_3..D_6$) and the last approximation coefficient (A_6). This feature set will be referred to as Wavelet Coefficients FS.

Figure 4.17 shows a 5-level wavelet decomposition of the EEG signal for Participant ID#4 in the password recall user study. This sample signal presents 6000 readings over the period of about 46 seconds, with a sampling rate of 128 Hz. The discrete wavelet transform is used to isolate signal components. An input signal is split into a low pass and high pass sub-bands, which are also known as approximation and detail sub-bands. A special filter is used for the first level of analysis. For subsequent analysis, the approximation sub-band is further split into approximation and detail sub-bands. The whole process is repeated at each level.

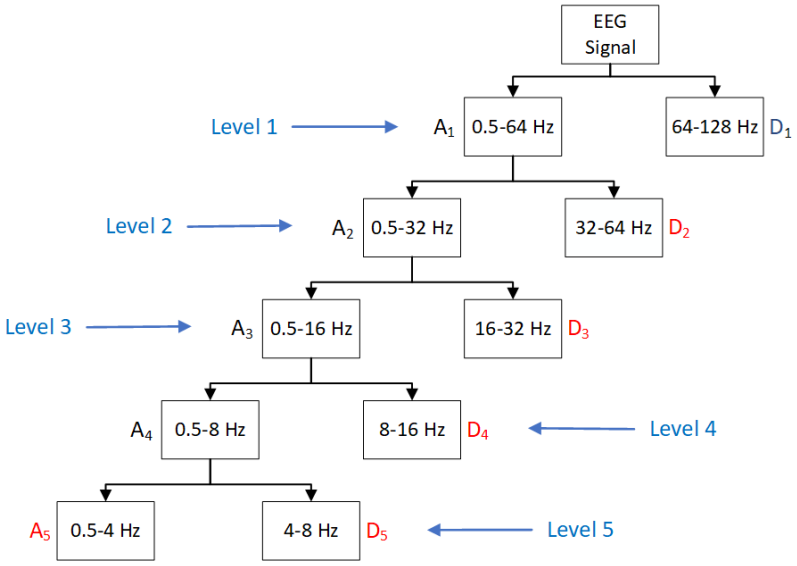


Figure 4.15: 5-level wavelet decomposition and corresponding frequencies. The EEG signals are decomposed into several frequency bands, where D_i is the detail coefficient and A_i is the approximation coefficient ($i = 1, 2, ..., 5$). The detail coefficients ($D_2, ..., D_5$) and the last approximation coefficient (A_5) are used as the feature set.

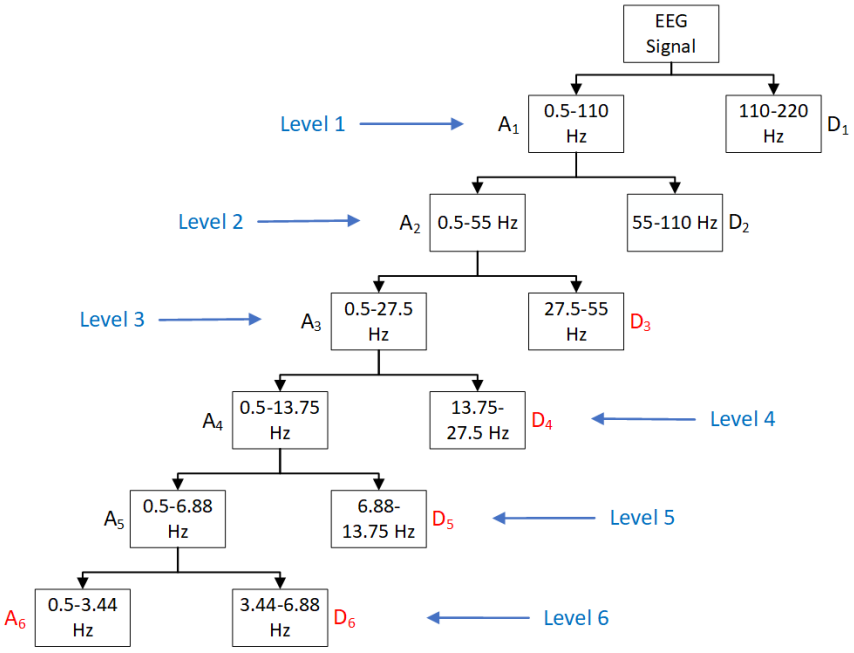


Figure 4.16: 6-level wavelet decomposition and corresponding frequencies. The EEG signals are decomposed into several frequency bands, where D_i is the detail coefficient and A_i is the approximation coefficient ($i = 1, 2, ..., 6$). The detail coefficients ($D_3, ..., D_6$) and the last approximation coefficient (A_6) are used as the feature set.

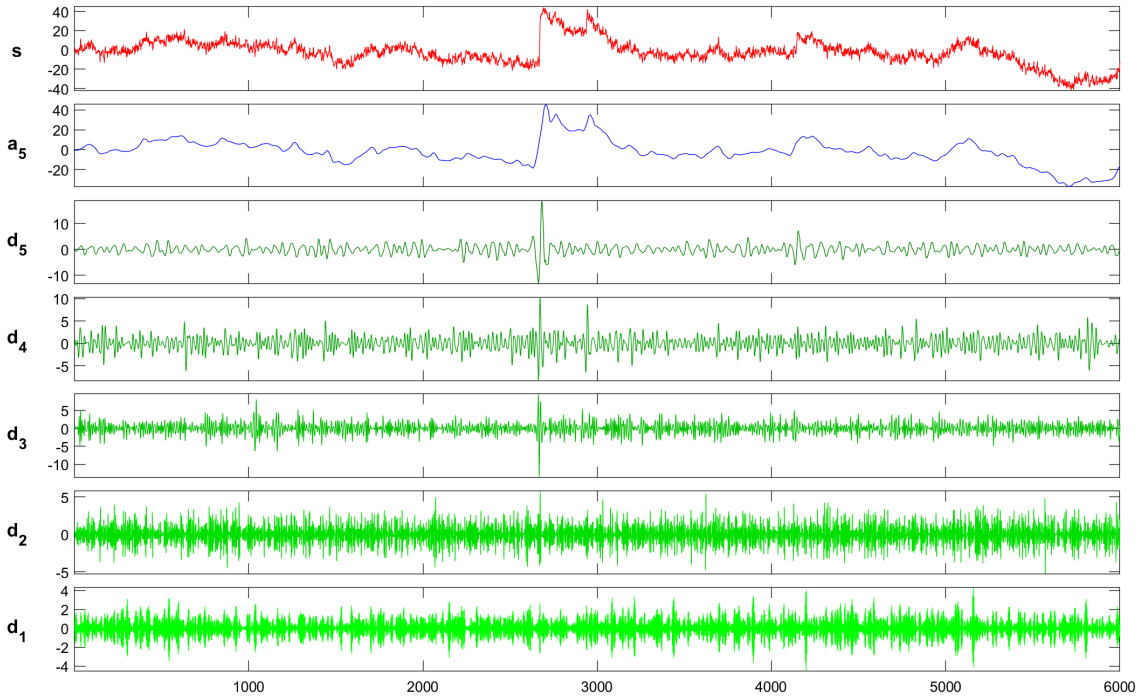


Figure 4.17: Decomposition at level 5: $s = a_5 + d_5 + d_4 + d_3 + d_2 + d_1$.

Table 4.11: Sample of signal decomposition of the first second - Participant ID#4.

s	1.08	2.05	-0.38	-1.80	0.21	-0.06	-3.63	-3.77	-1.14	-1.97	-3.69	-1.52	0.92	0.36	-0.80
$\mathbf{a}_5$	0.07	0.92	0.48	0.27	2.17	-4.34	4.92	25.51	24.48	-19.85					
$\mathbf{d}_5$	0.08	0.34	2.58	0.31	-8.11	3.60	-6.72	2.51	5.42	-4.34					
$\mathbf{d}_4$	0.02	0.30	2.91	3.00	8.21	-10.58	1.23	1.43	8.93	0.87	-2.39	1.19	1.94	-4.78	
$\mathbf{d}_3$	-0.97	-3.07	1.67	-0.53	0.89	8.24	9.37	3.26	2.08	4.70	-1.44	1.16	-8.56	-4.93	3.60
$\mathbf{d}_2$	0.22	0.48	-0.86	0.11	1.23	-2.64	-2.44	0.81	0.63	-2.70	-0.58	0.40	-1.56	1.96	1.24
$\mathbf{d}_1$	-0.57	-0.88	1.67	-2.12	2.18	-1.83	1.39	-0.83	0.16	0.38	-0.67	0.30	0.50	-0.67	0.37

Figure 4.18 depicts the original signal and the approximation at level 5, and Figure 4.19 shows the complete set of the signal and its approximations on the left side, and the details coefficients in the right side of the figure for the same participant.

Part of the signal and its coefficients are shown in Table 4.11. Note that this table shows only the first 15 values of each coefficient due to space limitations, and that each coefficient length is different. The length of the signal and its decompositions are 128, 10, 10, 14, 22, 37, and 67, which correspond to s, a_5, d_5, d_4, d_3, d_2, and d_1 respectively.

The three feature sets were combined to create a fourth feature set which will be referred to as Combined FS.

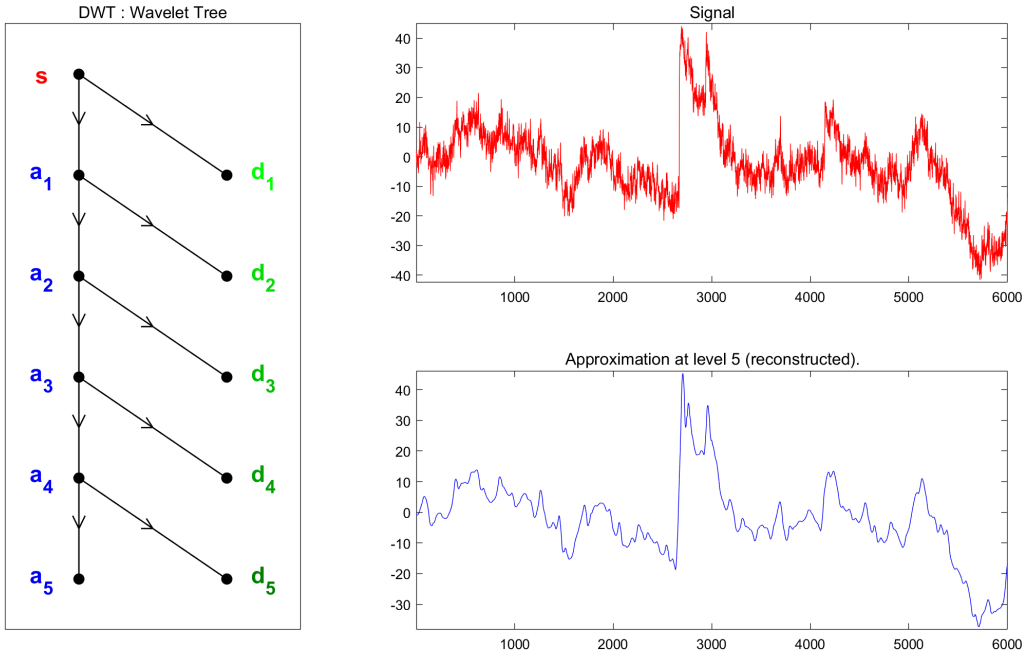


Figure 4.18: The original signal and the approximation at level 5 - Participant ID#4.

Once features are extracted, it is important to implement a feature selection method. Feature selection role is to eliminate redundant and insignificant features to improve prediction accuracy and reduce computational cost. A problem with feature selection of EEG signals is the high dimensional nature of the features space and often the small sample size. A number of feature selection methods has been proposed. Germán et al. [91] proposed a feature selection method that uses Least Angle Regression for ranking each feature; then, a Leave-One-Out estimation was used to choose the most relevant features of the EEG signals. Garrett et al. [46] and Flotzinger et al. [42] used a genetic algorithm approach, which minimizes the number of features taken for classification while maximizing classification performance. McFarland and Wolpaw [75] used multiple regression to select weighted combinations of features. They found that feature weights obtained from previous data generalized well to new datasets.

Two feature selection methods are implemented in this dissertation, lasso (least absolute shrinkage and selection operator), and stepwise regression.

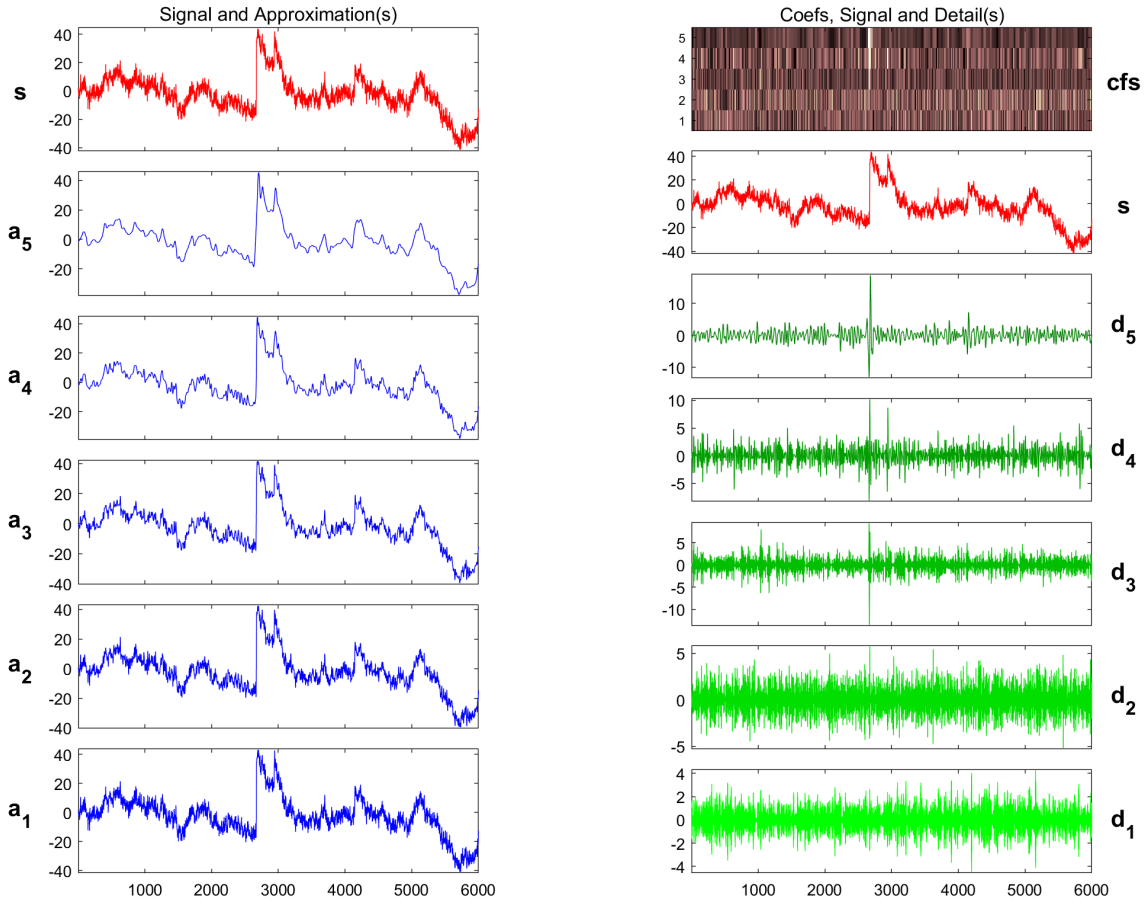


Figure 4.19: The set of approximations and details coefficients, each shown separately - Participant ID#4.

4.4.1 Lasso regression

Lasso is a regression analysis method that uses l_1 penalization [53]. It can be used to select features to enhance prediction performance [106]. In the lasso regularization, the coefficients of features which do not influence the prediction become zero or near to zero. A common practice in data analysis is to estimate the coefficients of a linear model via lasso and choose the regularization parameter via cross-validation [53]. To choose the tuning parameter λ for lasso, a 5-fold cross-validation was used.

The four feature sets (Power Spectrum, Statistics, Wavelet Coefficients, and Combined) were input to the lasso algorithm individually, and features selected by lasso from each set were used for classification. However, some feature sets yielded no features to be selected, and all coefficients of these feature sets were shrunken to zero or near to zero due to weak predictors.

4.4.2 Stepwise regression

Stepwise regression has been used in feature selection of EEG signals [63, 74, 109, 70]. A combination of forward and backward stepwise regression is implemented. Starting with no initial model coefficients, the most statistically significant predictor variable having a $p < 0.05$ is added to the model. After each new entry to the model, a backward stepwise regression is performed to remove the least significant variables. This process is repeated until no additional terms satisfy the entry/removal criteria [70]. It is important to note that stepwise regression is a computationally intensive process, and may be unfeasible to use when a large number of features is present.

The same procedure of feature selection using lasso was followed with stepwise regression, except that due to the number of features in each set, we only fed stepwise regression model with feature sets Statistics FS and Wavelet Coefficient FS. Once feature sets are extracted, and feature coefficients are calculated, features were fed into an SVM classifier. More details on each specific classification task are provided in Chapter 5. A summary of the preprocessing, feature extraction, feature selection, and classification process is shown in Figure 4.20.

4.4.3 Password Recall User Study

The number of features in each feature set in the Password Recall study is shown in Table 4.12. The number of features selected by lasso is reported in Table 4.13 based on password recall and password strength respectively. Stepwise regression did not select any features from Statistics FS,

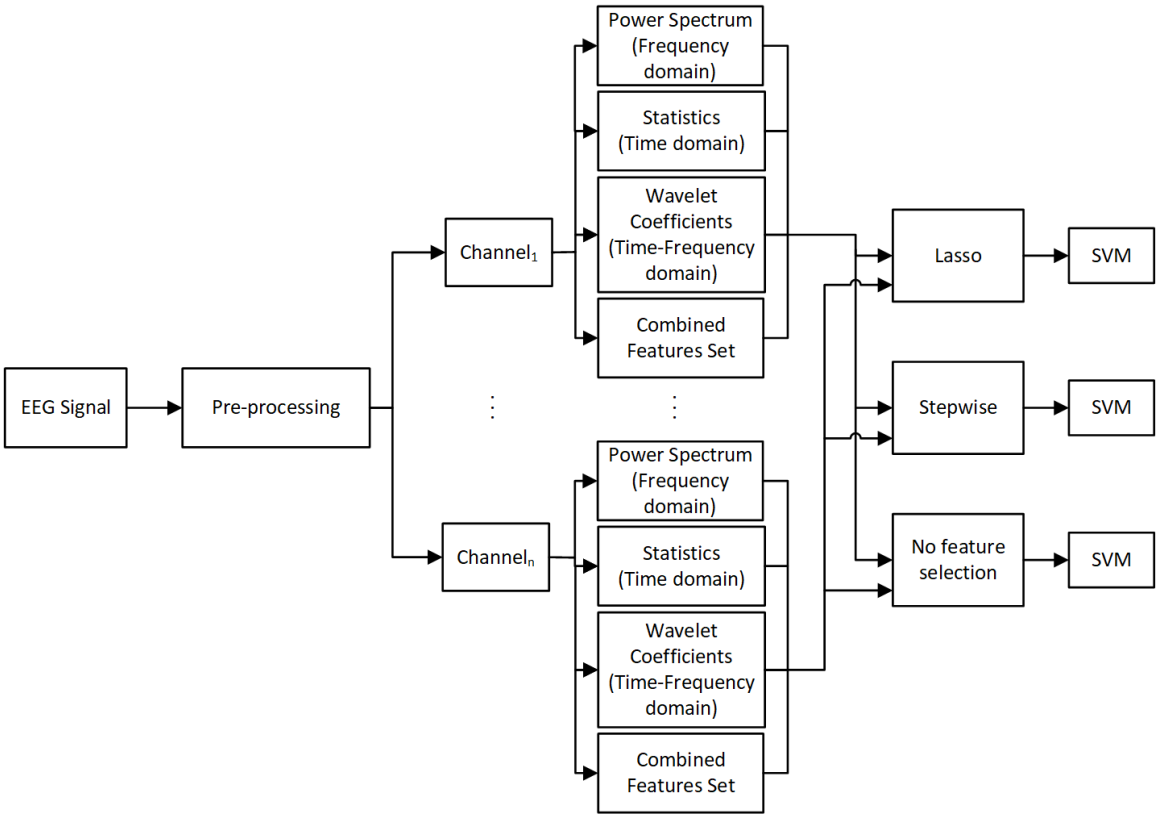


Figure 4.20: System diagram of data preprocessing, feature extraction, feature selection, and classification, where $n = 14$ for the data collected by the Emotiv Epoc, and $n = 4$ for the data collected by the Muse headband.

Table 4.12: The number of features in each set.

Feature Set	# of Features
Power Spectrum FS	516
Statistics FS	7
Wavelet Coefficients FS	95
Combined FS	618

Table 4.13: The number of features selected by lasso regression for short-term memory recall.

Participant ID	Power Spectrum FS	Statistics FS	Wavelet Coefficients FS	Combined FS
01	0	2	4	2
02	4	4	8	1
03	1	6	1	3
04	0	1	1	9
05	8	4	8	2
06	1	5	2	1
07	3	3	1	1
08	0	5	2	4
09	5	5	4	3
10	1	3	3	1
11	1	0	9	9
12	5	0	4	5
13	5	0	4	2
14	5	1	2	9
15	5	4	6	1
16	1	2	0	0
17	0	0	7	5
18	0	0	0	0
19	0	0	01	0

and selected the number of features shown in Table 4.14 from the Wavelet Coefficient FS based on password recall from short-term memory and password strength respectively.

Once feature selection was performed per participant, we looked at the selected features and chose the top five features that were repeatedly selected among most of the participants. These five features were then used as the feature set for classification. Note that the number of features is greatly reduced when using lasso regression, as well as stepwise regression. The reduction of the number of features helps improve performance and thus classification speed.

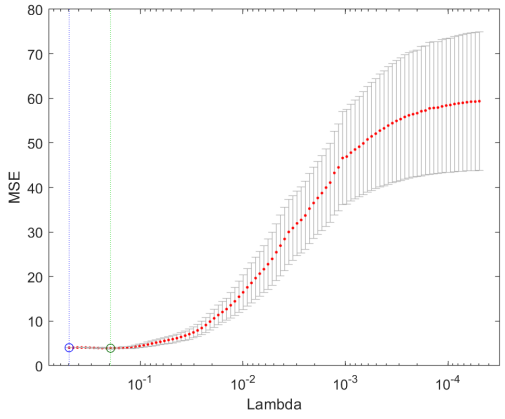
Table 4.14: The number of features selected by stepwise regression for short-term memory recall.

Participant ID	Wavelet Coefficients FS
01	4
02	3
03	1
04	2
05	5
06	3
07	2
08	2
09	5
10	1
11	1
12	4
13	3
14	1
15	3
16	1
17	2
18	2
19	4

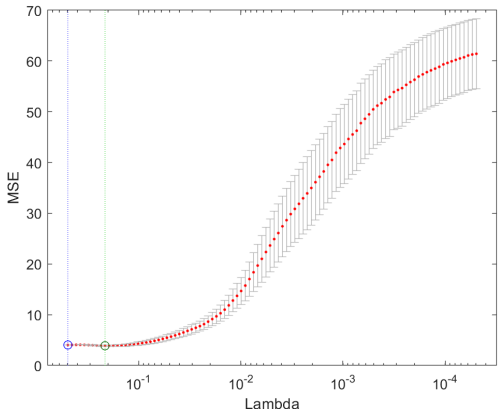
4.4.4 Perceived Memorability User Study

The λ values chosen for Wavelet Coefficients and Combined feature sets, based on the user password memorability ranking are shown in Figure 4.21 (a) and (b) respectively. Feature sets Power Spectrum and Statistics produced zero coefficients and had no features selected by lasso. For each λ in Figure 4.21, an estimate of the mean squared prediction error on new data for the model fitted by lasso with that value of λ, with the error bars of each estimate is shown. The green circle and dashed line indicate the value of λ with a minimum cross-validated mean squared error (MSE). The blue circle and dashed line indicate the greatest λ that is within one standard error of the minimum MSE. Figure 4.21 (c) and (d) shows the trace plots of the values in β against the $l1$ norm of β. Each colored line in the trace plot represents the value taken by a different coefficient in the model (the values in β of a single predictor variable).

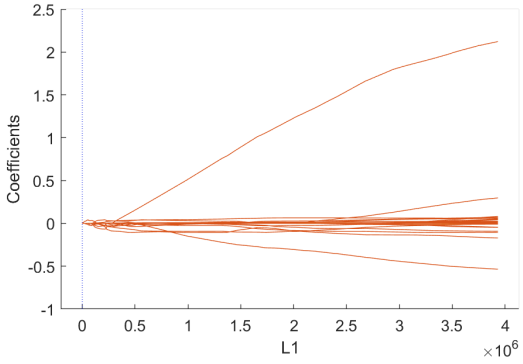
The λ values chosen for Power Spectrum and Wavelet Coefficients feature sets based on password strength ranking, are shown in Figure 4.22 (a) and (b) respectively, and Figure 4.22 (c) and (d) shows the trace plots based on the same. Feature sets Statistics and Combined produced zero coefficients and had no features selected by lasso.



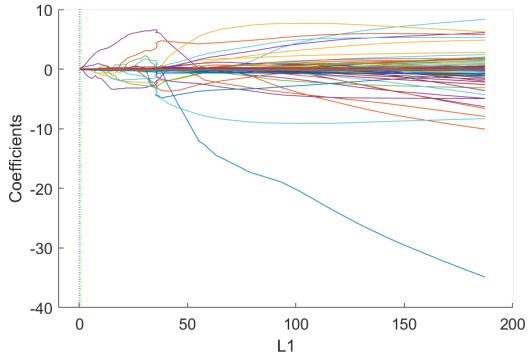
(a) Cross-validated MSE of lasso fit for the combined features set.



(b) Cross-validated MSE of lasso fit for the wavelet coefficients set.

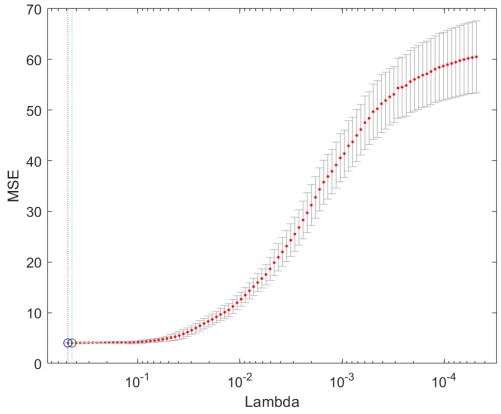


(c) Trace Plot of coefficients fit by lasso for the combined features set.

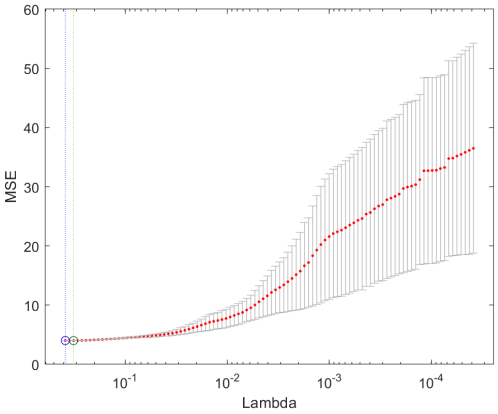


(d) Trace Plot of coefficients fit by lasso for the wavelet coefficients set.

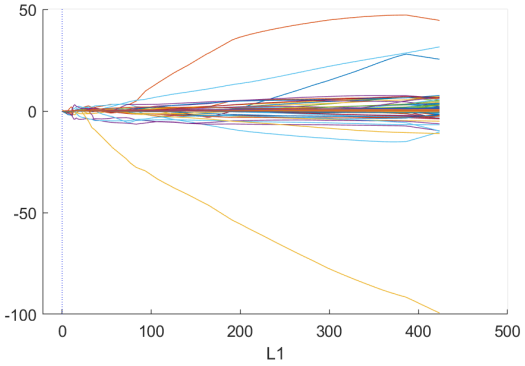
Figure 4.21: Cross-validation and trace plots for different feature sets based on perceived password memorability ranking.



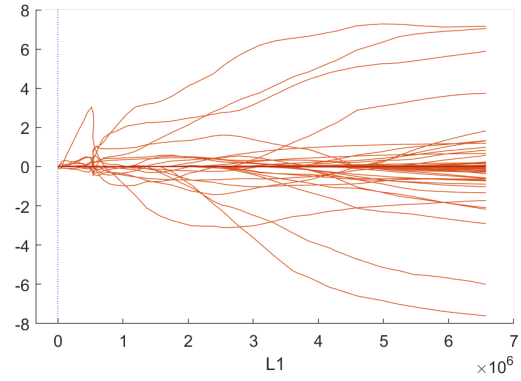
(a) Cross-validated MSE of lasso fit for the wavelet coefficients set.



(b) Cross-validated MSE of lasso fit for the power bands set.



(c) Trace Plot of coefficients fit by lasso for the wavelet coefficients set.



(d) Trace Plot of coefficients fit by lasso for the power bands set.

Figure 4.22: Cross-validation and trace plots for different feature sets based on zxcvbn password strength ranking.

Table 4.15: The number of features selected by lasso and stepwise regression based on perceived memorability and password strength, *Perceived Memorability* user study.

Feature Set	# of Features	lasso regression PM[1]	lasso regression PS[2]	Stepwise regression PM[1]	Stepwise regression PS[2]
Power Spectrum FS	645	0	3	*	*
Statistics FS	7	0	0	0	0
Wavelet Coefficients FS	102	12	3	4	6
Combined FS	754	23	0	*	*

* Computationally too intensive to be calculated in a single core.
[1] Perceived Memorability.
[2] Password Strength.

Stepwise regression did not select any features from Statistics FS, and selected 4 and 6 features from Wavelet Coefficient FS based on password memorability ranking and password strength ranking respectively.

The number of features in each feature set and the number of selected features by both methods are shown in Table 4.15.

Chapter 5

Results and Analysis

5.1 Introduction

In this chapter, we present the results obtained from running the two user studies explained in Chapter 4, along with the performance measures used to evaluate these results. It is worth mentioning that we present the results in this chapter based on the hypotheses testing in the user studies rather than the chronological order the studies were conducted in. The reason for this hypothesis based presentation is that we retested one hypothesis in the second user study, and it will be optimal to present the results of the hypothesis and its revisiting at the same time.

To determine if passwords that are more appealing to a user as being more memorable produce distinctive EEG signals, and to find if these signals are different enough to predict users' behaviour such as password recall (H1, H2, H2A) and perceived password memorability (H4), we performed password binary classification using a Support Vector Machine (SVM). To investigate password strength effect on its recall (H3) and memorability judgment (H5), a number of statistical tests were performed.

Different classification tasks were carried out to study password recall (successfully recalled vs. not recalled passwords), perceived password memorability (most memorable vs. least memorable), and password strength (weakest vs. strongest, and common vs. random), based solely on the EEG signals.

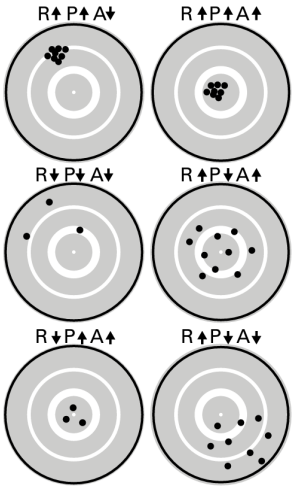


Figure 5.1: Bull's-eye illustration of the differences among recall (R), precision (P), and accuracy (A). Up-and-down arrows indicated high and low levels. Recall is illustrated by the number of dots, precision is illustrated by the spread of the dots, and accuracy is illustrated by the distance of the dots away from the center of the bull's eye.[1]

SVM classification was done using 10-fold cross-validation. Precision, recall, F-score and accuracy measures were used to evaluate the classifier performance; these are all common measures of performance evaluation, and are defined in equations 5.1, 5.2, 5.3, and 5.4 respectively.

Figure 5.1 illustrates the concept of precision and recall, and how they relate to accuracy. An ideal classifier will result in most of the dots being in the center of the bull's eye, indicating high recall, precision, and accuracy. Another effective measure to compare the performance of different features and classifiers is the area under the ROC curve (AUC).

$$Precision = \frac{TP}{TP + FP} \tag{5.1}$$

$$Recall = \frac{TP}{TP + FN} \tag{5.2}$$

$$F = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{5.3}$$

[1]Reprinted from "Analyzing Neural Time Series Data: Theory and Practice", Mike X. Cohen, p 25, Copyright (2014), with permission from The MIT Press.

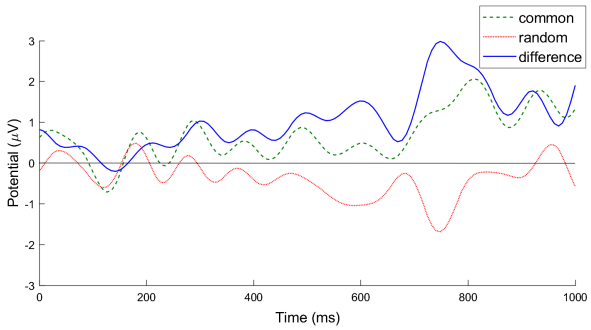


Figure 5.2: Grand average of the *common* and *random* elicited EEG signal and the difference between the two categories.

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{5.4}$$

5.2 Password Recall

We looked at the EEG data signals collected when users were presented with the *common* and *random* password lists. The average (over multiple participants) data for each class of passwords was calculated as the mean amplitude over one-second windows. This allowed us to determine if there is a human-visible difference to begin with before further analyzing the data. The grand average of *common* vs. *random* elicited EEG signals per stimulus type is shown in Figure 5.2. The result showed a distinction between the EEG signals elicited when the subjects saw *common* passwords and when they saw *random* passwords. Signals produced from *random* passwords were, in general, lower in voltage than those from *common* passwords. These EEG signal averages are done for all the common passwords shown to the subjects and the same for all the random passwords. Similar signal characteristics were seen in averaged signals per participant. Although there was a general pattern of differentiation across participants, it was not the same for every one.

To find out if a machine learning classifier could detect the difference between *common* and *random* passwords recall based on the EEG signals, we used a Support Vector Machine (SVM). Password recall was studied in both experiments. In the Recall Study, EEG data was collected

upon password presentation of *common* and *random* passwords. After which, participants did an immediate recall for the passwords they saw. Participants also did a recall from long-term memory for one password they chose over the periods of 24–48 hours and 8–10 days. In the Perception Study, participants were asked to memorize two passwords, at the beginning of the session, and were asked to recall them at the end of the session; and again in 24–48 hours and 8–10 days. Passwords are retrieved from memory by a process called recall, which requires the person to reproduce the password from memory. The following terms were defined for the classifier, based on the success or failure to recall a password by the participants:

- True Positive (TP): A *correctly recalled* password correctly classified as a *correctly recalled* password.
- True Negative (TN): An *incorrectly recalled* password correctly classified as an *incorrectly recalled* password.
- False Positive (FP): An *incorrectly recalled* password incorrectly classified as a *correctly recalled* password.
- False Negative (FN): A *correctly recalled* password incorrectly classified as an *incorrectly recalled* password.

5.2.1 Short-Term Memory Password Recall

Participants in the Recall Study performed a short-term password recall (within 10 seconds) for 68 passwords divided over two days. 48 passwords on Day 1, and 20 passwords on Day 2. The results shown in Table 5.1 are for the successful recall from short-term memory on Day 1. The percentages are by password category per participant. As noted before, the recall of passwords in the *common* category is higher than those in the *random* category. These results imply that random passwords are harder to recall even momentarily. Despite the general trend of *common* passwords having a much higher successful recall than the *random* passwords, individual differences in recall can be noticed clearly when it comes to *random* passwords. A quarter of the participants had a higher than 30% success rate of recall for random characters.

Due to the individual differences in password recall, the results are reported per participant rather than averaging the EEG data across participants. EEG data for the recall classification were

Table 5.1: Recall from Short-Term memory success percentage for passwords by category type - Day 1 (H1).

Participant ID	*common* passwords	*random* passwords	Participant ID	*common* passwords	*random* passwords
01	100%	30%	11	100%	36%
02	100%	18%	12	94%	0%
03	97%	21%	13	97%	48%
04	97%	12%	14	100%	43%
05	97%	27%	15	94%	9%
06	97%	30%	16	100%	36%
07	97%	15%	17	91%	9%
08	97%	3%	18	100%	15%
09	91%	3%	19	97%	15%
10	97%	12%			

averaged per participant, where EEG data for successfully recalled passwords were averaged over the 14 channels, and the same was done for unsuccessfully recalled passwords. This resulted in 28 rows per participant data file. Each data file was used to extract features in the three domains discussed earlier, and then each feature set was fed to a classifier three times, once with no feature selection methods, once using lasso feature selection, and using stepwise regression whenever applicable. We report here on the wavelet feature set only in Table 5.2, since it produced the highest results. Results from the other feature sets and the feature selection methods are reported in Appendix A.1.

5.2.2 Long-Term Memory Password Recall

To test long-term memorability, recall data collected from both experiments were used. The collected EEG data of the user chosen password $PWDCH$ upon presentation in the first experiment, coupled with the participant recall on days 2 and 8 are used to investigate password recall from long-term memory.

To test for correlation between passwords' EEG and long-term recall we ran two Pearson correlation tests with an $\alpha = 0.05$: First, test the correlation between the password mean EEG data and the participant success in recalling the password. We run the test for both days 1 and 8, for participants who completed the session on each day (excluding participants who reported writing down the password). Second, test the correlation between the password mean EEG data and the number of attempts it took the user to recall the password correctly. The number of attempts was set to

Table 5.2: SVM Classifier performance for classifying password short-term recall, *recalled* vs. *not recalled* - Wavelets FS (H1).

Participant ID	**Precision**	**Recall**	***F*-Score**	**Accuracy**	**AUC**
1	0.93	0.93	0.93	0.93	0.92
2	0.86	0.86	0.86	0.86	0.87
3	1.00	0.57	0.73	0.79	0.88
4	0.80	0.86	0.83	0.82	0.94
5	0.88	1.00	0.93	0.93	0.87
6	1.00	1.00	1.00	1.00	1.00
7	1.00	0.50	0.67	0.75	0.85
8	0.93	1.00	0.97	0.96	0.96
9	0.87	0.93	0.90	0.89	0.92
10	1.00	0.50	0.67	0.75	0.96
11	0.69	0.79	0.73	0.71	0.85
12	0.90	0.64	0.75	0.79	0.93
13	0.88	1.00	0.93	0.93	0.87
14	0.82	1.00	0.90	0.89	0.94
15	0.88	1.00	0.93	0.93	0.98
16	0.93	0.93	0.93	0.93	0.97
17	1.00	0.50	0.67	0.75	0.85
18	0.93	0.93	0.93	0.93	0.92
19	0.62	0.93	0.74	0.68	0.77
Average	0.89	0.83	0.84	0.85	0.91

Table 5.3: Pearson Correlation Coefficient r (H2).

	Day 2 (N=15)	Day 8 (N=10)
EEG Mean vs. Correct Recall	0.375	0.321
EEG Mean vs. # of Attempts	-0.356	-0.526

1, 2, and 3 when the participant authenticated successfully on the first, second, and third attempt respectively. If the participant did not recall the password successfully, the number of attempts was set to 5.

Table 5.3 shows the results of the Pearson Correlation test. A positive correlation was found between the EEG signals mean and the successful recall on both days 2 and 8, though the correlation weakens on day 8. A negative correlation with the number of attempts was noticed on days 2 and 8.

Although we found a positive correlation between the EEG signal mean and the password recall, and a negative correlation between the EEG signal mean and the number of attempts; the correlation was not significant when considering the sample size tested as will be discussed in Chapter 6.

Because the sample size was not large enough to make a conclusive decision, we decided to re-test the correlation of long-term recall to EEG data in the second experiment.

In the *Perceived Memorability* study, the collected EEG data upon presentation of the two passwords were used, coupled with the recall data on the three days 1, 2 and 8. To analyze the data, we started by testing the recall data on the first day, where participants were asked to memorize two passwords at the beginning of the session, and after performing the password memorability ranking task, they were asked to recall the passwords *Password1* and *Password2* at the end of the session. The session lasted for about 30 minutes, leaving around 25 minutes period between assigning the passwords and recalling them. Recall within that period of time is considered recall from the long-term memory store. The EEG and recall data of *Password1* and *Password2* on days 2 and 8 were also used to test recall from long-term memory.

A 10-fold cross-validation using the SVM classifier was conducted to determine if a password recall can be predicted using the EEG data of the two passwords along with their recall data collected in the experiment. The data set used for password recall prediction had imbalanced classes. For example, on the first day, the number of successfully recalled passwords among the 75 participants was 126 passwords, whereas the number of passwords that were not successfully recalled was 24. This caused the classifier to produce 100% accuracy but with low precision when first ran with the default cost options. The cost matrix for the classifier was modified to reflect the imbalanced classes, where a higher cost was assigned when misclassifying the forgotten passwords (recall=0), in comparison to the remembered passwords (recall=1). This causes the classifier algorithm to update the prior probabilities by incorporating the penalties described in the modified cost matrix.

The classifier results for predicting password recall from long-term memory are shown in Table 5.4, the data used in the password recall prediction task was 150 passwords, 2 passwords per participant. Each row consisted of that password EEG data averaged over the four channels. The labels of correctly recalled and incorrectly recalled are the same defined at the beginning of this section. The Wavelets FS set is used, and the ROC curves are shown in Figure 5.3.

5.3 Perceived Password Memorability

Data collected from the *Perceived Memorability* Study is used in this classification task. EEG data was collected from participants upon presenting 15 blocks of 5 passwords each, and asking the

Table 5.4: SVM Classifier performance for classifying password recall from long-term memory on Days 1, 2, and 8 - Wavelets FS (H2A).

Day	**Precision**	**Recall**	***F*-Score**	**Accuracy**	**AUC**
Day 1	0.93	0.84	0.88	0.81	0.86
Day 2	1.0	0.17	0.30	0.31	0.59
Day 8	1.0	0.07	0.13	0.22	0.54*

* 1-AUC.

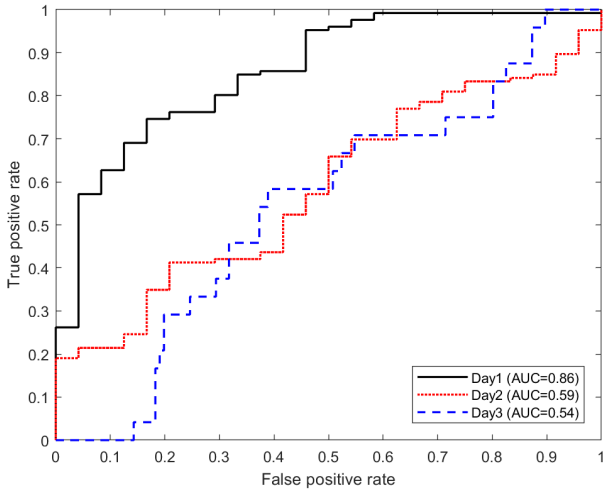


Figure 5.3: ROC curve for the recall from long-term of *Password1* and *Password2* - Wavelets FS (H2A).

participants to rank the passwords based on how they judge their memorability. To determine if there are distinctive EEG features of a password that makes it more appealing to a user as being more memorable, and to find if these features are present enough to predict perceived password memorability, password binary classification was done based on two labels, *most memorable* and *least memorable*. The following terms were defined for the classifier, based on the labels provided by the participants:

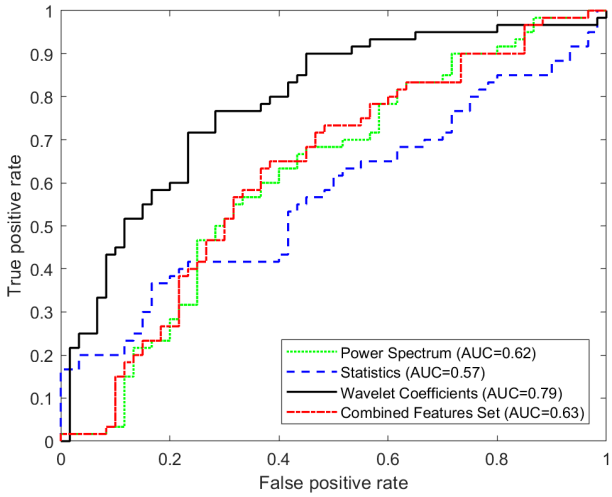
- True Positive (TP): A *most memorable* password correctly classified as a *most memorable* password.
- True Negative (TN): A *least memorable* password correctly classified as a *least memorable* password.
- False Positive (FP): A *least memorable* password incorrectly classified as a *most memorable* password.
- False Negative (FN): A *most memorable* password incorrectly classified as a *least memorable* password.

EEG data was used in feature extraction, and the four feature sets, Power Spectrum, Statistics, Wavelet Coefficients, and Combined were fed to the classifier. Each feature set was used three times, one time without any feature selection (labeled as none), the second time the features selected by lasso (labeled as lasso) were used, and the third time the features selected by stepwise regression were used (labeled as stepwise).

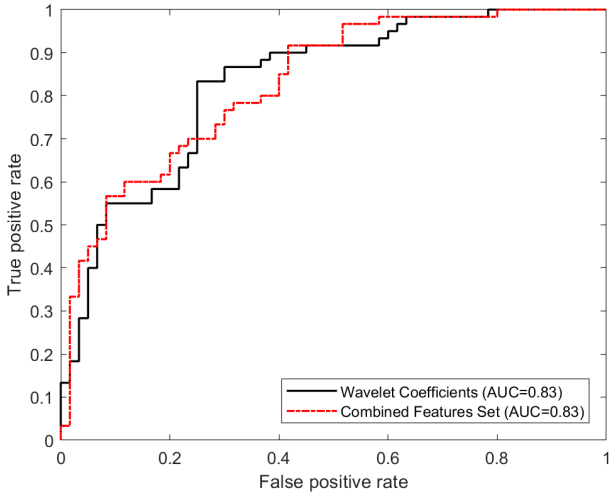
Table 5.5 shows the precision, recall, F-Score, and accuracy obtained from the SVM classifier. In addition to these performance measures, we report the AUC under the ROC curve in Figure 5.4, based on user memorability judgment ranking.

5.3.1 What makes a password more memorable?

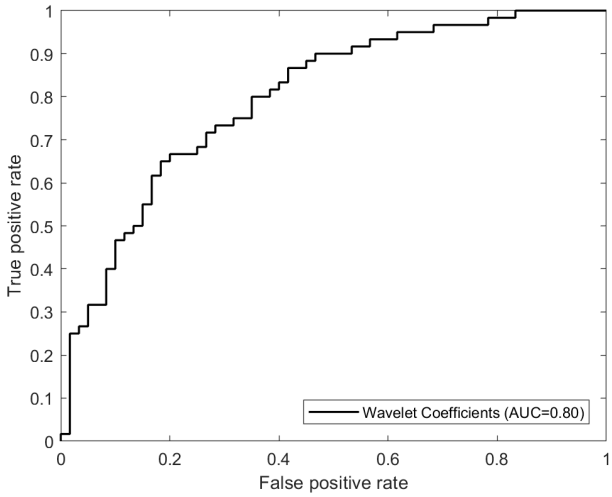
We asked the participants in the *Perceived Memorability* study, what makes a password more or less memorable to them? participants were generous in their replies. Their responses are included in Appendix D, after removing any identifying personal information. Responses collected from participants were grouped into five reasons that make a password more memorable, meaningful words,



(a) none (no feature selection methods applied)



(b) lasso



(c) stepwise

Figure 5.4: ROC curves and AUCs for the SVM classifier using different feature selection methods. ROC curves based on perceived password memorability (H4.)

Table 5.5: Classifier performance based on perceived password memorability ranking (H4).

Selection Method	Features Set	Precision	Recall	F-Score	Accuracy
none	Power Spectrum	0.70	0.23	0.35	0.57
	Statistics	0.51	0.58	0.55	0.52
	Wavelet Coefficients	0.79	0.55	0.65	0.70
	Combined Feature Set	0.71	0.25	0.37	0.58
lasso	Power Spectrum	n/a	n/a	n/a	n/a
	Statistics	n/a	n/a	n/a	n/a
	Wavelet Coefficients	**0.80**	**0.86**	**0.83**	**0.82**
	Combined Feature Set	0.90	0.64	0.75	0.79
stepwise	Power Spectrum	*	*	*	*
	Statistics	n/a	n/a	n/a	n/a
	Wavelet Coefficients	0.70	0.75	0.73	0.71
	Combined Feature Set	*	*	*	*

n/a Not applicable, as no features were selected by the algorithm (e.g., weak predictors.)
* Computationally too intensive to be calculated in one core.

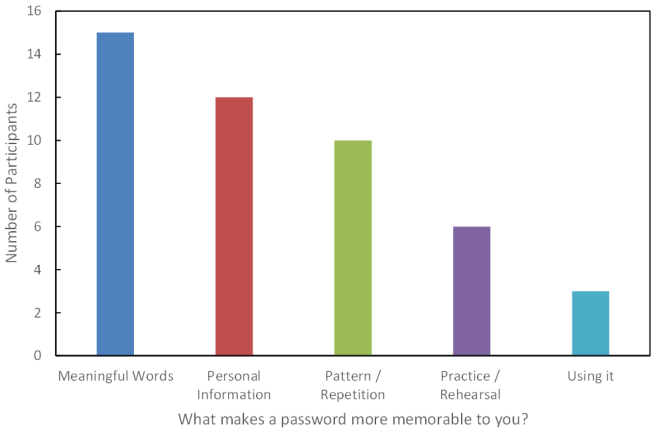


Figure 5.5: Characteristics that make a password more memorable as reported by participants.

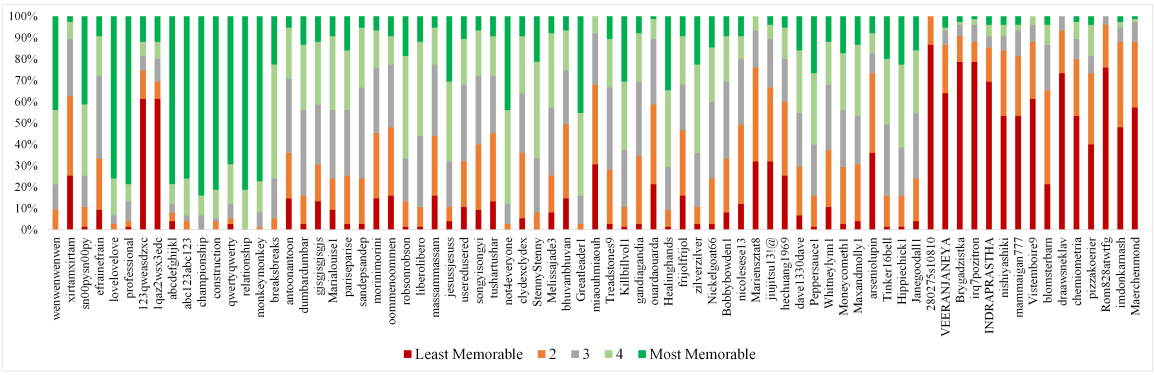


Figure 5.6: Users' ranking of password memorability per password.

personal information, patterns, practice, and password frequent use. The results are shown in Figure 5.5. Note that the question was an open question, and these categories were not presented to the participants. Meaningful words and common words were the number one reason to make a password more memorable reported by 15 participants; next was personal information which was mentioned 12 times. Patterns and repetition in a password were listed 10 times by the participants. Practicing typing the password made them more memorable for 6 participants, whereas 3 participants reported that using a password is what makes it more memorable.

For participants who chose to list what makes a password less memorable to them, a general trend was, as put by one of the participants, "Gibberish doesn't help with remembering." Symbols or special characters, numbers, random letters, and leet are reasons that make a password less memorable as reported by 9 participants.

Next, we looked at how participants ranked the passwords. Each of the 75 different passwords was ranked 75 times (by the 75 participants). Figure 5.6 shows how the individual passwords measured in terms of memorability. Passwords were next grouped into six categories, based on their lexical characteristics. These categories are shown in Figure 5.7.

5.4 Password Strength

5.4.1 Password Strength vs. Recall

In the password recall experiment, participants did password recall from short-term memory for the passwords they saw for 10 seconds period. They also did recall from long-term memory for one

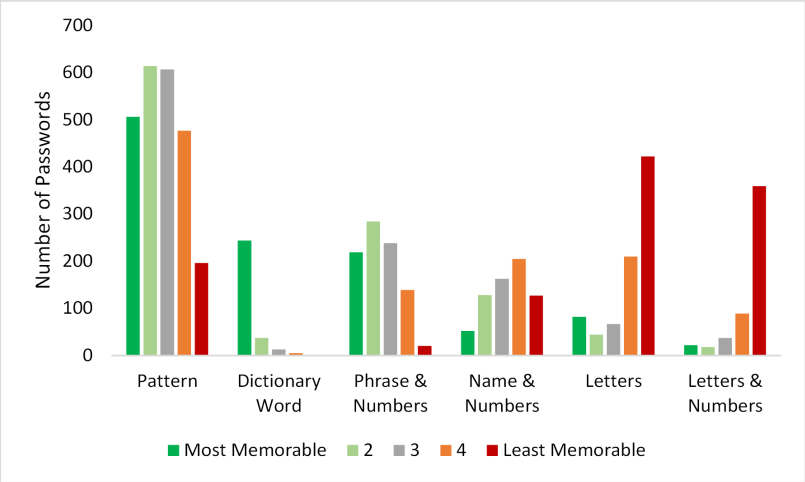


Figure 5.7: Users' ranking of password memorability per category. Note there was only one password with symbols in it (password = *jijitsu!@*).

password in the recall study, and for two passwords in the perceived memorability study. We wanted to explore if there is a correlation between password estimated strength and its recall.

We start by exploring if password strength can be predicted using EEG data. The following terms were defined for the classifier, based on the password strength in the *Password Recall* Study:

- True Positive (TP): A *common* password correctly classified as a *common* password.
- True Negative (TN): A *random* password correctly classified as a *random* password.
- False Positive (FP): A *random* password incorrectly classified as a *common* password.
- False Negative (FN): A *common* password incorrectly classified as a *random* password.

Note that the passwords in the *common* list are considered weakest or too guessable with a guessability score of mostly 0, and the passwords in the *random* list are considered somewhat weak or somewhat guessable with a score of 2. We report here on the wavelet feature set only in Table 5.2, since it produced the highest results. Results from the other feature sets and feature selection methods are reported in Appendix A.2.

Table 5.6 shows the precision, recall, F-Score, and accuracy obtained from the SVM classifier, using the wavelets feature set. Accuracies presented in Table 5.6 are based on single-trial analysis, which considers the variance within participants only [85]. However, when averaging the channels' data over all participants, a classification accuracy of 89.3% was achieved.

Table 5.6: SVM Classifier performance for classifying password strength *common* vs. *random* - Wavelets FS.

Participant ID	Precision	Recall	*F*-Score	Accuracy	AUC
1	1.00	1.00	1.00	1.00	1.00
2	1.00	0.57	0.73	0.79	0.94
3	1.00	0.50	0.67	0.75	0.77
4	1.00	0.93	0.96	0.96	1.00
5	0.76	0.93	0.84	0.82	0.95
6	0.93	1.00	0.97	0.96	1.00
7	0.91	0.71	0.80	0.82	0.89
8	1.00	0.93	0.96	0.96	0.97
9	0.91	0.71	0.80	0.82	0.94
10	0.93	1.00	0.97	0.96	0.99
11	0.82	1.00	0.90	0.89	0.94
12	0.88	0.50	0.64	0.71	0.94
13	0.79	0.79	0.79	0.79	0.88
14	0.91	0.71	0.80	0.82	0.89
15	0.76	0.93	0.84	0.82	0.95
16	0.75	0.64	0.69	0.71	0.84
17	0.78	0.50	0.61	0.68	0.82
18	0.87	0.93	0.90	0.89	0.97
19	0.81	0.93	0.87	0.86	0.95
Average	0.88	0.80	0.83	0.84	0.93

Once we found a correlation between EEG data and password strength indicating that participants were sensing the password strength, we wanted to investigate if that affected password recall. The correlation between password strength and recall was investigated without using EEG data.

To test the effect of password strength (or guessability) on password recall, a paired-samples t-test was used to determine whether there was a statistically significant difference between the recall of passwords in the *common* list compared to passwords in the *random* list. Paired samples statistics and results of the t-test are shown in Table 5.8 and Table 5.9 respectively, where N is the number of participants, t measures the size of the difference relative to the variation in the sample data compared to a t-distribution (t-test), *df* is the degrees of freedom, which is N–1, and Sig. is the statistical significance value. Participants recalled more passwords from the *common* list (0.97 $\pm$ 0.02) as opposed to passwords in the *random* bin (0.20 $\pm$ 0.12), a statistically significant increase of 0.77 (95% Confidence Interval, 0.710 to 0.819, note the confidence intervals do not contain the number zero between 0.710 and 0.819, which indicates a statistically significant mean difference), $t(18) = 29.59$, $p < .0005$.

Table 5.7: Interpretation of values of Cohen's d.

	small size	medium size	large size
d	0.2	0.5	0.8

Table 5.8: Paired Samples Statistics

	Mean	N	Std. Deviation	Std. Error Mean
common	0.97	19	0.02	0.01
random	0.20	19	0.12	0.03

To ensure that indeed there is a primary effect of strength on recall, the effect size was calculated using *Cohen's d* formula [27]. The standard value of Cohen's d for small, medium, and large sizes are shown in Table 5.7. An effect size of 6.8 was found, which means there is a strong influence of the password strength on the user ability to recall it.

5.4.2 Password Strength vs. Perceived Memorability

In the *Perceived Memorability* study, the EEG data recorded while presenting passwords from different guessability score bins were used. Specifically, the *weakest* and *strongest* bins. This is similar to the analysis performed when comparing *common* and *random* bins. However, the difference in the study of password strength in both experiments is that passwords in the *weakest* and *strongest* bins have the same 12 characters length in both bins, they have a large difference in password guessability scores of 0 and 4, and the correlation to password memorability judgment is investigated. Whereas, in the *common* and *random* bins, passwords have different lengths, the difference in their guessability score is 0 and 2, one password list was random characters, symbols, and numbers (not a dictionary word), and the correlation to passwords recall is investigated. Passwords in both studies are real-world passwords (from password leaks and worst passwords lists), except for the *random* list which was generated using an online tool.

Table 5.9: Paired Samples Test

	Mean	Std. Deviation	Std. Error Mean	95% Confidence Interval Lower	95% Confidence Interval Upper	t	df	Sig.(2-tailed)
common - random	0.765	0.112	0.025	0.710	0.819	29.593	18	.000

The same EEG data collected to predict perceived password memorability in Section 5.3 were used to predict password strength, but based on password strength ranking rather than the user ranking. The following terms were defined for the classifier, based on their guessability score:

- True Positive (TP): A *weakest* password correctly classified as a *weakest* password.
- True Negative (TN): A *strongest* password correctly classified as a *strongest* password.
- False Positive (FP): A *weakest* password incorrectly classified as a *strongest* password.
- False Negative (FN): A *strongest* password incorrectly classified as a *weakest* password.

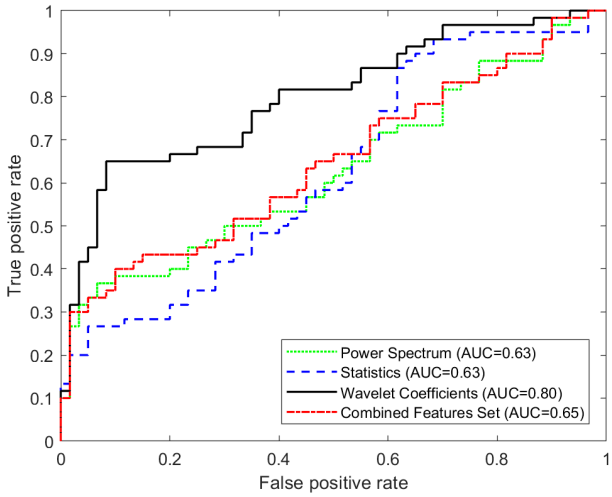
Note that in the classification of passwords based on password strength, no ranking data provided by the participants were used. The EEG data collected from participants were classified based solely on guessability ranking. The classification task was carried out following the same protocol in Section 5.3. The SVM classifier was fed with the four feature sets, Power Spectrum, Statistics, Wavelet Coefficients, and Combined. Each feature set was used three times, once with no feature selection, once with lasso feature selection, and once with stepwise feature selection (when applicable). The results obtained from the SVM classifier, based on the password strength, are reported in Table 5.10, and Figure 5.8 presents the ROC curves for each selection method.

Next, the correlation between participant ranking of how memorable a password is, and the password estimated strength is investigated. In this analysis, EEG data is not used. The perceived memorability ranking given by the participant is checked against the estimated guessability ranking.

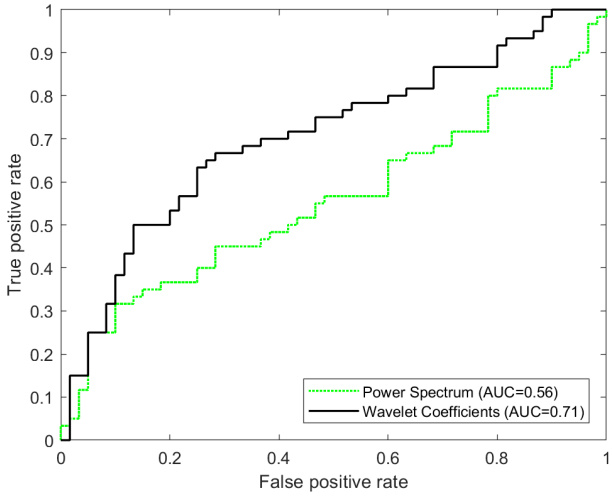
Figure 5.9 shows the ranking of the passwords based on their memorability as provided by the participants, versus the password strength ranking. The matrix shows that passwords in the two opposite bins of memorability are the most clear to participants, it also shows the correlation between how a user perceives password usability and its strength. That correlation, though, is not very clear within the matrix near the center where the difference in perceived memorability of passwords with close values of strength becomes less clear.

To test the effect of guessability on perceived memorability we look at the distribution of the data. The difference scores for the five bins were found to be normally distributed, as assessed by visual inspection of the Normal Q-Q Plot presented in Figure 5.10.

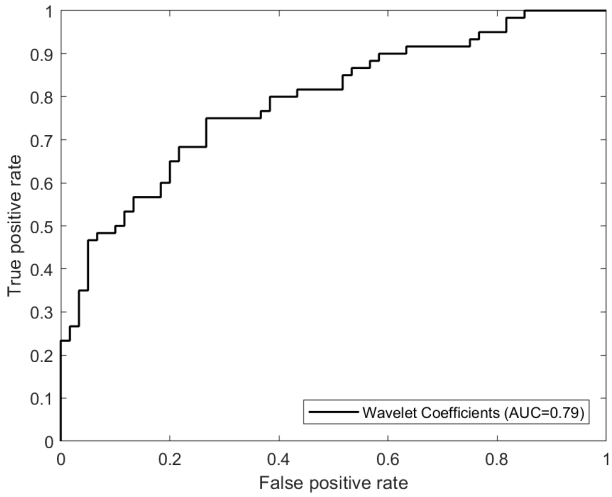
A paired-samples t-test was used to determine whether there was a statistically significant mean difference between the perceived password memorability when participants ranked passwords in the



(a) none (no feature selection methods applied)



(b) lasso



(c) stepwise

Figure 5.8: ROC curves and AUCs for the SVM classifier using different feature selection methods. ROC Curves based on password strength.

Table 5.10: Classifier performance based on password strength ranking.

Selection Method	Features Set	Precision	Recall	F-Score	Accuracy
none	Power Spectrum	0.79	0.38	0.52	0.64
	Statistics	0.59	0.32	0.41	0.55
	Wavelet Coefficients	**0.86**	**0.65**	**0.74**	**0.77**
	Combined Feature Set	0.77	0.35	0.48	0.63
lasso	Power Spectrum	0.76	0.32	0.45	0.61
	Statistics	n/a	n/a	n/a	n/a
	Wavelet Coefficients	0.70	0.58	0.64	0.67
	Combined Feature Set	n/a	n/a	n/a	n/a
stepwise	Power Spectrum	*	*	*	*
	Statistics	n/a	n/a	n/a	n/a
	Wavelet Coefficients	0.77	0.57	0.65	0.70
	Combined Feature Set	*	*	*	*

[n/a] Not applicable, as no features were selected by the algorithm, i.e., weak predictors are shrunken to zero.
[*] Computationally too intensive to be calculated in one core.

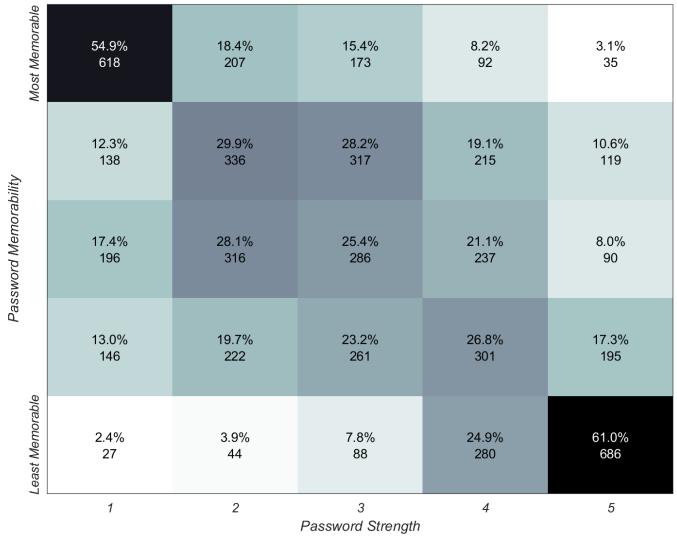


Figure 5.9: Password perceived memorability as ranked by the participants vs. password estimated strength.

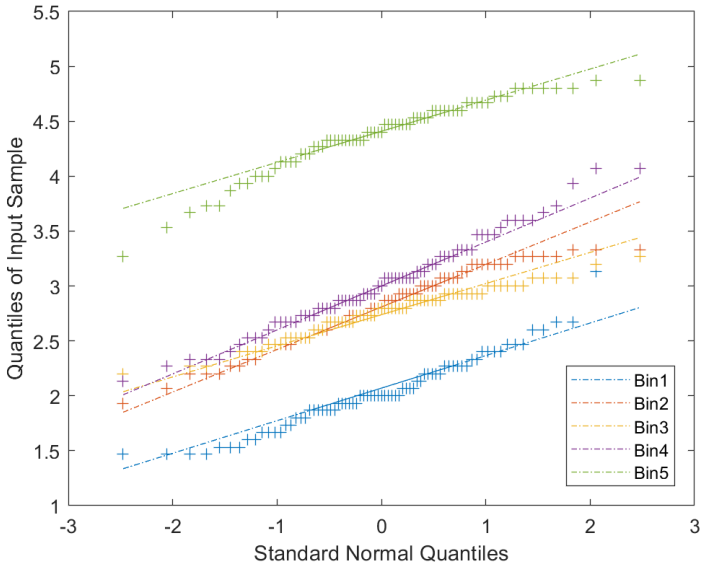


Figure 5.10: QQ Plot of Sample Data versus Standard Normal (H5).

Table 5.11: Paired Samples Statistics

	Mean	N	Std. Deviation	Std. Error Mean
weakest	4.38	75	.32	0.03
strongest	2.04	75	.36	0.04

weakest bin compared to passwords in the *strongest* bin. Paired samples statistics and results of the t-test are shown in Table 5.11 and Table 5.12 respectively. Participants ranked passwords in the *weakest* bin as more memorable (4.38 ± 0.32) as opposed to passwords in the *strongest* bin (2.04 ± 0.36), a statistically significant increase of 2.33 (95% Confidence Interval, 2.192 to 2.47), $t(74) = 33.013$, $p < .0005$.

To ensure that indeed there is a main effect of strength on recall, the effect size was calculated using *Cohen's d* formula, an effect size of 3.8 was found, which means there is a strong influence of the password strength on the user judgment of how memorable it is.

Table 5.12: Paired Samples Test

				95% Confidence Interval				
	Mean	Std. Deviation	Std. Error Mean	Lower	Upper	t	df	Sig.(2-tailed)
weakest - strongest	2.333	0.612	0.071	2.192	2.474	33.013	74	.000

Chapter 6

Discussion

In this chapter, the results obtained from the user studies and presented in Chapter 5 are discussed based on the hypothesis testing. The discussion of password recall, from both short-term and long-term memory, perceived password memorability, and password strength, is followed by an evaluation of the ecological validity and limitations of the studies.

6.1 Password Recall

Recall from short-term memory

Recall from short-term memory results in Table 5.1 showed that known easy-to-remember and hard-to-remember passwords have distinguishable EEG features that can be used to predict these passwords recall with an average accuracy of 85%, and an AUC of 0.91. It was noticed that during short-term memory recall for *common* and *random* passwords, participants were able to successfully recall passwords with higher EEG amplitudes more than those of lower amplitudes. This was true not only when comparing *common* to *random* passwords, but also within the random passwords category. Random passwords with higher EEG signals had a higher chance of being recalled than random passwords with lower EEG signals. Passwords that were recalled elicited higher voltages than passwords that were not recalled. This finding is in-line with the literature [92, 44, 83], where larger EEG amplitudes were noticed for words that had a higher chance of being subsequently remembered or recognized. It is worth noting here that there might be a number of reasons that are causing the EEG signals to be higher for the passwords that were successfully recalled later.

For example, the rise in mental activities, as detected by the BCI, could be attributed to the user attempting to mentally pronounce the password. Another reason for this rise, could be that the password is causing a mental imagery. These are valid reasons that we did not test, and require further investigation.

Recall from long-term memory

Initially, EEG signals were found to have a positive correlation with password recall from long-term memory on Day 2 with a Pearson coefficient of 0.375; this correlation weakened on Day 8 where Pearson coefficient was equal to 0.321. EEG signals also had a negative correlation with the number of attempts to recall the password with a coefficient value of -0.356, which implies that the lower the EEG signals amplitude of a password, the more the number of attempts to recall that password. The number of attempts to recall a password is related inversely to password memorability. In an ideal situation, a user will authenticate on their first attempt, whereas, more attempts to authenticate indicate the user did not remember the password.

Due to the small sample size used in calculating these correlations, the findings were not significant. For the EEG-Recall correlation finding of 0.375 to be significant, we needed $N = 19$, whereas we had $N = 15$ due to the removal of recall data for four participants who reported writing down their passwords. This motivated the revisiting of password recall from long-term memory over the one week period in the second study with a larger number of participants.

In the second experiment, password recall from long-term memory performed in the same session (within 30 minutes period) was successfully predicted with 81% accuracy and 0.86 AUC, as can be noted in Table 5.4. On the contrary to the high accuracy result of recall from long-term memory achieved on the same day of the experiment, predicting password recall from long-term memory over longer periods of time proved to be challenging. The classifiers performed poorly on days 2 and 8, where accuracy dropped to 31% and 22% respectively.

When comparing the accuracy of predicting password recall on Day 1 to that on Days 2 and 8, we notice the large drop in accuracy. This could be caused by the lack of averaging EEG signals when studying long-term password recall. Since each user was asked to remember one password only, and that password was different for each user (based on the seed word provided by them), that resulted in EEG data analysis being performed on single trials only. EEG data was collected one time from each participant upon initial presentation of their password, after which they were asked to recall that password on three different occasions. Due to the experiment design, it was

not possible to collect EEG data multiple times for the same password for the same participant (the EEG data being collected upon first presentation was a major component of our hypothesis for the long-term recall) or among different participants (due to customized passwords for different participants). Despite studies that argue single-trial analysis can be used and produce high accuracy in different scenarios [78, 64, 79], we acknowledge that it may have had a drawback effect on our results.

Password strength effect on recall

We studied brain signals of 19 participants while presented with passwords of different levels of strength. Half of the presented passwords were either too guessable (number of guesses needed to crack the password is less than 10^3) or very guessable (number of guesses is less than 10^6), and the other half was somewhat guessable (number of guesses is less than 10^8). Using only the EEG data collected from participants upon presenting the passwords, we were able to successfully classify password strength with an average accuracy of 84% per participant, and 89.3% when averaging data across participants.

The finding that EEG waves are different based on subsequent recognizability or recall of items is in-line with the literature [92, 44, 83, 6]. What this study adds is specific to passwords in two ways.

First, users are expected to memorize passwords. Hence, we asked the participants specifically to memorize the presented passwords, whereas prior work focused on recognition rather than recall. In this experiment, it was important to mimic the real world scenario of password use.

Second, we investigated the correlation between the password strength and its recall.

The results showed there was a statistically significant difference in password recall between the two lists. When testing the effect size, password strength had a strong influence on its recall ($d = 6.8$). Although, it has often been reported in the literature that users tend to choose weak passwords that are easy to remember (and hence easy to guess) [103, 17, 126, 116, 33, 125], there has been no study done to test if weak passwords are indeed easier to remember than stronger passwords. To the best of my knowledge, this study is the first to present empirical results that support existing speculation of users' behaviour.

The importance of directly linking password strength to its recall is two-fold. First, it establishes a base for what has been long known to be users' behaviour with no study that directly addresses the correlation between password strength and its recall. Data previously available linking the two is

based on observing leaked passwords' databases. Second, having an insight into how the brain reacts to weak versus stronger passwords, and what it can tell us about these passwords' recall, which may help provide insight for improving password security while acknowledging the human factor of the system and its instinctive behaviour [117, 94].

6.2 Password Perceived Memorability

Text passwords memorability as perceived by the users was investigated in the second study, where prediction of users' perception of password memorability based on the collected EEG signals was conducted.

Classification of EEG data with two labels, most memorable and least memorable, was performed using a 10-fold cross-validation SVM classifier. From Table 5.5, it is noted that the highest classification accuracy achieved was 82% with an AUC of 0.83. This result is considerably larger than random-guess, and suggests that there may be a difference in the elicited EEG signals for a password based on how the user perceives its memorability, despite all the passwords being the same length, and pulled from real passwords data leaks. This suggests that the users may be perceiving passwords memorability, hence usability, in an unaware manner, where they make decisions on password usability based on how they are mentally associating them.

The EEG dataset was used for a second classification task based on the passwords' estimated strength. In this task, no user perception data was used. The passwords were labeled based on their strength into two categories, weakest and strongest. Classification results were comparable to results based on the user perceived memorability. Classification accuracy of 77% with an AUC of 0.80 was obtained using password strength as shown in Table 5.10. In both classification tasks, performance was assessed by cross-validation and determined to fit the data to an acceptable degree based on several metrics (i.e., precision, recall, F-score, accuracy, and area under the ROC curve).

From Figures 5.4 and 5.8, and considering the performance metrics mentioned earlier, it was observed that classification based on perceived password memorability achieved higher accuracy and yielded in larger AUC (Accuracy = 82%, and AUC=0.83), compared to classification based on password strength (Accuracy = 77%, and AUC=0.80).

Next, we investigated the correlation between password strength and perceived password memorability. Using paired-samples t-test, a statistically significant difference in perceived memorability

of passwords depending on their strength was found, with $p < 0.005$, and a strong effect with an effect size of $d = 3.8$. However, this strong effect should be taken with caution. As can be noticed in the *strongest* passwords list of Table 4.9, some of these passwords –although considered very strong with a score of 4– they might be interpreted differently by different people based on their background. For example, passwords like *Blomsterbarn* and *Indraprastha* translate into flower child in Danish and an ancient city mentioned in Buddhist texts respectively. Although they were ranked as least memorable by the participants, due to their English background, these passwords may be perceived as very memorable for people familiar with them. In that case, the password strength will not be the only factor that influences the users perspective on usability.

These results of significant difference between *least memorable* and *most memorable*, as well as between *weakest* and *strongest* provide insight on how users perceive password memorability and password strength. It helps explain users' behaviour –commonly reported in the literature– of choosing weaker passwords over stronger ones since they perceive them as more usable.

6.3 Feature extraction, selection, and classification

Different methods of feature extraction were investigated, we generated three feature sets, in the time, frequency, and time-frequency domains. A fourth set was created by combining the three sets. For feature selection, the four feature sets were fed into two selection models: lasso regularization model and stepwise linear regression model. An SVM classifier was used to classify passwords based on both their memorability and strength. The four feature sets were fed to the classifier first without any feature selection methods applied, then with lasso regression, and finally with stepwise regression.

When comparing the performance of the different feature sets, it was observed through Tables 5.5 and 5.10 that the wavelet coefficients feature set yielded in larger areas than those of the power spectrum, statistics, and the fusion of these three methods whenever applicable.

When classification is based on the dataset of all participants, it can be observed from Figures 5.4 and 5.8 that –with the exception of wavelet coefficients feature set– using feature sets without any feature selection methods performed worse compared to using a feature selection method.

Comparison of stepwise vs. lasso performance is only limited to the wavelet coefficients feature set, and can be noticed that both selection methods took turns in performing better in different

situations. Stepwise outperformed lasso with an accuracy of 70% compared to lasso accuracy of 67% when classifying passwords based on strength. Whereas, lasso outperformed stepwise with an accuracy of 82% compared to an accuracy of 71% when classifying passwords based on perceived memorability.

6.4 Ecological Validity and Limitations

One challenge in password studies is that studying real-world passwords of the participants poses privacy and security concerns, so instead, we gather password usage data in a controlled laboratory environment. This results in a challenge in the quality of data gathered. Since passwords created or assigned during password studies do not carry importance for the participants. A password created for the purpose of a study does not guard any information of value to the user and is not needed to access information through the study period. Hence, users may not put much effort into memorizing it and sometimes choose to write it down instead. These are known challenges in studying password recall through user password studies. Komanduri et al. [68] described the ecological validity of studies as being difficult to demonstrate in any password study where participants know they are creating a password for a study, instead of creating a password for an account they value and expect to access repeatedly over time. Fahl et al. [40] compared user study passwords of 645 students to their real passwords created for their university's system, and found that 29.9% of participants did not behave as they normally do, while 46.1% percent offered comparable data and 24.0% offered somewhat comparable data, concluding that password studies create useful data. Although there are some participants who do not behave realistically during password studies, on the whole, they recommended that more research is needed to be done to find out how to best interpret the results. Despite Fahl et al. findings, a lab study with the use of BCIs carries more challenges, resulting in a small number of participants, in which case the 46% comparable data will not be enough to produce reliable results.

Due to these challenges, we chose to use perceived password memorability measure to study password features in the lab, in addition to password recall; despite password recall being the best choice to go to when studying password memorability. Another limitation of the study is not accounting for password entry errors such as incorrect capitalization, missing characters, or failing to press shift when typing a special character. Unsuccessful logins may not represent memory recall

failure but simply typographical errors [67]. Stanton and Greene [97] studied password recall and analyzed password entry errors, where they found that increasing the length of a password also increases the possibility of making typographical errors. However, it cannot be fully determined whether those errors were memory errors or motor execution errors (or a combination of both) [101]. There is no agreed upon method to distinguish password entry errors from memory errors, especially that the users in our experiment had two online sessions, where they could have used a smart phone platform rather than the desktop platform they used in the lab. This could have contributed to touchscreen errors.

The sample size was also a limitation of the *Password Recall* study recall analysis, 14 subjects finished the three sessions of the experiment, thus affecting statistical analyses. However, this only affected the long-term memorability study, not the classification of *common* vs. *random* or the short-term memorability study.

Some other limitations were related to the BCI headset itself. The fitting of wet sensors can sometimes be uncomfortable, making it not practical for user applications. Using wet-sensor headset in a study is feasible, but for everyday use, more robust headsets are needed. For example, a fewer number of sensors, the use of dry sensors, more flexible designs that allow for the headset to be quickly and easily fitted and used for longer periods of time, and the quality of collected data are areas where improvement could be sought. The Muse headband used dry sensors. However, the design of the headband was a challenge in some cases, the round nature of the headband prevented it from making proper contact for two forehead shapes.

Chapter 7

Conclusions and Future Work

7.1 Introduction

Through the experiments in this thesis, different password attributes were studied, using off-the-shelf brain-computer interfaces. Utilizing feature extraction and selection methods, the collected EEG data -upon presentation of different categories of passwords- were analyzed. Analyses of users' behaviour of password recall and anticipated memorability have been performed, and the effect the password strength has on recall and how users perceive and remember passwords was investigated.

In this chapter, we present a summary of the findings, our contribution along with the hypotheses tested and their results, then discuss areas of interest for future work.

7.2 Findings

Four classification tasks of EEG data collected by off-the-shelf BCIs were carried out over two experiments:

(1) *Recalled* vs. *Not recalled*, using labels provided by the participants upon password recall.

(2) *Most memorable* vs. *Least memorable*, using labels provided by the participants on perceived password memorability.

(3) *Weakest* vs. *Strongest*, using labels of estimated password strength, (weakest score = 0, and strongest score = 4).

(4) *Common* vs. *Random*, using labels of estimated password strength (common score = 0, and random score = 2).

The recall of passwords from short-term memory was studied by presenting users with a number of passwords for a period of 10 seconds each, and then asking the participants to recall or reproduce them. Using different feature extraction and selection methods, and a 10-fold cross-validation SVM algorithm, classification of data was done using binary labels of *Recalled=1* and *Not Recalled=0* collected from the participants' data. We were able to predict a password recall from short-term memory with an unknown label correctly, on average, 85% (AUC = 0.91) of the time.

Password recall from long-term memory was also studied. First, users were asked to remember one password over the study period. A positive correlation between password recall and the mean of EEG signals was found on the second day of the study, however, it weakened on the eighth day. A negative correlation was noted between the number of attempts a user needed to recall a password and that password's EEG signal mean. Due to the small sample size ($N = 15$ on Day 2, and $N = 10$ on Day 8), no significant results could be concluded. This led us to continue to investigate password recall from long-term memory with a larger sample population. Password recall for the second study period, similar to the first study (8–10 days), was investigated with $N = 75$. The users were asked to remember two different passwords for the period of the study. In the second study, we tested password recall by the end of the session on Day 1. Using the labels based on users recall, a classification accuracy of 81% (AUC = 0.86) was achieved. When testing the same hypothesis of predicting password recall based on the EEG data on Days 2 and 8, low accuracies were found. This led us to suspect that it may not be possible to use EEG data to predict password recall from long-term memory.

Users perception of password usability was studied. We collected EEG data from users while presenting them with different passwords, then we asked them to rank these passwords based on their perception of most to least memorable. Labels of ranking as provided by the users, *most memorable* and *least memorable*, were used to classify passwords achieving a classification accuracy of 82% (AUC = 0.83).

Next, we investigated password strength, running two classification tasks, one in each experiment. First, the classification of password strength based on the labels *common* and *random* was done. These two labels translate into password strength scores of 0 and 2, respectively. Classification based on password strength was achieved with an average accuracy of 84% (AUC = 0.93). We were

skeptical of the result as the two password lists are completely different in lexical terms, one being the worst passwords and commonly used words, and the other being all random characters and numbers. We decided to re-test the password classification based on password strength, this time with all passwords being real-world passwords drawn from the same datasets. Labels used in this classification task were *weakest* passwords vs. *strongest* passwords. These two labels translate into password strength scores of 0 and 4, respectively. The classification of $2,250$ passwords, with half labeled as weakest and half as strongest, achieved 77% accuracy (AUC = 0.80).

Finally, the effect of password strength on password recall and user perception was studied. Recall and user perception data collected from users during both experiments were studied in correlation to estimated password strength. No EEG data was used in this task. In the first experiment, we studied the effect of password strength on the recall from short-term memory. The test result showed a strong effect size of $d = 6.8$ on password recall. In other words, the stronger the password, the less chance there is to recall it, even momentarily, within short-term memory.

The effect of password strength on user perception of password memorability was studied in the second experiment. The test result showed a strong effect size of $d = 3.8$ on users perception of password memorability. The effect size of password strength is found to be larger on password recall than on perceived password memorability (6.8 vs. 3.8 effect size).

7.3 Contributions

This dissertation contributes to the usable security community by presenting much needed human performance data that are difficult to obtain in the real world. The main contribution of this work is to further the understanding of users' behaviour with regard to password memorability and security.

The findings seem to indicate that the users may be able to sense the password strength upon presentation and make decisions based on that. EEG signals were different between password categories based on password recall or even the perception of password memorability. Password strength had a strong effect on the recall of passwords and the perception of memorability, even with the users being unaware of it. This contributes to understanding users' behaviour such as choosing weak passwords as they perceive them as more memorable. Recalling passwords with higher strengths also proved to be more difficult than recalling weaker passwords, which supports the idea that asking users to create easy-to-remember and hard-to-guess passwords produces a cognitive load.

Below, we address the hypotheses tested in this thesis and summarize the findings.

(1) Is it possible to use machine learning and EEG data collected upon presentation of passwords to predict short-term memorability of passwords?

Hypothesis 1 (H1): It is possible to predict passwords' recall from **short-term memory** based on EEG data collected from participants upon presentation of the passwords.

We found evidence that suggests it may be possible to predict password recall from short-term memory, achieved accuracy = 85%

(2) Is it possible to use machine learning and EEG data collected upon presentation of passwords to predict long-term memorability for passwords?

Hypotheses 2 and 2A (H2 and H2A): It is possible to predict passwords' recall from **long-term memory** based on EEG data collected from participants upon presentation of the passwords.

We did not find evidence that suggests it may be possible to predict password recall from long-term memory, achieved accuracy on Day1= 81*%, Day2=*31*%, and Day8=*22*%.*

(3) Are stronger passwords always harder to recall?

Hypothesis (H3): There is a correlation between password strength and its recall.

We found evidence that suggests password strength affects its recall, negative correlation with an effect size of 6.8.

(4) Is it possible to predict a password's perceived memorability as judged by the users based on EEG signals elicited upon presenting that password?

Hypothesis (H4): It is possible to predict how a user **perceives a password memorability** based on EEG data collected using a BCI.

We found evidence that suggests it may be possible to predict perceived password memorability, achieved accuracy = 82%

(5) Does password strength affect the way users judge a password memorability? In other words, does the human brain recognize password strength and perceive stronger passwords as less memorable, hence less usable?

Hypothesis (H5): There is a correlation between password strength and its perceived memorability.

We found evidence that suggests password strength affects users perception of that password memorability, negative correlation with an effect size of 3.8.

7.4 Future Work

There are a number of areas throughout this thesis that were intentionally not explored due to their scope and scale but would be source of future research and interesting areas to be explored; some are discussed below. A further step that would make an interesting subject for future work is to build a password usability meter. Based on our findings, we believe there is the potential to develop a practical system to help users make better decisions when choosing passwords, through predicting the password memorability at the time of creation. Designing and testing an EEG-based password memorability meter, which has the potential to be integrated with password generators the same way password strength meters are displayed on many online sign-up forms. In principle, this app would require the user to fit a BCI, and perform calibration. After that, the user will be prompted to choose if they want to generate a password based on a seed-word they provide, or they prefer a password generated by the application. Based on the decision, a number of passwords will be shown to the user, while data is collected and the app is using that data for training. This will allow the classifier to choose a password that has a higher probability of being remembered by the user. This application will aim at helping the user choose a more usable password personalized for them, but provides no guarantee or replacement for user practicing the password and putting effort into memorizing it. Once the user decides on a password, they are given a rehearsal tool to practice it.

Another area, that may be of interest based on our work, is to replicate the experiments but assign the participants two passwords with different strengths. Next, study the participant's password long-term recall behaviour, and compare recall of the two passwords to explore the effect of password strength on long-term recall. In our experiment we found a strong effect of password strength on password recall from short-term memory, and it will be interesting to explore the correlation over long-term.

Another experiment of interest, is using EEG data to speculate a user password, for example, have participants create a password on the first session, and in the next session show them a number of passwords on the screen which contains their password, while monitoring their EEG as their

password appears on the screen. Next, explore if using the EEG data can inform us of when they saw their password.

7.5 Conclusion

Password memorability is a challenging subject to study, since the most established metric of password memorability is recall, and recall over long periods of time is not feasible in user studies, where passwords do not carry importance to participants and are not used frequently. It is hard to predict how users relate to or encode a password, and the strategies they use in remembering them. Some passwords may have a number or a combination of letters or numbers that are meaningful to one user but not to another.

In an effort to understand how users perceive passwords, we studied the brain waves of 19 participants using an Emotiv Epoc headband, while they were presented with a number of passwords, which they were asked to recall later. Results seem to indicate the possibility of predicting password recall from short-term memory based on the EEG data. Prediction of password recall from long-term memory was not possible over long periods of time.

A second empirical study was conducted, where we studied the brain waves of 75 participants using a Muse headband, while they were presented with a number of real passwords created by people and leaked to the Internet. Participants were then asked to rank how they think each password measured in terms of memorability. During the experiment explanation to the participants, they were explicitly asked not to consider if a password was secure enough, but only how memorable the password was for them.

Our findings suggest that there is a correlation between the EEG signals measured during presentation of passwords, and how users perceive the password memorability.

A correlation was also found between the EEG signals and the password strength ranking, which led us to explore the effect the password strength has on the users' perceived memorability. Results showed a strong effect of password strength on how users perceive its memorability.

The findings indicate that the users have a slightly higher capacity of estimating password memorability than solely based on its strength. The classification of EEG signals based on user ranking achieved higher results than the classification of the same signals based on password strength rank-

ing. This would indicate that users, while able to sense password strength, may be influenced by other factors that affect the way they perceive a password.

Appendices

Appendix A

Password Recall Study Classifier Performance

A.1 Recalled vs. Not Recalled

Results from Password Recall Study for classification of *Recalled* vs. *Not Recalled* passwords.

Table A.1: SVM Classifier performance of password recall from short-term memory - Power Spectrum FS.

Participant ID	Precision	Recall	*F*-Score	Accuracy	AUC
1	0.52	0.93	0.67	0.54	0.32
2	0.64	0.64	0.64	0.64	0.68
3	0.63	0.86	0.73	0.68	0.79
4	1.00	0.14	0.25	0.57	0.33
5	0.81	0.93	0.87	0.86	0.96
6	0.78	1.00	0.88	0.86	0.98
7	1.00	0.50	0.67	0.75	0.80
8	0.62	0.93	0.74	0.68	0.11
9	0.56	1.00	0.72	0.61	0.11
10	1.00	0.50	0.67	0.75	0.64
11	0.60	0.21	0.32	0.54	0.44
12	0.80	0.29	0.42	0.61	0.12
13	0.81	0.93	0.87	0.86	0.96
14	0.81	0.93	0.87	0.86	0.96
15	0.71	0.86	0.77	0.75	0.87
16	1.00	0.21	0.35	0.61	0.14
17	1.00	0.50	0.67	0.75	0.80
18	0.55	0.86	0.67	0.57	0.89
19	0.58	1.00	0.74	0.64	0.27
Average	0.76	0.70	0.66	0.69	0.59

Table A.2: SVM Classifier performance of password recall from short-term memory - Statistics FS.

Participant ID	Precision	Recall	*F*-Score	Accuracy	AUC
1	0.86	0.86	0.86	0.86	0.14
2	0.67	0.86	0.75	0.71	0.24
3	0.67	0.86	0.75	0.71	0.73
4	0.82	0.64	0.72	0.75	0.86
5	0.74	1.00	0.85	0.82	0.96
6	0.93	1.00	0.97	0.96	0.99
7	1.00	0.50	0.67	0.75	0.86
8	0.93	0.93	0.93	0.93	0.97
9	0.76	0.93	0.84	0.82	0.85
10	0.67	1.00	0.80	0.75	0.66
11	0.79	0.79	0.79	0.79	0.80
12	0.92	0.86	0.89	0.89	0.92
13	0.74	1.00	0.85	0.82	0.96
14	0.59	0.71	0.65	0.61	0.32
15	0.90	0.64	0.75	0.79	0.21
16	0.60	0.64	0.62	0.61	0.30
17	1.00	0.50	0.67	0.75	0.86
18	0.55	0.86	0.67	0.57	0.85
19	0.53	0.71	0.61	0.54	0.30
Average	0.77	0.80	0.77	0.76	0.67

Table A.3: SVM Classifier performance of password recall from short-term memory - Wavelet FS.

Participant ID	Precision	Recall	*F*-Score	Accuracy	AUC
1	0.93	0.93	0.93	0.93	0.92
2	0.86	0.86	0.86	0.86	0.87
3	1.00	0.57	0.73	0.79	0.88
4	0.80	0.86	0.83	0.82	0.94
5	0.88	1.00	0.93	0.93	0.87
6	1.00	1.00	1.00	1.00	1.00
7	1.00	0.50	0.67	0.75	0.85
8	0.93	1.00	0.97	0.96	0.96
9	0.87	0.93	0.90	0.89	0.92
10	1.00	0.50	0.67	0.75	0.96
11	0.69	0.79	0.73	0.71	0.85
12	0.90	0.64	0.75	0.79	0.93
13	0.88	1.00	0.93	0.93	0.87
14	0.82	1.00	0.90	0.89	0.94
15	0.88	1.00	0.93	0.93	0.98
16	0.93	0.93	0.93	0.93	0.97
17	1.00	0.50	0.67	0.75	0.85
18	0.93	0.93	0.93	0.93	0.92
19	0.62	0.93	0.74	0.68	0.23
Average	0.89	0.83	0.84	0.85	0.88

Table A.4: SVM Classifier performance of password recall from short-term memory - Combined FS.

Participant ID	Precision	Recall	*F*-Score	Accuracy	AUC
2	0.54	0.50	0.52	0.54	0.36
3	0.65	0.93	0.76	0.71	0.80
4	1.00	0.14	0.25	0.57	0.89
5	0.82	1.00	0.90	0.89	0.98
6	0.78	1.00	0.88	0.86	1.00
7	1.00	0.50	0.67	0.75	0.80
8	0.74	1.00	0.85	0.82	0.96
9	0.56	1.00	0.72	0.61	0.10
10	1.00	0.50	0.67	0.75	0.64
11	0.91	0.71	0.80	0.82	0.10
12	1.00	0.64	0.78	0.82	0.93
13	0.82	1.00	0.90	0.89	0.98
14	0.82	1.00	0.90	0.89	0.94
15	0.87	0.93	0.90	0.89	0.97
16	1.00	0.43	0.60	0.71	0.08
17	1.00	0.50	0.67	0.75	0.80
18	0.55	0.86	0.67	0.57	0.89
19	0.61	1.00	0.76	0.68	0.19
Average	0.80	0.77	0.73	0.74	0.67

Table A.5: SVM Classifier performance of password recall from short-term memory - Power Spectrum FS - Lasso.

Participant ID	**Precision**	**Recall**	***F*-Score**	**Accuracy**	**AUC**
2	0.63	0.86	0.73	0.68	0.66
3	0.63	0.86	0.73	0.68	0.79
5	1.00	1.00	1.00	1.00	1.00
6	0.76	0.93	0.84	0.82	0.94
7	1.00	0.50	0.67	0.75	0.78
9	0.72	0.93	0.81	0.79	0.85
10	1.00	0.50	0.67	0.75	0.84
11	0.75	0.43	0.55	0.64	0.48
12	0.60	0.86	0.71	0.64	0.69
13	0.83	0.36	0.50	0.64	0.72
14	0.75	0.43	0.55	0.64	0.68
15	1.00	0.14	0.25	0.57	0.53
16	1.00	0.79	0.88	0.89	0.95
Average	0.82	0.66	0.68	0.73	0.76

Table A.6: SVM Classifier performance of password recall from short-term memory - Statistics FS - Lasso.

Participant ID	**Precision**	**Recall**	***F*-Score**	**Accuracy**	**AUC**
1	0.78	1.00	0.88	0.86	0.81
2	0.69	0.79	0.73	0.71	0.17
3	0.63	0.86	0.73	0.68	0.71
4	0.80	0.57	0.67	0.71	0.79
5	0.74	1.00	0.85	0.82	0.97
6	0.93	1.00	0.97	0.96	0.98
7	0.83	0.71	0.77	0.79	0.82
8	0.87	0.93	0.90	0.89	0.96
9	0.76	0.93	0.84	0.82	0.85
10	1.00	0.50	0.67	0.75	0.98
14	0.85	0.79	0.81	0.82	0.78
15	0.56	0.71	0.63	0.57	0.68
16	0.55	0.86	0.67	0.57	0.84
Average	0.77	0.82	0.78	0.77	0.80

Table A.7: SVM Classifier performance of password recall from short-term memory - Wavelet FS - Lasso.

Participant ID	Precision	Recall	*F*-Score	Accuracy	AUC
1	1.00	1.00	1.00	1.00	1.00
2	0.93	1.00	0.97	0.96	1.00
3	0.79	0.79	0.79	0.79	0.89
4	1.00	0.93	0.96	0.96	0.98
5	1.00	1.00	1.00	1.00	1.00
6	1.00	1.00	1.00	1.00	1.00
7	0.86	0.86	0.86	0.86	0.93
8	0.76	0.93	0.84	0.82	0.95
9	0.92	0.86	0.89	0.89	0.89
10	1.00	0.50	0.67	0.75	0.96
11	0.88	1.00	0.93	0.93	0.99
12	0.90	0.64	0.75	0.79	0.94
13	0.87	0.93	0.90	0.89	0.97
14	0.93	1.00	0.97	0.96	0.98
15	0.88	1.00	0.93	0.93	0.96
17	0.58	0.50	0.54	0.57	0.69
Average	0.89	0.87	0.87	0.88	0.95

Table A.8: SVM Classifier performance of password recall from short-term memory - Combined FS - Lasso

Participant ID	Precision	Recall	*F*-Score	Accuracy	AUC
1	0.78	1.00	0.88	0.86	0.83
2	0.78	1.00	0.88	0.86	0.00
3	0.75	0.86	0.80	0.79	0.84
4	1.00	0.93	0.96	0.96	0.98
5	0.88	1.00	0.93	0.93	0.96
6	1.00	1.00	1.00	1.00	1.00
7	0.87	0.93	0.90	0.89	0.91
8	0.93	0.93	0.93	0.93	0.95
9	1.00	0.86	0.92	0.93	0.97
10	1.00	0.50	0.67	0.75	0.96
11	0.88	1.00	0.93	0.93	0.99
12	1.00	1.00	1.00	1.00	1.00
13	0.87	0.93	0.90	0.89	0.97
14	0.92	0.79	0.85	0.86	0.91
15	0.93	1.00	0.97	0.96	1.00
17	0.58	0.50	0.54	0.57	0.69
Average	0.88	0.89	0.88	0.88	0.87

Table A.9: SVM Classifier performance of password recall from short-term memory- Wavelet FS - Stepwise

Participant ID	**Precision**	**Recall**	***F*-Score**	**Accuracy**	**AUC**
1	1.00	0.78	0.88	0.89	1.00
2	0.93	1.00	0.96	0.96	0.99
3	0.75	0.85	0.8	0.78	0.84
4	1.00	0.92	0.96	0.96	0.98
5	0.73	0.78	0.75	0.75	0.85
6	0.93	1.00	0.96	0.96	0.97
7	0.86	0.92	0.89	0.89	0.94
8	0.78	0.78	0.78	0.78	0.91
9	0.93	1.00	0.96	0.96	1.00
10	0.91	0.78	0.84	0.85	0.95
11	0.92	0.85	0.88	0.89	0.98
12	0.85	0.85	0.85	0.85	0.96
13	0.92	0.92	0.92	0.92	0.97
14	0.93	1.00	0.96	0.96	0.99
15	0.81	0.92	0.86	0.85	0.95
16	0.65	0.92	0.76	0.71	0.77
17	0.58	0.50	0.53	0.57	0.69
18	0.73	0.78	0.75	0.75	0.85
19	0.65	0.92	0.76	0.71	0.77
Average	0.83	0.86	0.84	0.84	0.91

A.2 Common vs. Random

Results from Password Recall Study for classification of *common* vs. *random* passwords.

Table A.10: SVM Classifier performance of *common* vs. *random* passwords - Power Spectrum FS.

Participant ID	**Precision**	**Recall**	***F*-Score**	**Accuracy**	**AUC**
1	0.61	1.00	0.76	0.68	0.88
2	1.00	0.21	0.35	0.61	0.29
3	1.00	0.50	0.67	0.75	0.83
4	0.80	0.29	0.42	0.61	0.83
5	0.63	0.36	0.45	0.57	0.32
6	1.00	0.79	0.88	0.89	0.95
7	0.80	0.57	0.67	0.71	0.85
8	0.56	1.00	0.72	0.61	0.39
9	0.86	0.43	0.57	0.68	0.61
10	0.82	1.00	0.90	0.89	0.98
11	0.83	0.36	0.50	0.64	0.68
12	1.00	0.29	0.44	0.64	0.77
13	1.00	0.43	0.60	0.71	0.77
14	0.79	0.79	0.79	0.79	0.87
15	0.80	0.57	0.67	0.71	0.81
16	0.61	1.00	0.76	0.68	0.80
17	1.00	0.29	0.44	0.64	0.79
18	0.63	0.36	0.45	0.57	0.68
19	0.80	0.57	0.67	0.71	0.85
Average	0.82	0.57	0.62	0.69	0.73

Table A.11: SVM Classifier performance of *common* vs. *random* passwords - Statistics FS.

Participant ID	**Precision**	**Recall**	***F*-Score**	**Accuracy**	**AUC**
1	0.91	0.71	0.80	0.82	0.90
2	1.00	0.57	0.73	0.79	0.87
3	0.65	0.93	0.76	0.71	0.83
4	1.00	0.79	0.88	0.89	0.89
5	0.64	0.50	0.56	0.61	0.26
6	1.00	0.71	0.83	0.86	0.87
7	0.81	0.93	0.87	0.86	0.88
8	0.77	0.71	0.74	0.75	0.71
9	0.60	0.43	0.50	0.57	0.72
10	0.58	0.79	0.67	0.61	0.72
11	0.92	0.86	0.89	0.89	0.94
12	0.67	0.71	0.69	0.68	0.84
13	0.89	0.57	0.70	0.75	0.76
14	0.60	0.64	0.62	0.61	0.75
15	0.83	0.71	0.77	0.79	0.14
16	0.71	0.71	0.71	0.71	0.82
17	0.69	0.64	0.67	0.68	0.27
18	0.64	0.50	0.56	0.61	0.26
19	0.81	0.93	0.87	0.86	0.88
Average	0.77	0.70	0.73	0.74	0.70

Table A.12: SVM Classifier performance of *common* vs. *random* passwords - Wavelet FS.

Participant ID	**Precision**	**Recall**	***F*-Score**	**Accuracy**	**AUC**
1	1.00	1.00	1.00	1.00	1.00
2	1.00	0.57	0.73	0.79	0.94
3	1.00	0.50	0.67	0.75	0.77
4	1.00	0.93	0.96	0.96	1.00
5	0.76	0.93	0.84	0.82	0.95
6	0.93	1.00	0.97	0.96	1.00
7	0.91	0.71	0.80	0.82	0.89
8	1.00	0.93	0.96	0.96	0.97
9	0.91	0.71	0.80	0.82	0.94
10	0.93	1.00	0.97	0.96	0.99
11	0.82	1.00	0.90	0.89	0.94
12	0.88	0.50	0.64	0.71	0.94
13	0.79	0.79	0.79	0.79	0.88
14	0.91	0.71	0.80	0.82	0.89
15	0.76	0.93	0.84	0.82	0.95
16	0.75	0.64	0.69	0.71	0.16
17	0.78	0.50	0.61	0.68	0.18
18	0.87	0.93	0.90	0.89	0.97
19	0.81	0.93	0.87	0.86	0.95
Average	0.88	0.80	0.83	0.84	0.86

Table A.13: SVM Classifier performance of *common* vs. *random* passwords - Combined FS.

Participant ID	**Precision**	**Recall**	***F*-Score**	**Accuracy**	**AUC**
1	0.64	1.00	0.78	0.71	0.97
2	0.76	0.93	0.84	0.82	0.08
3	1.00	0.50	0.67	0.75	0.85
4	1.00	0.64	0.78	0.82	0.99
5	0.75	0.43	0.55	0.64	0.76
6	0.88	1.00	0.93	0.93	0.99
7	1.00	0.57	0.73	0.79	0.90
8	1.00	0.64	0.78	0.82	0.05
9	0.82	0.64	0.72	0.75	0.70
10	0.82	1.00	0.90	0.89	0.99
11	0.83	0.71	0.77	0.79	0.93
12	1.00	0.21	0.35	0.61	0.93
13	1.00	0.57	0.73	0.79	0.91
14	0.75	0.43	0.55	0.64	0.24
15	1.00	0.57	0.73	0.79	0.90
16	0.64	1.00	0.78	0.71	0.83
17	1.00	0.29	0.44	0.64	0.84
18	0.86	0.43	0.57	0.68	0.85
19	0.82	1.00	0.90	0.89	0.98
Average	0.87	0.66	0.71	0.76	0.77

Table A.14: SVM Classifier performance of *common* vs. *random* passwords - Power Spectrum FS - Lasso.

Participant ID	**Precision**	**Recall**	***F*-Score**	**Accuracy**	**AUC**
1	0.63	0.86	0.73	0.68	0.83
2	1.00	0.29	0.44	0.64	0.61
3	0.80	0.57	0.67	0.71	0.85
4	0.83	0.36	0.50	0.64	0.80
5	0.73	0.57	0.64	0.68	0.67
6	0.90	0.64	0.75	0.79	0.82
7	0.80	0.57	0.67	0.71	0.88
8	0.89	0.57	0.70	0.75	0.82
9	1.00	0.64	0.78	0.82	0.88
10	1.00	1.00	1.00	1.00	1.00
11	0.86	0.43	0.57	0.68	0.77
12	0.86	0.43	0.57	0.68	0.68
13	1.00	0.50	0.67	0.75	1.00
14	0.86	0.43	0.57	0.68	0.84
15	1.00	0.43	0.60	0.71	0.78
16	0.59	0.93	0.72	0.64	0.76
17	1.00	0.29	0.44	0.64	0.62
Average	0.87	0.56	0.65	0.72	0.80

Table A.15: SVM Classifier performance of *common* vs. *random* passwords - Statistics FS - Lasso.

Participant ID	**Precision**	**Recall**	***F*-Score**	**Accuracy**	**AUC**
1	0.90	0.64	0.75	0.79	0.87
2	1.00	0.71	0.83	0.86	0.89
3	0.69	0.79	0.73	0.71	0.80
4	1.00	0.86	0.92	0.93	0.91
5	0.65	0.79	0.71	0.68	0.66
6	0.91	0.71	0.80	0.82	0.85
7	0.87	0.93	0.90	0.89	0.89
10	0.58	0.79	0.67	0.61	0.28
11	0.86	0.86	0.86	0.86	0.94
12	0.71	0.71	0.71	0.71	0.85
13	0.88	0.50	0.64	0.71	0.77
14	0.60	0.64	0.62	0.61	0.25
15	0.83	0.71	0.77	0.79	0.86
16	0.71	0.71	0.71	0.71	0.82
17	0.71	0.86	0.77	0.75	0.74
Average	0.79	0.75	0.76	0.76	0.76

Table A.16: SVM Classifier performance of *common* vs. *random* passwords - Wavelet FS - Lasso.

Participant ID	**Precision**	**Recall**	***F*-Score**	**Accuracy**	**AUC**
1	1.00	1.00	1.00	1.00	1.00
3	0.85	0.79	0.81	0.82	0.86
4	1.00	1.00	1.00	1.00	1.00
5	0.81	0.93	0.87	0.86	0.88
6	1.00	1.00	1.00	1.00	1.00
7	0.92	0.86	0.89	0.89	0.95
8	1.00	0.93	0.96	0.96	0.99
9	0.93	0.93	0.93	0.93	0.97
10	0.93	0.93	0.93	0.93	0.99
11	0.78	1.00	0.88	0.86	1.00
12	0.80	0.86	0.83	0.82	0.88
13	0.92	0.79	0.85	0.86	0.88
14	1.00	1.00	1.00	1.00	1.00
15	1.00	0.86	0.92	0.93	0.98
17	0.75	0.86	0.80	0.79	0.85
Average	0.91	0.91	0.91	0.91	0.95

Table A.17: SVM Classifier performance of *common* vs. *random* passwords - Combined FS - Lasso

Participant ID	**Precision**	**Recall**	***F*-Score**	**Accuracy**	**AUC**
1	1.00	1.00	1.00	1.00	1.00
2	0.92	0.79	0.85	0.86	0.98
4	0.93	0.93	0.93	0.93	0.97
5	0.80	0.57	0.67	0.71	0.78
6	1.00	1.00	1.00	1.00	1.00
7	1.00	0.86	0.92	0.93	0.94
8	0.93	1.00	0.97	0.96	1.00
9	0.93	0.93	0.93	0.93	0.97
10	0.87	0.93	0.90	0.89	0.97
11	0.93	1.00	0.97	0.96	1.00
12	0.91	0.71	0.80	0.82	0.90
13	0.92	0.79	0.85	0.86	0.88
14	0.93	0.93	0.93	0.93	0.99
15	1.00	0.86	0.92	0.93	0.98
Average	0.93	0.88	0.90	0.91	0.95

Table A.18: SVM Classifier performance of *common* vs. *random* passwords - Wavelet FS - Stepwise

Participant ID	**Precision**	**Recall**	***F*-Score**	**Accuracy**	**AUC**
1	1.00	1.00	1.00	1.00	1.00
2	0.61	0.79	0.69	0.64	0.76
3	0.89	0.57	0.70	0.75	0.86
4	1.00	0.93	0.96	0.96	0.99
5	0.61	0.79	0.69	0.64	0.76
6	1.00	1.00	1.00	1.00	1.00
7	0.86	0.86	0.86	0.85	0.94
8	0.93	0.93	0.93	0.92	0.99
9	1.00	1.00	1.00	1.00	1.00
10	0.93	0.93	0.93	0.92	0.98
11	1.00	0.93	0.96	0.96	0.99
12	0.70	0.50	0.58	0.64	0.69
13	0.87	0.93	0.90	0.89	0.97
14	0.93	1.00	0.97	0.96	1.00
15	0.88	1.00	0.93	0.92	0.97
16	0.62	0.93	0.74	0.67	0.74
17	0.89	0.57	0.70	0.75	0.87
18	0.61	0.79	0.69	0.64	0.76
19	0.86	0.86	0.86	0.85	0.94
Average	0.85	0.86	0.85	0.84	0.91

Appendix B

Eligibility Questionnaire

- Are you 18 years of old or older as of today?
- Do you have normal or corrected to normal vision? (i.e. normal meaning you are able to see and read clearly, or corrected to normal meaning you are able to see and read clearly using glasses or contact lens)
- Do you have any condition or are taking any substance that may hinder your ability to read (e.g., drowsy prescription drugs)?
- Are you willing to remove any headgear to allow placement of sensors?
- 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))

Appendix C

Consent Form

C.1 Recall Study

Research Personnel

Ruba Alomari	Miguel Vargas Martin	Chris Bellman
Graduate Student	Faculty	Graduate Student
Ruba.Alomari@uoit.ca	Miguel.VargasMartin@uoit.ca	Christopher.Bellman@uoit.net

Ramiro Liscano	Shane Macdonald
Faculty	Student
Ramiro.Liscano@uoit.ca	Shane.Macdoanld@uoit.net

Purpose: The purpose of this study is to conduct research on BCI devices (devices that take signals produced by the human brain and convert them into data that a computer can read and interpret) and how the brain interacts with memorability-based tasks.

Task Requirements: This study is composed of three sessions in the lab, in the first session you will be asked to fill a short pre-questionnaire, and to provide a word that we can use later to create a password for you to use during this study period, you will also be asked to look at passwords on a screen for a period of time, and try to remember them, after each password you will be asked to type it after a short period of time, this task will be repeated for a number of passwords, at the end

of Session 1 you will be assigned a password that you will be asked to remember and re-enter in two different sessions over the course of 8-10 days. You can practice typing the assigned password as long as you wish to in order to aid in remembering it.

In Session 2 you will be asked to come to the lab to enter the study password within 1-2 days after Session 1, and in Session 3 you will be asked to return to the lab to enter the study password within 8-10 days after Session 1. At the end of Session 3 you will complete a short questionnaire.

As a token of appreciation, you will be compensated for your time helping in this research $5 after the completion of Session 1, $2 after the completion of Session 2 and $3 after the completion of Session 3, you can choose to collect the compensation after each session or choose to collect your compensation of the complete amount of the $10 after session 3 is completed.

Duration and Location: All sessions will be conducted in the lab at a time decided and agreed upon based on the participants availability, the first session is expected to last 30 minutes, the second and third sessions will last 5-10 minutes each. The total expected duration of the experiment should be approximately 45 minutes.

Potential Risks/Discomfort: Due to the design of the BCI headset used for the experiment, there is slight discomfort after wearing the device for an extended period of time, if you begin to feel discomfort, a short break will be given to remove the headset and alleviate the discomfort. It should be noted that you may withdraw at any time, should you experience an allergic reaction, you will be brought to the Campus Wellness Center in case of minor reactions, major reactions will warrant the activation of emergency services contacted through campus security. The experiment also uses a small amount of contact lens solution to moisten the sensors on the headset. Participants who are allergic to general multipurpose contact lens solution may experience a reaction, the solution used in the experiment list the following ingredients in the package: hydranate (hydroxyalkylphosphonate), boric acid, edetate disodium, poloxamine, sodium borate, sodium chloride, dymed (polyaminopropyl biguanide).

Anonymity/Confidentiality: All collected data will be held completely confidential. The data will only be made available to the researchers who are involved in the study. Data will be coded for identification purposes, which means the data will be associated with an arbitrary identifier (such

as "6532") which cannot be linked to a specific person. The record that will map the participant to the randomly coded ID will be kept securely by the primary investigator throughout the study and will be permanently destroyed upon completion of the study.

Right to Withdraw: You have the right to withdraw at any time during this study, before or during any of the sessions. To request that none of your data is used, you may tell us during the lab session or email us at any time using the email address specified below, without any explanation as to the reason for withdrawing from the testing. Please note that your data cannot be destroyed after completion of Session 3 as it will have been anonymized.

Secondary Use of Data: Please note, if you agree to participate (and do not withdraw from the study), your anonymized data (no identifiers linking the information to you) may also be used (for future studies relating to BCI and password attributes systems, etc.).

Participant Concerns and Reporting: If you have any questions concerning the research study or experience any discomfort related to the study, please contact the researcher at Ruba.Alomari@uoit.ca

Any questions regarding your rights as a participant, complaints or adverse events may be addressed to Research Ethics Board through the Ethics and Compliance Officer – researchethics@uoit.ca or 905.721.8668 x. 3693. This study has been approved by the UOIT Research Ethics Board REB #14057 on Sep 13th, 2016.

I have read and understand the above terms of testing and I understand the conditions of my participation, by consenting to participate I do not waive any legal rights or recourse.

Name (printed) **Signature** **Date**

C.2 Password Perceived Memorability

Research Personnel

Ruba Alomari	Miguel Vargas Martin	Chris Bellman
Graduate Student	Faculty	Graduate Student
Ruba.Alomari@uoit.ca	Miguel.VargasMartin@uoit.ca	Christopher.Bellman@uoit.net

Ramiro Liscano	Shane Macdonald
Faculty	Student
Ramiro.Liscano@uoit.ca	Shane.Macdoanld@uoit.net

Purpose: The purpose of this study is to conduct research on BCI devices (devices that take signals produced by the human brain and convert them into data that a computer can read and interpret) and how the brain interacts with memorability-based tasks.

Task Requirements: This study is composed of three sessions, one in the lab and two online. In the first session you will be asked to fill a short pre-questionnaire, and to provide two words that will be used to create two passwords. You will use the two created passwords during this study period. Next you will be asked to study a number of passwords appearing in front of you on a screen. You will be asked to rank these passwords based on how memorable you think that password is (1 being the least memorable password and 5 the most memorable). Once you finish studying the passwords you will be asked to login to the study website by entering the two passwords that were created at the beginning of the session. You will be asked to remember and re-enter these two passwords in two different sessions over the course of 8-10 days.

Sessions 2 and 3 are online sessions. You will be emailed a link to the study website, where you will login and update some information as guided by the website. Sessions 2 is within 24-48 hours after Session 1, Session 3 is within 8-10 days after Session 1. At the end of session 3 you will complete a short questionnaire and by that the study is completed. As a token of appreciation, you will be compensated for your time helping in this research $5 after the completion of Session 1, $5 after the completion of Session 3, you can choose to collect the compensation after each session or choose to collect your compensation of the complete amount of the $10 after session 3 is completed.

Duration and Location: Session 1 will be conducted in the lab at a time decided and agreed upon based on the participants availability, the first session is expected to last 30 minutes, the second and third sessions will last 5 minutes each. The total expected duration of the experiment should be approximately 40 minutes.

Potential Risks/Discomfort: Due to the design of the BCI headset used for the experiment, there is slight discomfort after wearing the device for an extended period of time, if you begin to feel discomfort, a short break will be given to remove the headset and alleviate the discomfort. It should be noted that you may withdraw at any time, should you experience an allergic reaction, you will be brought to the Campus Wellness Center in case of minor reactions, major reactions will warrant the activation of emergency services contacted through campus security. The experiment also uses a small amount of contact lens solution to moisten the sensors on the headset. Participants who are allergic to general multipurpose contact lens solution may experience a reaction, the solution used in the experiment list the following ingredients in the package: hydranate (hydroxyalkylphosphonate), boric acid, edetate disodium, poloxamine, sodium borate, sodium chloride, dymed (polyaminopropyl biguanide).

Anonymity/Confidentiality: All collected data will be held completely confidential. The data will only be made available to the researchers who are involved in the study. Data will be coded for identification purposes, which means the data will be associated with an arbitrary identifier (such as "6532") which cannot be linked to a specific person. The record that will map the participant to the randomly coded ID will be kept securely by the primary investigator throughout the study and will be permanently destroyed upon completion of the study.

Right to Withdraw: You have the right to withdraw at any time during this study, before or during any of the sessions. To request that none of your data is used, you may tell us during the lab session or email us at any time using the email address specified below, without any explanation as to the reason for withdrawing from the testing. Please note that your data cannot be destroyed after completion of Session 3 as it will have been anonymized.

Secondary Use of Data: Please note, if you agree to participate (and do not withdraw from the study), your anonymized data (no identifiers linking the information to you) may also be used (for future studies relating to BCI and password attributes systems, etc.).

Participant Concerns and Reporting: If you have any questions concerning the research study or experience any discomfort related to the study, please contact the researcher Ruba Alomari at (647)390-1056 or Ruba.Alomari@uoit.ca

Any questions regarding your rights as a participant, complaints, or adverse events may be addressed to Research Ethics Board through the Research Ethics Coordinator – researchethics@uoit.ca or 905.721.8668 x. 3693.

This study has been approved by the UOIT Research Ethics Board REB #14407 on 06/06/2017.

I have read and understand the above terms of testing and I understand the conditions of my participation, by consenting to participate I do not waive any legal rights or recourse.

Name (printed) **Signature** **Date**

Appendix D

Post-Study Survey Responses

"They were all very memorable even with symbols involved (although the symbols made it a little harder to remember). Remembering the number 0 not being a Uppercase "O" was hard though."

"Repetition of phrase or one-worded passwords are easier to remember."

"The original passwords and the relationship between them and the transformed ones."

"I think using meaningful words in your password makes it more memorable."

"What makes a password memorable is using it. The more one gets used to a code, the easier it is to remember it. It doesn't have to be memorable or easy, with enough practice and use, any password can become memorable."

"The password is memorable because it is personal to me and I practiced it until it felt like I got the muscle memory of it during the first session. It is less memorable because of the different case and character substitutions."

"Practice repeating the passwords in the first session was very helpful. It allowed for the visual picture to present itself when I was required to enter the passwords again. I also chose words that had meaning, or that I come across everyday, which helped me to remember them."

"What makes a password more memorable is if it's a word or something that symbolizes what you like. For me I just chose my favourite colour and fantasy animal."

"Oddly enough, I think the fact that my second password had a (in it made it more memorable. Possibly because it was more novel?"

"Repetition or common words make a password more memorable."

"Consistent patterns make remembering passwords easier."

"Less memorable if there is astrik, numbers, upper case letter etc.."

"If some memory/object which I frequently use/do/read can be directly connected in my brain using a word or combination of words, it gets stored permanently in my memory."

"Words of everyday meaning to me (what I study, names, food, places I visit, dates I know) are memorable passwords."

"Chosen familiar Words and memorized the spaces where the special characters are used during Password Construction."

"To me, a password is more memorable if it's a pattern of words that I can recognize, regardless of length."

"In my opinion I think that the most memorable passwords are those including names, doubled words. And the less memorable are the ones including mixture of numbers and letters. For me, I feel it is more hard to remember numbers than it is to remembering names."

"A password is more memorable if it is modified from english like my passwords were "H" to "#". It is also more memorable if it is a repeated series of numbers/letters or common repeated series of numbers/letters like "123" or "qwerty"."

"What made the passwords most memorable for me was that the 3 changed characters were not spread out."

"I believe that a password is more memorable in my opinion, by special characteristics or the uncomforting formatting of the passwords."

"I think what made my password more memorable was that I used names that I was well associated with. For instance, *password removed by researcher* is the name I gave to my first desktop computer. *Password removed by researcher* is the name of my fiancée. However, I understand that using just a name without any tricks to it would probably be the least secure password ever. Adding symbols or random CAPS would be helpful in assuring a safe and secure account password."

"Choosing a word that has some relevance in your life or is just really odd helps create a good password."

"I always find a password to be more memorable based on a few factors:

1. Relatability to myself. 2. Keeping it related to the website. 3. Not using special characters or numbers but if I do, I keep it similar or standard with most of my logins for each site. For example if I used $ in one, I'd use it more often. 4. Capitalizing letters within a word is tricky. 5. Gibberish doesn't help with remembering. 6. No matter what, if I haven't used a login for a few weeks, I will most likely press forgot password. 7. The practice within session 1 was helpful for remembering the password."

"Making a password from something you use often, do often, or say often can help you remember a password. That being said, passwords created and remembered but never used can be forgotten quite easily."

"I did not write down my password at any point but I did go over it in my head about once or twice a day in detail. By detail I mean I would tell myself where the capitals and symbols were in the password and thought of visuals that I was reminded of when repeating the passwords."

"A password that has meaning, significant or memorable letters are different, symbols resemble the letter that they replace."

"I think the more relatable a password is to a word the more it sticks."

"In my personal opinion, I think what makes a password memorable is the connection between whatever the password is, and the person. For example, the passwords for this study I had to remember was *passwords removed by researcher*."

"If it's a word with no extra numbers or random letters it is the most memorable."

"These password [sic] were very annoying to memorize I suggest just using longer passwords with multiple easy to remember words like instead of wINd0w and TurtL3 turtlewindow is a lot easier to remember and probably harder to guess."

"I think that interesting but comfortable keystrokes made for more memorable passwords. Personally, I also make passwords by making 'codexes' to convert the base idea for the password into something else completely (e.g., different languages + 1337 + binary)."

"Passwords with similar, repetitive symbols are harder to remember (e.g. 0opdb0o vs. 98ewr98). As well, passwords that have distinct patterns can be easily remembered through muscle memory (e.g. 123qweasdzxc or qwerty123). More variation in a password makes it less memorable (e.g. b@nAn4$ vs. banana)."

"The multiple symbols were difficult to remember."

"What makes a password memorable is whether or not it has any meaning to you outside of the login screen or if it has any correlation to what you are signing into provided it still being a secure string of text.

For example my passwords were the words "school" and "studies" and why I chose those is because I was taking part in a study at my school. This way I didn't actually have to remember my passwords because the act of participating in the study was all the remembering required. I also think they were good passwords because your algorithm to make them more secure took the hard part out of my hands and I had no reasons to believe someone would guess those words anyways."

"I think what made me remember the passwords was typing them out repeatedly and occasionally getting them wrong, so I learned from my mistakes."

"The more special characters used to substitute ordinary letters, the harder sometimes it becomes to remember unless all possible characters have been substituted by special characters. For instance, hotrod is easier to remember as h0+r0d than ho+r0d. One can use a password word [sic] that contains only one letter that can be substituted as a special character like cat $\rightarrow$ c@t but its difficult to remember a password in which some but not all possible letters have been substituted by special characters (redundant point). Thank you for this opportunity to help research @ UOIT"

"What makes password more memorable for me if its a word or a phrase [sic]. I find it easier to remember those kinds of passwords even if they have numbers and capital letters in them."

"Changing a letter to a similar number (like letter 'O' to number 'zero') is much memorable than to a symbol."

"I think it's good to have a password that you can relate back to something. A phrase is more memorable than a string of numbers."

"I can remember the password if its a word that is related to my day to day activity or a word that i use very often or sometimes names of my family members makes it easy for me to remember the password."

"A password is more memorable when the words are repeated. For instance, hellohello, would be an example of a memorable password."

"The passwords that were generated were pretty easy to memorize, I memorized them by splitting them into parts. The first password was *password removed by researcher*. I guess what makes a password more memorable is simplicity. Common words "superhero123" or "lighthouse3030". However, passwords are also more memorable depending on the person's personal life. What they're familiar with. Latin-sounding words might be more memorable than even common words. So there is a variety."

"I remembered my passwords somewhat like an image, so recalling it the first time was a bit difficult. However took a piece of paper to try spelling out my password, and it became much easier to recall the passwords. This was for session 2. Session 3 was much easier for me as, at this point I have already memorized my passwords."

"The more relatable/personal the password is, the easier it is to remember."

Bibliography

[1] Secure Password Generator, 2017. http://passwordsgenerator.net.

[2] Abujelala, M., Abellanoza, C., Sharma, A., and Makedon, F. Brain-ee: Brain enjoyment evaluation using commercial EEG headband. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments* (New York, NY, USA, 2016), PETRA '16, ACM, pp. 33:1–33:5.

[3] Adams, A., and Sasse, M. A. Users are not the enemy. *Commun. ACM 42*, 12 (Dec. 1999), 40–46.

[4] Al-Fahoum, A. S., and Al-Fraihat, A. A. Methods of EEG signal features extraction using linear analysis in frequency and time-frequency domains. *ISRN Neuroscience 2014* (2014), 267–288.

[5] Allan, A. Passwords are near the breaking point. In *Gartner Research* (2004).

[6] Alomari, R., Martin, M. V., MacDonald, S., Bellman, C., Liscano, R., and Maraj, A. What your brain says about your password: Using brain-computer interfaces to predict password memorability. In *2017 15th Annual Conference on Privacy, Security and Trust (PST)* (Aug 2017).

[7] Alotaiby, T., El-Samie, F. E. A., Alshebeili, S. A., and Ahmad, I. A review of channel selection algorithms for EEG signal processing. *EURASIP Journal on Advances in Signal Processing 2015*, 1 (Aug 2015), 66.

[8] Amin, H. U., Malik, A. S., Ahmad, R. F., Badruddin, N., Kamel, N., Hussain, M., and Chooi, W.-T. Feature extraction and classification for EEG signals using wavelet

transform and machine learning techniques. *Australasian Physical & Engineering Sciences in Medicine 38*, 1 (Mar 2015), 139–149.

[9] Atkinson, R. C., and Shiffrin, R. M. Human memory: A proposed system and its control processes. *Psychology of Learning and Motivation 2* (1968), 89–195.

[10] Baddeley, A. D., and Hitch, G. Working memory. *Psychology of learning and motivation 8* (1974), 47–89.

[11] Baddeley, A. D., Thomson, N., and Buchanan, M. Word length and the structure of short-term memory. *Journal of Verbal Learning and Verbal Behavior 14*, 6 (1975), 575–589.

[12] Bartolic, E., Basso, M., Schefft, B., Glauser, T., and Titanic-Schefft, M. Effects of experimentally-induced emotional states on frontal lobe cognitive task performance. *Neuropsychologia 37*, 6 (1999), 677 – 683.

[13] Benjamin, A. S. Predicting and postdicting the effects of word frequency on memory. *Memory & Cognition 31*, 2 (2003), 297–305.

[14] Berger, H. *Hans Berger on the electroencephalogram of man: The fourteen original reports on the human electroencephalogram.* Elsevier, 1969.

[15] Biddle, R., Chiasson, S., and Van Oorschot, P. Graphical passwords: Learning from the first twelve years. *ACM Comput. Surv. 44*, 4 (Sept. 2012), 19:1–19:41.

[16] Blocki, J., Komanduri, S., Procaccia, A., and Sheffet, O. Optimizing password composition policies. In *Proceedings of the Fourteenth ACM Conference on Electronic Commerce (EC 2013)* (New York, USA, 2013), ACM, pp. 105–122.

[17] Bonneau, J. The science of guessing: Analyzing an anonymized corpus of 70 million passwords. In *2012 IEEE Symposium on Security and Privacy* (May 2012), pp. 538–552.

[18] Bonneau, J., Herley, C., v. Oorschot, P. C., and Stajano, F. The quest to replace passwords: A framework for comparative evaluation of web authentication schemes. In *2012 IEEE Symposium on Security and Privacy* (May 2012), pp. 553–567.

[19] Bonneau, J., and Shutova, E. *Linguistic Properties of Multi-word Passphrases.* Springer Berlin Heidelberg, Berlin, Heidelberg, 2012, pp. 1–12.

[20] Bridger, E. K., Bader, R., and Mecklinger, A. More ways than one: ERPs reveal multiple familiarity signals in the word frequency mirror effect. *Neuropsychologia 57* (2014), 179–190.

[21] Brown, A. S., Bracken, E., Zoccoli, S., and Douglas, K. Generating and remembering passwords. *Applied Cognitive Psychology 18*, 6 (2004), 641–651.

[22] Burbidge, R., Trotter, M., Buxton, B., and Holden, S. Drug design by machine learning: Support vector machines for pharmaceutical data analysis. *Computers & Chemistry 26*, 1 (2001), 5–14.

[23] Burr, W. E., Dodson, D. F., and Polk, W. T. *Electronic authentication guideline*. US Department of Commerce, Technology Administration, National Institute of Standards and Technology, 2004.

[24] Chee, M. W., Westphal, C., Goh, J., Graham, S., and Song, A. W. Word frequency and subsequent memory effects studied using event-related fMRI. *NeuroImage 20*, 2 (2003), 1042–1051.

[25] Cisar, P., and Cisar, S. M. Password - a form of authentication. In *2007 5th International Symposium on Intelligent Systems and Informatics* (Aug 2007), pp. 29–32.

[26] Cohen, A., and Kovacevic, J. Wavelets: The mathematical background. *Proceedings of the IEEE 84*, 4 (Apr 1996), 514–522.

[27] Cohen, J. Statistical power analysis for the behavioral sciences 2nd edn, 1988.

[28] Cortes, C., and Vapnik, V. Support-vector networks. *Machine Learning 20*, 3 (Sep 1995), 273–297.

[29] Cowan, N. Chapter 20 what are the differences between long-term, short-term, and working memory? In *Essence of Memory*, W. S. Sossin, J.-C. Lacaille, V. F. Castellucci, and S. Belleville, Eds., vol. 169 of *Progress in Brain Research*. Elsevier, 2008, pp. 323 – 338.

[30] Curran, E. A., and Stokes, M. J. Learning to control brain activity: A review of the production and control of EEG components for driving brain–computer interface (BCI) systems. *Brain and Cognition 51*, 3 (2003), 326 – 336.

[31] Darwin, C. J., Turvey, M. T., and Crowder, R. G. An auditory analogue of the sperling partial report procedure: Evidence for brief auditory storage. *Cognitive Psychology 3*, 2 (1972), 255–267.

[32] Das, A., Bonneau, J., Caesar, M., Borisov, N., and Wang, X. The tangled web of password reuse. In *NDSS* (2014), vol. 14, pp. 23–26.

[33] De Alvaré, A. M. How crackers crack passwords or what passwords to avoid. Tech. rep., Lawrence Livermore National Lab., CA (USA), 1988.

[34] de Zubicaray, G. I., McMahon, K. L., Eastburn, M. M., Finnigan, S., and Humphreys, M. S. fMRI evidence of word frequency and strength effects in recognition memory. *Cognitive Brain Research 24*, 3 (2005), 587–598.

[35] DeLosh, E. L., and McDaniel, M. A. The role of order information in free recall: Application to the word-frequency effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition 22*, 5 (1996), 1136.

[36] 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* (2013), 56.

[37] Ekanayake, H. Research use of Emotiv Epoc, year = 2017, day = 22, note = http://neurofeedback.visaduma.info/emotivresearch_o.htm, annote = URL.

[38] EMOTIV Inc., 2017. http://emotiv.com/epoc.

[39] Fahl, S., Harbach, M., Acar, Y., and Smith, M. On the ecological validity of a password study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security (SOUPS 2013)* (New York, USA, 2013), ACM, pp. 13:1–13:13.

[40] Fahl, S., Harbach, M., Acar, Y., and Smith, M. On the ecological validity of a password study. In *Proceedings of the Ninth Symposium on Usable Privacy and Security* (New York, NY, USA, 2013), SOUPS '13, ACM, pp. 13:1–13:13.

[41] Florêncio, D., and Herley, C. A large-scale study of web password habits. In *Proceedings of the 16th International Conference on World Wide Web* (New York, NY, USA, 2007), WWW '07, ACM, pp. 657–666.

[42] Flotzinger, D., Pregenzer, M., and Pfurtscheller, G. Feature selection with distinction sensitive learning vector quantisation and genetic algorithms. In *Neural Networks, 1994. IEEE World Congress on Computational Intelligence., 1994 IEEE International Conference on* (June 1994), vol. 6, pp. 3448–3451 vol.6.

[43] Forget, A., Chiasson, S., van Oorschot, P. C., and Biddle, R. Improving text passwords through persuasion. In *Proceedings of the 4th Symposium on Usable Privacy and Security (SOUPS 2008)* (New York, USA, 2008), ACM, pp. 1–12.

[44] Friedman, D., and Trott, C. An event-related potential study of encoding in young and older adults. *Neuropsychologia 38*, 5 (2000), 542–557.

[45] Furey, T. S., Cristianini, N., Duffy, N., Bednarski, D. W., Schummer, M., and Haussler, D. Support vector machine classification and validation of cancer tissue samples using microarray expression data. *Bioinformatics 16*, 10 (2000), 906–914.

[46] Garrett, D., Peterson, D. A., Anderson, C. W., and Thaut, M. H. Comparison of linear, nonlinear, and feature selection methods for EEG signal classification. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 11*, 2 (June 2003), 141–144.

[47] Gevins, A., Smith, M. E., Leong, H., McEvoy, L., Whitfield, S., Du, R., and Rush, G. Monitoring working memory load during computer-based tasks with EEG pattern recognition methods. *Human Factors 40*, 1 (1998), 79–91.

[48] Gregg, V. Word frequency, recognition and recall.

[49] Guttentag, R., and Carroll, D. Memorability judgments for high-and low-frequency words. *Memory & Cognition 26*, 5 (1998), 951–958.

[50] hardware specifications, I. M., 2017. http://developer.choosemuse.com/hardware-firmware/hardware-specifications.

[51] Hayashi, E., and Hong, J. A diary study of password usage in daily life. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, USA, 2011), CHI '11, ACM, pp. 2627–2630.

[52] Herley, C., and Oorschot, P. V. A research agenda acknowledging the persistence of passwords. *IEEE Security Privacy 10*, 1 (Jan 2012), 28–36.

[53] Homrighausen, D., and McDonald, D. The lasso, persistence, and cross-validation. In *International Conference on Machine Learning* (2013), pp. 1031–1039.

[54] Horcher, A. M., and Tejay, G. P. Building a better password: The role of cognitive load in information security training. In *2009 IEEE International Conference on Intelligence and Security Informatics* (June 2009), pp. 113–118.

[55] Howard, M. W., and Kahana, M. J. A distributed representation of temporal context. *Journal of Mathematical Psychology 46*, 3 (2002), 269–299.

[56] Hu, D., Li, W., and Chen, X. Feature extraction of motor imagery EEG signals based on wavelet packet decomposition. In *The 2011 IEEE/ICME International Conference on Complex Medical Engineering* (May 2011), pp. 694–697.

[57] Huh, J. H., Kim, H., Bobba, R. B., Bashir, M. N., and Beznosov, K. On the memorability of system-generated pins: Can chunking help? In *Proceedings of the Eleventh Symposium On Usable Privacy and Security (SOUPS 2015)* (Ottawa, ON, Canada, 2015), USENIX Association, pp. 197–209.

[58] Hulme, C., Maughan, S., and Brown, G. D. Memory for familiar and unfamiliar words: Evidence for a long-term memory contribution to short-term memory span. *Journal of memory and language 30*, 6 (1991), 685–701.

[59] Hwang, H.-J., Kim, S., Choi, S., and Im, C.-H. EEG-Based Brain-Computer Interfaces: A Thorough Literature Survey. *International Journal of Human-Computer Interaction 29*, 12 (2013), 814–826.

[60] Inglesant, P. G., and Sasse, M. A. The true cost of unusable password policies: Password use in the wild. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2010), CHI '10, ACM, pp. 383–392.

[61] InteraXon Inc., 2017. http://www.choosemuse.com/.

[62] Jasper, H. H. The ten twenty electrode system of the international federation. *Electroencephalography and Clinical Neurophysiology 10* (1958), 371–375.

[63] Johnson, R. R., Popovic, D. P., Olmstead, R. E., Stikic, M., Levendowski, D. J., and Berka, C. Drowsiness/alertness algorithm development and validation using synchronized EEG and cognitive performance to individualize a generalized model. *Biological Psychology 87*, 2 (2011), 241 – 250.

[64] Jung, T.-P., Makeig, S., Westerfield, M., Townsend, J., Courchesne, E., and Sejnowski, T. J. Analysis and visualization of single-trial event-related potentials. *Human brain mapping 14*, 3 (2001), 166–185.

[65] Kamp, P.-H. Linkedin password leak: Salt their hide. *Queue 10*, 6 (June 2012), 20:20–20:22.

[66] KeePass Password Safe, 2017. http://keepass.info.

[67] Keith, M., Shao, B., and Steinbart, P. J. The usability of passphrases for authentication: An empirical field study. *International Journal of Human-Computer Studies 65*, 1 (2007), 17 – 28. Information security in the knowledge economy.

[68] Komanduri, S., Shay, R., Kelley, P. G., Mazurek, M. L., Bauer, L., Christin, N., Cranor, L. F., and Egelman, S. Of passwords and people: Measuring the effect of password-composition policies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (New York, NY, USA, 2011), CHI '11, ACM, pp. 2595–2604.

[69] Krigolson, O. E., Williams, C. C., and Colino, F. L. *Using Portable EEG to Assess Human Visual Attention*. Springer International Publishing, Cham, 2017, pp. 56–65.

[70] Krusienski, D., Sellers, E., McFarland, D., Vaughan, T., and Wolpaw, J. Toward enhanced p300 speller performance. *Journal of Neuroscience Methods 167*, 1 (2008), 15 – 21. Brain-Computer Interfaces (BCIs).

[71] Lotte, F., Congedo, M., Lécuyer, A., Lamarche, F., and Arnaldi, B. A review of classification algorithms for EEG-based brain–computer interfaces. *Journal of Neural Engineering 4*, 2 (2007), R1.

[72] Mallat, S. G. A theory for multiresolution signal decomposition: The wavelet representation. *IEEE Transactions on Pattern Analysis and Machine Intelligence 11*, 7 (Jul 1989), 674–693.

[73] Martin, M. V., Cho, V., and Aversano, G. Detection of subconscious face recognition using consumer-grade brain-computer interfaces. *ACM Trans. Appl. Percept. 14*, 1 (Aug. 2016), 7:1–7:20.

[74] McFarland, D. J., Sarnacki, W. A., and Wolpaw, J. R. Electroencephalographic (EEG) control of three-dimensional movement. *Journal of neural engineering 7*, 3 (2010), 036007.

[75] McFarland, D. J., and Wolpaw, J. R. Sensorimotor rhythm-based brain-computer interface (BCI): Feature selection by regression improves performance. *IEEE Transactions on Neural Systems and Rehabilitation Engineering 13*, 3 (Sept 2005), 372–379.

[76] McKenzie, W. A., and Tiberghien, G. Context effects in recognition memory: The role of familiarity and recollection. *Consciousness and Cognition 13*, 1 (2004), 20–38.

[77] Miller, G. A. The magical number seven, plus or minus two: Some limits on our capacity for processing information. *Psychological Review 63*, 2 (1956), 81.

[78] Müller, K.-R., Tangermann, M., Dornhege, G., Krauledat, M., Curio, G., and Blankertz, B. Machine learning for real-time single-trial eeg-analysis: From brain–computer interfacing to mental state monitoring. *Journal of Neuroscience Methods 167*, 1 (2008), 82 – 90. Brain-Computer Interfaces (BCIs).

[79] Mustafa, M., Guthe, S., and Magnor, M. Single-trial eeg classification of artifacts in videos. *ACM Transactions on Applied Perception (TAP) 9*, 3 (2012), 12.

[80] Nie, A., Griffin, M., Keinath, A., Walsh, M., Dittmann, A., and Reder, L. ERP profiles for face and word recognition are based on their status in semantic memory not their stimulus category. *Brain Research 1557* (2014), 66–73.

[81] Noh, E., Herzmann, G., Curran, T., and de Sa, V. R. Using single-trial EEG to predict and analyze subsequent memory. *NeuroImage 84* (2014), 712–723.

[82] O'Gorman, L. Comparing passwords, tokens, and biometrics for user authentication. *Proceedings of the IEEE 91*, 12 (Dec 2003), 2021–2040.

[83] Paller, K. A., McCarthy, G., and Wood, C. C. ERPs predictive of subsequent recall and recognition performance. *Biological Psychology 26*, 1 (1988), 269–276.

[84] Parkin, A. J. *Memory: Phenomena, experiment and theory.* Routledge, 2016.

[85] Pernet, C. R., Sajda, P., and Rousselet, G. A. Single-trial analyses: Why bother? *Frontiers in Psychology 2* (2011).

[86] Peterson, L., and Peterson, M. J. Short-term retention of individual verbal items. *Journal of experimental psychology 58*, 3 (1959), 193.

[87] Pion-Tonachini, L., Makeig, S., and Kreutz-Delgado, K. Crowd labeling latent dirichlet allocation. *Knowledge and Information Systems 53*, 3 (Dec 2017), 749–765.

[88] Raaijmakers, J. G. The story of the two-store model of memory: Past criticisms, current status, and future directions.

[89] Raaijmakers, J. G., and Shiffrin, R. M. Search of associative memory. *Psychological review 88*, 2 (1981), 93.

[90] Renaud, K. Quantifying the quality of web authentication mechanisms: A usability perspective. *Journal of Web Engineering 3*, 2 (2004), 95–123.

[91] Rodríguez-Bermúdez, G., García-Laencina, P. J., Roca-González, J., and Roca-Dorda, J. Efficient feature selection and linear discrimination of EEG signals. *Neurocomputing 115*, Supplement C (2013), 161 – 165.

[92] Sanquist, T. F., Rohrbaugh, J. W., Syndulko, K., and Lindsley, D. B. Electrocortical signs of levels of processing: Perceptual analysis and recognition memory. *Psychophysiology 17*, 6 (1980), 568–576.

[93] Shannon, C. E. A mathematical theory of communication, part i, part ii. *Bell System Technical Journal 27* (1948), 623–656.

[94] Shirley, J., and Evans, D. The user is not the enemy: Fighting malware by tracking user intentions. In *Proceedings of the 2008 New Security Paradigms Workshop* (New York, NY, USA, 2008), NSPW '08, ACM, pp. 33–45.

[95] Sperling, G. The information available in brief visual presentations. *Psychological Monographs: General and Applied 74*, 11 (1960), 1.

[96] SplashData, 2017. http://www.splashdata.com.

[97] Stanton, B. C., and Greene, K. K. Character strings, memory and passwords: What a recall study can tell us. In *Human Aspects of Information Security, Privacy, and Trust* (Cham, 2014), T. Tryfonas and I. Askoxylakis, Eds., Springer International Publishing, pp. 195–206.

[98] Subasi, A. EEG signal classification using wavelet feature extraction and a mixture of expert model. *Expert Systems with Applications 32*, 4 (2007), 1084 – 1093.

[99] Subasi, A., and Gursoy, M. I. EEG signal classification using PCA, ICA, LDA and support vector machines. *Expert Systems with Applications 37*, 12 (2010), 8659–8666.

[100] Surangsrirat, D., and Intarapanich, A. Analysis of the meditation brainwave from consumer EEG device. In *SoutheastCon 2015* (April 2015), pp. 1–6.

[101] Tamborello, F. P., and Green, K. K. Memory and motor processes of password entry error. In *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (2015), vol. 59, SAGE Publications Sage CA: Los Angeles, CA, pp. 672–676.

[102] Tan, D., and Nijholt, A. *Brain-Computer Interfaces and Human-Computer Interaction.* Springer London, London, 2010, pp. 3–19.

[103] Taneski, V., Heričko, M., and Brumen, B. Password security – no change in 35 years? In *2014 37th International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)* (May 2014), pp. 1360–1365.

[104] Telling Components Apart, 2017. http://labeling.ucsd.edu/tutorial/labels.

[105] Thorpe, J., Al-Badawi, M., MacRae, B., and Salehi-Abari, A. The presentation effect on graphical passwords. In *Proceedings of the 32Nd Annual ACM Conference on Human Factors in Computing Systems* (New York, NY, USA, 2014), CHI '14, ACM, pp. 2947–2950.

[106] Tibshirani, R. Regression shrinkage and selection via the lasso. *Journal of the Royal Statistical Society. Series B (Methodological) 58*, 1 (1996), 267–288.

[107] Tousignant, C., and Bodner, G. E. Test context affects recollection and familiarity ratings: Implications for measuring recognition experiences. *Consciousness and cognition 21*, 2 (2012), 994–1000.

[108] Tousignant, C., Bodner, G. E., and Arnold, M. M. Effects of context on recollection and familiarity experiences are task dependent. *Consciousness and cognition 33* (2015), 78–89.

[109] Travis, F., Tecce, J., Arenander, A., and Wallace, R. Patterns of EEG coherence, power, and contingent negative variation characterize the integration of transcendental and waking states. *Biological Psychology 61*, 3 (2002), 293 – 319.

[110] Treisman, A. Monitoring and storage of irrelevant messages in selective attention. *Journal of Verbal Learning and Verbal Behavior 3*, 6 (1964), 449–459.

[111] Übeyli, E. D. Analysis of EEG signals by combining eigenvector methods and multiclass support vector machines. *Computers in Biology and Medicine 38*, 1 (2008), 14–22.

[112] Uktveris, T., and Jusas, V. *Comparison of Feature Extraction Methods for EEG BCI Classification*. Springer International Publishing, Cham, 2015, pp. 81–92.

[113] Ur, B., Kelley, P. G., Komanduri, S., Lee, J., Maass, M., Mazurek, M. L., Passaro, T., Shay, R., Vidas, T., Bauer, L., et al. How does your password measure up? The effect of strength meters on password creation. In *21st USENIX Security Symposium (USENIX Security 12)* (2012), pp. 65–80.

[114] Ur, B., Segreti, S. M., Bauer, L., Christin, N., Cranor, L. F., Komanduri, S., Kurilova, D., Mazurek, M. L., Melicher, W., and Shay, R. Measuring real-world accuracies and biases in modeling password guessability. In *24th USENIX Security Symposium (USENIX Security 15)* (Washington, D.C., 2015), USENIX Association, pp. 463–481.

[115] Vega-Escobar, L., Castro-Ospina, A. E., and Duque-Muñoz, L. Feature extraction schemes for BCI systems. In *2015 20th Symposium on Signal Processing, Images and Computer Vision (STSIVA)* (Sept 2015), pp. 1–6.

[116] Veras, R., Thorpe, J., and Collins, C. Visualizing semantics in passwords: The role of dates. In *Proceedings of the Ninth International Symposium on Visualization for Cyber Security (VizSec 2012)* (New York, USA, 2012), ACM, pp. 88–95.

[117] Vidyaraman, S., Chandrasekaran, M., and Upadhyaya, S. Position: The user is the enemy. In *Proceedings of the 2007 Workshop on New Security Paradigms* (New York, NY, USA, 2008), NSPW '07, ACM, pp. 75–80.

[118] Volf, N. V., and Razumnikova, O. M. Sex differences in EEG coherence during a verbal memory task in normal adults. *International Journal of Psychophysiology 34*, 2 (1999), 113 – 122.

[119] Weirich, D., and Sasse, M. A. Pretty good persuasion: A first step towards effective password security in the real world. In *Proceedings of the 2001 Workshop on New Security Paradigms* (New York, NY, USA, 2001), NSPW '01, ACM, pp. 137–143.

[120] Wheeler, D. L. zxcvbn: Low-budget password strength estimation. In *25th USENIX Security Symposium (USENIX Security 16)* (Austin, TX, 2016), USENIX Association, pp. 157–173.

[121] Wiechert, G., Triff, M., Liu, Z., Yin, Z., Zhao, S., Zhong, Z., Zhaou, R., and Lingras, P. Identifying users and activities with cognitive signal processing from a wearable headband. In *2016 IEEE 15th International Conference on Cognitive Informatics Cognitive Computing (ICCI*CC)* (Aug 2016), pp. 129–136.

[122] Wiedenbeck, S., Waters, J., Birget, J.-C., Brodskiy, A., and Memon, N. Authentication using graphical passwords: Effects of tolerance and image choice. In *Proceedings of the 2005 Symposium on Usable Privacy and Security* (New York, NY, USA, 2005), SOUPS '05, ACM, pp. 1–12.

[123] Wolpaw, J. R., Birbaumer, N., Heetderks, W. J., McFarland, D. J., Peckham, P. H., Schalk, G., Donchin, E., Quatrano, L. A., Robinson, C. J., and Vaughan, T. M. Brain-computer interface technology: A review of the first international meeting. *IEEE Transactions on Rehabilitation Engineering 8*, 2 (Jun 2000), 164–173.

[124] Worthen, J. B., and Roark, B. Free recall accuracy for common and bizarre verbal information. *The American journal of psychology 115*, 3 (2001), 377–394.

[125] Yampolskiy, R. V. Analyzing user password selection behavior for reduction of password space. In *Proceedings 40th Annual 2006 International Carnahan Conference on Security Technology* (Oct 2006), pp. 109–115.

[126] Yan, J., Blackwell, A., Anderson, R., and Grant, A. Password memorability and security: Empirical results. *IEEE Security & Privacy 2*, 5 (Sept 2004), 25–31.

[127] Zheng, Z., Li, J., Xiao, F., Broster, L. S., Jiang, Y., and Xi, M. The effects of unitization on the contribution of familiarity and recollection processes to associative recognition memory: Evidence from event-related potentials. *International Journal of Psychophysiology 95*, 3 (2015), 355–362.