"Your Eyes Tell You Have Used This Password Before": Identifying Password Reuse from Gaze and Keystroke Dynamics

Yasmeen Abdrabou
Bundeswehr University Munich
Munich, Germany &
University of Glasgow
Glasgow, United Kingdom
yasmeen.essam@unibw.de

Johannes Schütte
Bundeswehr University Munich
Munich, Germany
johannes.schuette@unibw.de

Ahmed Shams
Fatura LLC
Cairo, Egypt
ahmed.shams@fatura-egypt.com

Ken Pfeuffer
Aarhus University
Aarhus, Denmark &
Bundeswehr University Munich
Munich, Germany
ken@cs.au.dk

Daniel Buschek
University of Bayreuth
Bayreuth, Germany
daniel.buschek@uni-bayreuth.de

Mohamed Khamis
University of Glasgow
Glasgow, United Kingdom
mohamed.khamis@glasgow.ac.uk

Florian Alt
Bundeswehr University Munich
Munich, Germany
florian.alt@unibw.de

ABSTRACT

A significant drawback of text passwords for end-user authentication is password reuse. We propose a novel approach to detect password reuse by leveraging gaze as well as typing behavior and study its accuracy. We collected gaze and typing behavior from 49 users while creating accounts for 1) a webmail client and 2) a news website. While most participants came up with a new password, 32% reported having reused an old password when setting up their accounts. We then compared different ML models to detect password reuse from the collected data. Our models achieve an accuracy of up to 87.7% in detecting password reuse from gaze, 75.8% accuracy from typing, and 88.75% when considering both types of behavior. We demonstrate that using gaze, password reuse can already be detected during the registration process, before users entered their password. Our work paves the road for developing novel interventions to prevent password reuse.

CCS CONCEPTS

• **Security and privacy → Usability in security and privacy**.

KEYWORDS

Passwords, Gaze Behavior, Keystroke Dynamics, Machine Learning

CHI '22, April 29-May 5, 2022, New Orleans, LA, USA

ACM ISBN 978-1-4503-9157-3/22/04…$15.00
https://doi.org/10.1145/3491102.3517531

ACM Reference Format:
Yasmeen Abdrabou, Johannes Schütte, Ahmed Shams, Ken Pfeuffer, Daniel Buschek, Mohamed Khamis, and Florian Alt. 2022. "Your Eyes Tell You Have Used This Password Before": Identifying Password Reuse from Gaze and Keystroke Dynamics. In *CHI Conference on Human Factors in Computing Systems (CHI '22), April 29-May 5, 2022, New Orleans, LA, USA.* ACM, New York, NY, USA, 16 pages. https://doi.org/10.1145/3491102.3517531

1 INTRODUCTION

After more than six decades, passwords remain a ubiquitous approach to authentication. While their end has been repeatedly predicted and other forms of authentication, such as fingerprint, facial recognition, and behavioral biometrics, have gained substantial popularity we are far from getting rid of passwords anytime soon [8]. The main reason is that passwords currently present a Pareto equilibrium between usability, security, and administrability [11], i.e. there is no other mechanisms providing an equally good trade-off between the effort required for implementation, ease of administration (e.g., reset / changing credentials), ease of use, and security.

At the same time, as a result of still having to remember too many and too complex passwords, users develop coping strategies (using simple passwords, writing down passwords) of which many compromise security. A particularly problematic strategy is the reuse of passwords. One reason is that if a reused password is leaked, attackers can easily gain access to other accounts of the user for which the same password is being used [23].

Having recognized this issue, both researchers and practitioners worked towards solutions. One popular approach is password managers. However, a substantial number of users are hesitant to use such password managers: a recent survey[1] ran by PasswordManager.com and YouGov among 1280 US American citizens showed

[1]Password Manager Survey: https://www.passwordmanager.com/password-manager-trust-survey/

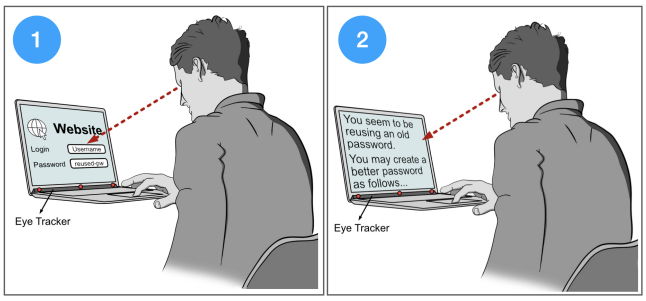


Figure 1: We investigate an approach to identify whether a user reuses a prior password during the registration process. In particular, we analyze eye movement and keystroke data while a user creates a password (1). We infer whether the user created a new password or reused an old one from the behavioral data only, without the need to know the actual password. The approach can serve as a basis for interventions to support users in creating more secure passwords (2).

that almost two thirds of participants do not trust password managers. Furthermore, prior work also showed that password managers not necessarily solve the issue, as a substantial number of password manager users still reuse passwords [39].

Preventing people from reusing passwords is a challenging task for several reasons: First, it requires knowledge about whether or not a user is reusing a password. One approach is *comparing the just created password to a database of known, breached passwords*. Yet, this does not prevent cases in which users are reusing a password that has, so far, not been leaked. Another approach is *comparing all passwords in use by a person* – this becomes possible as people are using a service to centrally manage their passwords (e.g., the aforementioned browser-based or standalone password managers). Such analyses are offered, as part of Google's password checkup[2] or as features of common password managers, such as LastPass's Security Challenge[3]. The drawback, again, is that a substantial number of people are not using password managers and post-hoc alerts on password breaches are often ignored by many users [49]. Furthermore, convincing people to post-hoc change their password is not easy. Prior work showed that even in cases where their passwords were verifiably breached, only 13% of users changed their passwords in the three months following the breach [10].

To overcome the aforementioned issues, we explore a novel approach to detect the password reuse based on sensing physiological user information. In particular, we assess users' gaze to infer the reuse of passwords (a) independent of people's password history, (b) without access to the actual password, and (c) already during the password creation process. Our approach is based on the assumption that cognition and behavior are different when reusing or creating a new password. For instance, users might "think harder" about a new password (which would affect fixations) and be required to direct their gaze to the input device more often, due to not having developed a motor memory of the password as a result of frequent use (which would affect the gaze path).

[2]Google Security Checkup: https://passwords.google.com/
[3]LastPass Security Challenge: http://blog.lastpass.com/2016/06/protecting-lastpass-users-from-password-reuse/

To investigate this concept, we collected data on gaze behavior and keystroke dynamics from 49 participants. In particular, we asked participants to create passwords for two types of accounts (a news website and a webmail client), protecting data of different sensitivity. We did not log participants' passwords, but asked them post-hoc, whether or not they reused any passwords. Similar to prior work, participants in about 30% of cases reused passwords.

Based on the collected data, we built prediction models using different machine learning classifiers. More specifically, we look at the different phases of the password registration process – (1) preparing for the registration (orientation), (2) entering the login / ID (identification), (3) entering the password (password), and (4) confirming the password (confirmation) – and analyze users' eye gaze during those phases as well as calculate prediction accuracy.

Our results show that by analyzing typing behavior only, an accuracy of up to 76% can be achieved, which is similar to accuracy in the literature. Predictions based on gaze increase the accuracy up to 88%. We also found that using gaze we can assess password reuse before users enter the password with an accuracy of 86%.

Contribution Statement. The contribution of our work is twofold. Firstly, we lay out and investigate the novel concept of assessing password reuse based on gaze data. Secondly, we provide an in-depth analysis of the approach. In particular, we (a) provide a comparison of gaze to other behavior data commonly available during password creation (that is typing behavior) and (b) analyze the behavior of users in the different phases of the password registration process as well as the possibility to predict password reuse during these phases using a Machine Learning-based approach.

Eye tracking is increasingly finding its way into users' everyday life and the value of (real-time) information on users' gaze behavior has been recognized by the usable security community [29]. Hence, we believe the research community as well as practitioners can benefit from our work in several ways. We envision that our approach can inspire researchers and designers to come up with novel concepts that better address password reuse. We see a particular potential of our approach in its *independence from the authentication interface*, in contrast to existing techniques where users have to enter their password first for it to be assessed. Our approach does not require any knowledge about the actual password, hence minimizing the attack surface. Furthermore, concepts can be implemented in a technology-independent way. For example, by using a mobile eye tracker, the system could detect password reuse on arbitrary devices, such as laptops, tablets, smartphones, or other surfaces. Interventions educating the user or helping them compose a better, unique password could be provided to the user via a smart watch or AR interface. Another strength is that through our concept of using gaze behavior as a means to detect password reuse, it will become feasible to *recognize password reuse instantly* and, in some cases, even before entering the password. This is not possible using keystroke dynamics. In this way, chances can be increased that users follow recommendations of not reusing passwords – compared to many current approaches hinting at password reuse post-hoc.

2 RELATED WORK

Our work draws from prior work on users' password habits and work on typing and gaze behavior in security contexts.

2.1 Users' Passwords Habits

People have on average 80 accounts for which they use 3.5 passwords. This makes password memorability challenging [23]. Coping strategies are choosing easy to remember passwords (e.g., 'password' or '123456'), reusing passwords, and writing down passwords. According to a survey by Google, 65% of users reuse passwords for some or all of their accounts[4]. Hence, the community focused on better understanding user behavior regarding password reuse and concepts to mitigate such behavior.

Wash et al. have studied users' password reuse behavior [52]. The authors created a Web browser plugin to collect user passwords across frequently used websites. Their results showed that people reuse strong passwords more frequently across different websites. Pearman et al. conducted an in-situ study to understand users' password managing behavior [38]. The authors found that the larger the number of accounts a user has, the higher the chances are that they reuse parts or all of their passwords across their accounts. This was also confirmed by another study done in 2006 by Florencio et al. [20]. Here, the authors assessed the average number of passwords and accounts users have and conducted a large scale study over 3 months to understand how many passwords users type per day, how often passwords are shared across sites, and how often users forget passwords. Findings show that on average participants have 6.5 passwords, each of which is shared across 3.9 different sites. In 2011, Campbell et al. [14] investigated the impact of imposing restrictive password composition rules on password choices made by users, such as requiring a minimum number of special or upper and lower case characters. They found that imposing password policies had a positive affect on password reuse, i.e. less people reused passwords if policies were enforced. The same was confirmed by Abbott et al., [1] in a study involving several US Universities. They found that stricter password policies led to a lower rate in reused passwords.

Researchers looked at users behavior when registering and using passwords. Shay et al. [46] show that more than half of participants modify an old password or reuse a password when signing up. Von Zezschwitz et al. [51] found through user interviews that 45% of users reuse the exact passwords. Hanamsagar et al. [23] found that after registration, participants 98% of the time reused the same passwords and in 2% of cases modified them. Data was collected using a Chrom extension, capturing passwords upon each attempt.

Reusing passwords can become a considerable threat for users as attackers get access to the server on which the password or a hash thereof is stored. As a result, attackers may use this information to impersonate the user for getting access to another account [23]. Prior work has investigated approaches to address this from a system perspective. For example, Das et al. [17] show how client-side password hashing can be used to generate unique passwords for different websites, thus helping mitigate the risk of password reuse. In addition, some systems enforce that passwords are not used beyond a certain time span, require minimum password length, or do not accept a password containing a sub string of a blacklisted password [45]. In the same direction, Seitz et al. suggested using dynamic password policies which adjust the password policy if a system detects a password that could be widely used [44].

[4]Google Survey: https://services.google.com/fh/files/blogs/google_security_infographic.pdf

Another counter-measure for password reuse is two- or multi-factor authentication. These solutions accept that passwords have weaknesses and try to mitigate this by requiring users to perform additional forms of authentication (e.g., entering a TAN). However, this comes at the expense of additional effort each time the user seeks to access an account. In contrast, our approach addresses the root cause, that is the password being insecure as a result of reuse. Rather than adding a burden upon each authentication attempt, our approach enables concepts that require additional effort only once, that is upon password registration. Note, that generally our approach can also be combined with multi-factor authentication. The result is that the password factor becomes stronger.

2.2 Gaze and Typing Behavior

Prior research looked into how knowledge on users' behavior can serve to enhance security mechanisms. We will particularly review work on typing and gaze behavior.

Much of prior work on *typing behavior* was motivated by the endeavor of building new authentication mechanisms based on behavioral biometrics. An early example is the work of Monrose et al. [36]. The authors showed that the way people type on a keyboard can be used to identify them. In particular, the authors identified latency between keystrokes, keystroke pressing duration, finger position on the keyboard and applied pressure on the keys as suitable features to build a classifier, based on which a user can be predicted. Buch et al. [13] looked at how users can be authenticated while writing longer texts, comparing copying text and entering free text. Similarly, Tappert et al. [48] built an authentication system based on free text entry, comparing different lengths entered on both laptop and desktop computers. The results suggest that the keyboard affects the classification accuracy. Hereby, typing on desktop keyboards led to a higher accuracy compared to laptops. Also the keyboard layout was shown to have a strong impact on typing behavior. Researchers compared different keyboards and languages [6, 7, 22, 35].

More recently, *gaze behavior* has moved into the focus of research. An ever-increasing number of mobile devices and laptops are being equipped with eye trackers [29]. Research showed how gaze behavior can be leveraged in different ways, for example, to detect personality traits [26] and to measure cognitive load [25]. At the same time, gaze has also been used for continuous verification [4, 16, 53] and for implicit identification [9, 15, 50]. In 2018, Katsini et al. [30], investigated users' visual behavior and how it relates to the strength of the created picture passwords. The authors used cognitive style theories to interpret their results. They show that users with different cognitive styles followed different patterns of visual behavior, affecting the strength of the created passwords. The findings introduce a new perspective for improving password strength in graphical user authentication. Furthermore, the authors looked at whether the strength of user-created graphical passwords can be estimated based on eye gaze behavior during password composition [31]. They analyzed unique fixations per area of interest (AOI) and the total fixation duration per AOI. Their results revealed a strong positive correlation between the strength of the passwords and the mentioned gaze features.

Abdrabou et al. showed that creating strong passwords increases users' cognitive load, reflected in users' pupil diameter [2]. They followed up by showing that gaze behavior can indicate password strength without revealing the actual password [3]. In both studies, participants created 12 weak and 12 strong passwords and entered half of them on a smartphone and the other half on a laptop.

2.3 Summary

From prior work we learn that password reuse is still a major challenge in usable security research. There are several reasons for this. Firstly, detecting password reuse is difficult. If a system has access to users' passwords, reused passwords can be detected by comparing them to corpus of leaked passwords or to other passwords of the user. Secondly, when designing concepts for password reuse mitigation, the time of the intervention plays an important role as, when being asked at a later point in time, people are rather unwilling to change their password [23]. We conclude, that being able to know as early as possible that users are about to reuse a password can be valuable when designing mitigation concepts.

Of particular interest is prior research that tried to infer password reuse from keystroke dynamics [28], achieving an accuracy of up to 81.71%. At the same time, prior work showed that the keyboard layout has a considerable influence on accuracy, suggesting that using other modalities might further increase the accuracy and the time at which a reasonable prediction can be made as well as enable novel opportunities for interventions. In addition, prior work has shown that gaze behavior differs between weak and strong graphical and text-based passwords. This led us to assume that reusing passwords might equally be reflected in users' gaze behavior.

Next, we will lay out the concept for using gaze as a means to detect reuse of text-based passwords and discuss study design considerations. We then present a proof-of-concept implementation and evaluation. To compare our work to prior research, we included detection password reuse from keystroke dynamics as a baseline.

3 CONCEPT AND RESEARCH QUESTIONS

We explore the concept of identifying the reuse of text-based passwords from gaze and typing behavior. The objective of our work is (1) to improve state-of-the-art by showing that the use of gaze can enhance the prediction accuracy, (2) to investigate how the prediction accuracy changes across different phases of the password creation process, and (3) to understand how the sensitivity of the data being protected by the passwords influences the approach.

We first provide background information on eye gaze analysis. Then we explain the different steps of the password creation process. Finally, we present the main research questions driving our work.

3.1 Gaze Behavior Analysis

Eye tracking research showed that from gaze, information can be derived on the user's state, intentions, and behavior. We explain how, based on different metrics, password reuse might be inferred.

Eye tracking provides information on where the user looks in the form of gaze points (fixations) and the transition between these (saccades). *Fixations* might provide valuable hints as to whether or not people are reusing passwords. The reason is that when reusing passwords, people can likely draw from motor memory (i.e. they know without looking how to enter the password). As a result, one can expect that people reusing a password fixate less on the input device (keyboard fixation count). Furthermore, the need to think about a new password is likely to result in a longer average fixation duration (fixation duration / average fixation duration) similar to literature where Katsini et al. found that users fixate longer while creating strong passwords [30]. Closely related is the *distribution of fixations*. We expect that users might, while trying to come up with a new password, differently distribute their gaze on the screen, resulting in longer/shorter saccades (saccadic length / average saccadic length) and in more/less time spent on transitioning between fixations (saccadic duration / average saccadic duration). In addition, we define two areas of interest (AOI): the screen with the authentication interface and the input device (here a keyboard).

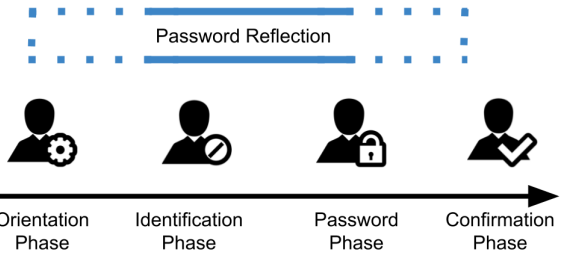


Figure 2: Phases of password registration: People first get familiar with the registration interface, then provide their ID and enter the password, and finally confirm their password. In parallel, they reflect on the password.

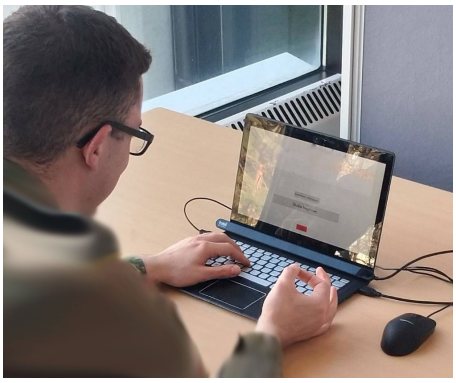


Figure 3: Study Setup: Participants were asked to register for two web services on a laptop. We logged keystroke dynamics and gaze using an eye tracker.

3.2 Phases of Password Creation

One important aspect of our work is *when* a system could predict password reuse based on gaze data. To investigate this, we decompose the password registration process:

Orientation Phase (O Phase) The authentication process begins with a phase of orientation, where the user is exposed

to the authentication interface. During this phase, the user not only gets familiar with the interface, but might already start to think about the password they will use. This phase begins when the authentication interface is displayed, and ends when the user begins to enter their ID.

Identification Phase (ID Phase) In the second phase, the user enters their user ID, which can be a user name or email address. Users might still continue to think about their password while they are already entering their identification information. The phase begins with the first keystroke of the user, as they start entering their ID and ends as the cursor is moved to the password field.

Password Phase (P Phase) In this phase, the user enters the password they thought about. It begins as the cursor is moved into the password field and ends as the user moves the cursor to the password confirmation field.

Confirmation Phase (C Phase) In the final phase, the user re-enters the password. This phase begins as the cursor is moved to the password confirmation field and ends as the user moves the cursor to the register button.

Figure 2 depicts the process. Note that users might have different strategies of when they think about the password they want to use. Whereas some users might think about the password already during the orientation phase, others might do so only after they entered their ID. Also, this reflection might span across multiple phases and it could be that users even during the identification phase think about the password.

3.3 Research Questions

Prior work used keystroke dynamics to detect if a password entered is new or reused [28]. We hypothesize that physiological signals better indicate password reuse. Hence, the first driving research question is: *How well can we predict the reuse passwords from gaze behavior, keystroke dynamics, or both (*RQ1*)?* We investigate the best gaze and typing features reflecting password reuse.

Second, we expect the sensitivity of protected data to play a role, resulting in the second driving question: *Is password reuse behavior different as passwords protect data with a different degree of sensitivity (*RQ2*)?* We compare behavior while creating a password for 1) a webmail client and 2) a customer account for a news website.

4 DATA COLLECTION

We conducted a data collection study in which we recorded users' gaze and typing behavior while creating passwords for two fictitious accounts, protecting data of different sensitivity.

4.1 Study Design Considerations

Our study design was driven by a number of considerations, most importantly how to observe natural user behavior, how to preserve privacy by not storing users' passwords, and how to minimize influences from the hardware.

Observing Natural User Behavior Haque et al. [24] showed the sensitivity of the data being protected by a password to have an influence on password choice. Participants create shorter and less secure password when registering a password for a website protecting less sensitive data. As a result, we followed common practice from the literature [2, 3], investigating both cases where users were to chose passwords protecting a web mail account (more sensitive data) and a news website account (less sensitive data).

Password Privacy Our study had two objectives regarding password use: (a) ensuring users chose reasonable passwords they could remember and (b) not storing the actual passwords (which would be necessary for password verification). To address this we only store password characteristics. For example, as users chose A!3, we would store the following information <upper case letter><special character><digit>. We used this information later to verify whether the re-entered password matched those characteristics. The trade-off is that we could not exactly verify the password. However, as this was not the purpose of this approach, we prioritized privacy.

Influence of Hardware Prior work on keystroke dynamics showed that the keyboard hardware has an influence on user behavior [43]. Hence, we decided to collect data from all participants using the same hardware and setup.

4.2 Study Design and Apparatus

We designed a within-subjects study with one independent variable (authentication interface), resulting in two conditions: 1) *Webmail Client* – a web-based authentication interface, meant to protect sensitive, personal email data. The interface resembled the webmail client of our University. 2) *News Website* – a web-based authentication interface, protecting less sensitive data. The interface resembled the authentication interface of a popular regional news website (see Figure 4).

All participants experienced both conditions in a counter-balanced order. We measured 8 dependent variables: duration for the password registration process, gaze metrics, keyboard metrics, time spent on each form field, password characteristics, and perceived password memorability. We did not store the raw password, but instead its length and the characteristics of each character, i.e., whether it was lowercase, uppercase, a number, or a symbol). For the apparatus we used a Lenovo Yoga 900s 12ISK laptop with a 12,5" screen (3200 × 1800 pixels) and off-the-shelf Tobii 4C eye tracker with a framerate of 90 Hz. We also implemented a demographics questionnaire at the end of the study. The questionnaire had questions about, age, gender, background, profession, experience with eye tracking and experience with IT security.

4.3 Study Setting, Procedure and Recruiting

We setup a booth in a quiet area of one of our local university's cafeteria (Figure 3). We approached people on campus and asked them to participate in the study. When participants agreed, we went with them to the cafeteria and asked them to sit at the booth. Participants were facing the booth wall to eliminate the influence of people in the vicinity.

We first asked participants to fill in a brief demographic questionnaire and a consent form. They were then told that we conducted a usability test of a slightly updated version of the University's web mail's password registration interface. Hereby, we specifically told them that the interface was not connected to the actual web mail

Figure 4: We rebuilt the webmail registration interface of the local University (left) and of a regional news website (right) to investigate differences in user behavior when creating passwords for accounts with more and less sensitive data.

system of the University. Furthermore, we explained that we compared it to the password registration interface of a regional news website. We also told them that we recorded gaze data to identify issues with the interface. After that, the eye tracker was calibrated using Tobii's 5 point calibration. Participants were asked to register an account for both websites. Participants were told that we did not store their passwords but that they had to remember them as they would be asked to later sign on with them. Participants were then shown the first registration page with three fields – one each for ID, password and password confirmation – and a register button (Figure 4). After participants had filled in the ID and passwords, they clicked the register button and were directed to the second interface, following the same procedure. Afterwards, participants were asked for each of the passwords how memorable they thought it was (5-Point Likert scale; 1=not memorable at all; 5=very memorable). Then, they were asked to log into both interfaces again in the order of registration. Finally, we wanted to know from participants whether they reused a password or created a new one. At the end of the study, we explained participants the true objective of the study and asked them to explain their strategy behind creating the passwords. On this occasion we were also able to clarify what password reuse means, if needed.

The experiment took around 10 minutes and participants were compensated with chocolates/treats. The study complied with our university's ethics requirements.

4.4 Limitations

We acknowledge the following limitation. Firstly, we cannot verify whether participants truthfully answered the questions regarding password reuse. Participants might have lied about non-compliant, insecure behavior. We tried to minimize any such influence by running the study in a completely anonymized way where no personal information was collected so as to establish trust. Furthermore, the percentage of reused passwords aligns with the literature, suggesting that participants mostly answered in a truthful way. Secondly, while the number of participants is in line with much similar prior work, we acknowledge the rather small size of our sample.

5 FEATURE EXTRACTION AND CLASSIFICATION

We describe our step-by-step process to evaluate eye gaze and keystroke dynamics for password reuse detection. First, we analyzed the collected passwords' characteristics and evaluated the effect of password type on password characteristics. Second, we extracted keystroke and gaze features required for classification and tested their statistical significance for the two types of passwords. Third, we built and tested different classifiers based on these features. We distinguish two categories: new and reused passwords. All features below were extracted for both categories.

5.1 Feature Extraction

We extracted a feature set describing keystroke dynamics and gaze behavior from the collected data in addition to password characteristics. We also analyze perceived password memorability.

5.1.1 Password Characteristics. We extracted the following password characteristics: password length, number of upper-case letters, number of lower-case letters, number of digits, and number of symbols. We also tracked the study duration, i.e. time in seconds from when the UI was shown until the 'Register' button was pressed.

5.1.2 Gaze Features. From the collected raw gaze data (X and Y positions on the screen), we derived the following characteristic eye movement features [27, 41]:

Fixations Count: Number of fixations performed during task.
Fixation Duration: Time for which users dwelled with their eyes on the laptop screen as well as on the keyboard.
Saccadic Length: Euclidian distance between two consecutive fixations with the eyes, determined in pixel.
Saccadic Duration: Duration between consecutive fixations.
Screen Fixation Count: Number of fixations on screen.
Keyboard Fixation Count: Number of fixations on keyboard.

The features are computed and analysed for each password phase, as well as over all phases.

5.1.3 Keystroke Dynamics Features. We collected 5 keystroke dynamics, informed by the literature [21, 28, 37].

Total Duration: Duration for typing email and password in milliseconds (not considering password confirmation).
Password Typing Duration: Time taken by the participant to enter the password in milliseconds.
Password Keystrokes Count: Number of keystrokes needed to type the passwords (including insertion, deletion).
Flight Time: Average latency between key presses in ms.
Pre-input Time: Time from the moment the interface was shown until the first key was pressed in milliseconds.

5.2 Classification Approach

The goal of our classifier is to map a feature vector computed from a time window of data to one of the classes corresponding to the password type (new vs reused). We first built an interface-dependent classifier, accounting for data sensitivity (webmail client vs news website). The classifier is trained on the data from different users but on the same interface. We then built an interface-independent classifier, not accounting for data sensitivity.

We used 3 feature sets: 1) keystroke features + password characteristics, 2) gaze features, and 3) both features combined. Keyboard and gaze data were saved and synchronized using the timestamp.

We compared the performance of three classifiers: Support Vector Machines (SVM), decision trees, and random forest, as done by Abdrabou et al. when detecting password strength [3]. To optimize performance, hyper parameters for each classifier were empirically optimized on a small set of values.

5.2.1 Interface-Dependent Classifier: Webmail Client vs. News Website. To understand how generalizable our approach is across different interfaces, we created interface-dependent classifiers by training the models on all users' data for each of the two interfaces separately. For each of the previously mentioned phases, we created one classifier. We implemented a two-fold cross validation. Figure 5 shows the steps for creating the classifier. We start with cleaning the data by removing the data outside our areas of interest (i.e., the screen and keyboard). During the pre-processing we assign the label 'new' or 'reused' to each sample, according to the participants' responses. After that we calculate the features for both gaze and keystroke dynamics. The collected data is synchronized using the timestamp for the analysis. This is followed by assigning the data to the 2 folds and running the classification. These steps are repeated for each phase. At the end, we report the AUC (Area Under the Curve) score which measures the ability of a classifier to distinguish between the two classes ('new' and 'reused') and is used as a summary of the ROC curve[5].

5.2.2 Interface-Independent Classifier: Both Interfaces. To understand whether a classifier working for interfaces protecting data of different sensitivity could be built, we created models that were independent of the data to be protected – in our case the web mail and the news page data. To do so, the classifier is trained on the data of all users and both interfaces. We split the data similar to the interface-dependent classifier into a training set and a test set.

[5]AUC: https://www.analyticsvidhya.com/blog/2020/06/auc-roc-curve-machine-learning/

6 RESULTS

In this section, we present and analyze the collected data.

6.1 Participants

A total of 52 participants (10 females) were recruited. The study ran over two weeks. Participants' age varied between 17 and 54 years ($M = 25.27$; $SD = 6.76$). 30 participants were students, 10 academic staff and the remaining 12 administrative staff. Most participants stated to be rather inexperienced with IT security (5-Point Likert scale; 1=no experience at all; 5=strong experience; $M = 2.23$; $SD = .35$). 23 participants wore glasses.

6.2 Data Pre-Processing and Overview

We removed data from 2 participants due to poor calibration quality. We lost data from one participant due to technical issues while saving. Overall we collected 98 passwords, half of which were created on the news website interface and the other half on the webmail interface. Table 1 shows the number of the newly created and reused password for each interface. As can be seen, participants reuse more passwords for the news website than for the webmail client. Participants needed on average 52 seconds to create a new password for the webmail interface and 42 seconds for the news website. In contrast, for the reused passwords, participants needed on average 38 seconds for the webmail interface and 25 seconds for the news website. A Wilcoxon test, revealed statistically significant differences between the study duration for reused and new passwords for the news website ($Z = -2.85, P = .004$) but not for the webmail client ($P > .05$). For both gaze and keystroke data, we sampled data at 90 Hz from the eye tracker and from key input events. This led to an average of 3149 samples per password, resulting in overall 340K samples for all participants for both interfaces.

6.3 New vs. Reused Passwords

We analyzed and compared cases where passwords were newly created or re-used.

Regarding *password memorability*, we found a statistically significant difference between reused ($M = 4.8; SD = .6$) and new passwords' memorability ($M = 3.9; SD = 1.1$) for the webmail client, ($Z = -2.226, P = .026$). This show that reused passwords (at least those protecting sensitive data), are more memorable than newly generated ones. Table 2 presents characteristics of passwords obtained during the study, and their distribution over conditions.

No statistically significant differences were found between the two interfaces regarding *password characteristics* (password length, number of digits / special characters / upper-case letters).

Table 3 summarizes findings regarding *keystroke features*. Our results indicate that participants took more time to think about and type new passwords compared to when reusing passwords. This includes shorter times when reusing passwords for pre-input time, typing duration and flight time.

Regarding *eye movement features*, we found several statistically significant differences between new and reused passwords (Table 4). The password type has a significant effect on several features for both interfaces. Furthermore, it shows that when considering both interfaces, for the reused passwords, users gaze was characterized by significantly shorter fixation times, shorter saccadic duration,

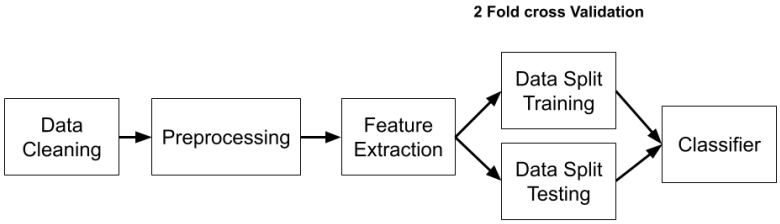


Figure 5: ML Classification Steps from data preparation until sending the data to the classifier.

Table 1: Number of new and reused passwords and task completion time.

	Webmail Client		News Website	
	New Passwords	Reused Passwords	New Passwords	Reused Passwords
Number of Passwords	35	14	31	18
Task Completion Time	52.28	37.89	42.07	25.99

Table 2: Wilcoxon signed-rank tests for both new and reuse password features on both interfaces. The results show that there is no statistically significant differences for the password characteristics between new and reuse passwords.

Password Characteristics	Email Interface			News Interface			Both Interfaces		
Feature	New Mean	Reuse Mean	Wilcoxon	New Mean	Reuse Mean	Wilcoxon	New Rank	Reuse Rank	Wilcoxon
Password Length	9.5	10.6	Z=-1.517, P>.05	9.5	10.3	Z=-.573, P>.05	10.4	10.3	Z=-1.154, P>.05
Upper-case Letters	1	1.1	Z=-.583, P>.05	.6	.6	Z=-.372, P>.05	0.8	0.8	Z=-.655, P>.05
Digits	3.3	3.2	Z=-.394, P>.05	2	3.3	Z=-1.800, P>.05	3.2	2.7	Z=-.892, P>.05
Symbols	.29	.71	Z=-1.403, P>.05	.3	.1	Z=-1.134, P>.05	0.4	0.3	Z=-.573, P>.05

Table 3: Wilcoxon signed-rank tests for keystroke features. For Webmail there is a significant effect of password type on the password typing duration. For the news website the password type had significant effects on flight time and thinking time.

Keystroke Feature	Webmail Client			News Website			Both Interfaces		
	New Mean	Reused Mean	Wilcoxon	New Mean	Reused Mean	Wilcoxon	New Mean	Reused Mean	Wilcoxon
Typing Duration	33.7	25.2	Z=-1.664, P >.05	27.5	16.9	Z=-1.764, P >.05	30.8	20.5	**Z=-2.711, P=.007**
Password Keystroke Count	16.5	13	Z=-.345, P >.05	13.6	12.3	Z=-.980, P >.05	15.1	12.6	Z=-.841, P >.05
Password Typing Duration	23	13.7	**Z=-2.103, P=.035**	15.8	10.2	Z=-1.851, P >.05	19.6	11.8	**Z=-3.048, P=.002**
Flight Time	1.7	1.1	Z=-1.852, P >.05	1.3	.9	**Z=-2.025, P=.043**	1.5	1	**Z=-3.160, P=.002**
Thinking Time	14.6	8.5	Z=-1.782, P >.05	7.4	4.2	**Z=-3.027, P=.002**	11.3	6	**Z=-3.586, P<.001**

less fixations, shorter saccades and less fixations on both the screen and keyboard. Overall, the many significant differences suggest eye movement features to be well suitable to accurately identify password reuse. We discuss practical implications in Section 8.

6.4 Gaze Path

As a complementary analysis, we visually inspected the eye movements in the form of the gaze path. Figure 6 shows some selected examples. We found that participants fixate more often on the screen (area 1) and keyboard (area 2) while creating new passwords, compared to when entering a reused password. This was independent of the interface on which passwords were created.

6.5 Classifier Performance

We compared the performance of three different models: SVM, random forest, and decision trees. We conducted two classifications: *phase-based classification* (i.e. per phase of the password registration) and *multiple phases classification.*

6.5.1 Phase-based Classification. We use data from the different registration phases (cf. Figure 2) to build the model. The phase-based model helped us understand how each phase contributes to the model. To understand which features are best for our classification task, we ran the classifier on gaze features only, keystroke features only, and both. Random forest and SVM classifiers resulted in a similar AUC (Area Under the Curve) score. However, SVM resulted in a better AUC score in most cases. Hence, the remainder of our analysis will focus on and report the SVM results.

For the *interface-dependent classifier*, Table 5 shows the overall performance of classification for each interface for all classifiers across the different phases. For webmail, the AUC is best when combining all phases. The highest AUC is 87.73% for gaze features and 88.75% for the combination of gaze and keystroke features. This means that users' behavior is more reflected in their gaze behavior features than in their typing behavior. Also, gaze features better reflect users' password behavior across the different phases. For the news website, similar to the webmail client, the best AUC is

Table 4: Wilcoxon signed-rank tests for the gaze features. The results show that for both Webmail and the News Website, the password type had a significant effect on serveral gaze features.

Gaze Feature	Webmail Client			News Website			Both Interfaces		
	New Mean	Reused Mean	Wilcoxon	New Mean	Reused Mean	Wilcoxon	New Mean	Reused Mean	Wilcoxon
Fixation Duration	28041.9	15728.1	**Z=-2.542, P =.011**	20143.4	13631.6	**Z=-2.330, P =.020**	24497.3	14548.7953	**Z=-3.964, P <.001**
Avg. Fixation Duration	222.8	203.1	Z=-1.66, P >.05	210.9	208.8	Z=-.152, P >.05	219.9	206.2945	**Z=-2.375, P =.018**
Saccadic Duration	20850.4	18704.9	Z=-.471, P >.05	18896.4	10490.1	**Z=-2.199, P =.028**	19988.7	14084.1108	**Z=-2.001, P =.045**
Avg Saccadic Duration	174.6	257	**Z=-2.982, P =.003**	196.6	171.1	Z=-.370, P >.05	186.2	207.3771	**Z=-2.618, P =.009**
Fixation Count	2595.8	1458	**Z=-2.542, P =.011**	1862.1	1265.7	**Z=-2.330, P =.020**	2266.4	1349.7500	**Z=-3.927, P <.001**
Avg. Fixation Count	.6	.5	**Z=-2.982, P =.003**	.6	.6	Z=-1.067, P >.05	.6	.5327	**Z=-3.385, P =.001**
Saccadic Length	1677.9	1539	Z=-.282, P >.05	1436	960.3	**Z=-2.199, P =.028**	1574.5	1213.2813	**Z=-2.094, P =.036**
Avg. Saccadic Length	.4	.5	**Z=-2.982, P =.003**	.4	.4	Z=-1.067, P>.05	.4	.4673	**Z=-3.385, P =.001**
Screen Fixation Count	2149.5	1193.9	**Z=-2.668, P =.008**	1690.7	1122.7	**Z=-2.461, P =.014**	1947.7	1153.8437	**Z=-3.843, P <.001**
Keyboard Fixation Count	446.3	264	Z=-1.915, P >.05	173.2	142.9	Z=-1.918, P >.05	318.8	195.9063	**Z=-2.786, P =.005**

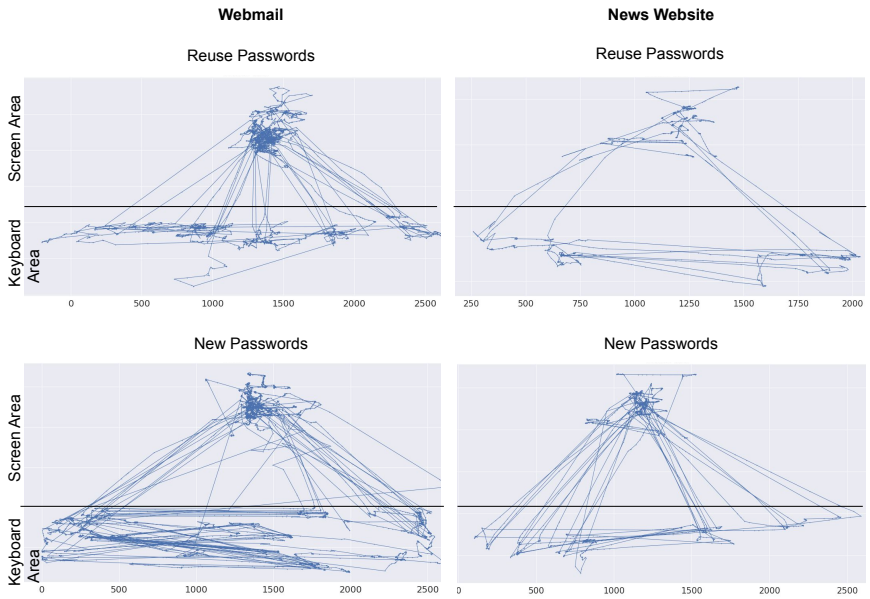


Figure 6: Visualization of selected users' gaze paths, both for the webmail (left) and news website (right) interface: In both cases, fixations are primarily focused on the input fields in the middle of the screen. Yet, for cases in which participants created new passwords, more transitions between screen and keyboard occur and more fixations are located in the keyboard area.

achieved when considering gaze features and the combination of gaze and keystroke features. The accuracy here is highest in the "identification phase" (84.56%). Our interpretation of this is that the password choice is primarily made during this phase. The keystroke features allow for an equally good prediction, but only when considering all phases. This means that for interfaces protecting sensitive content, password reuse is more accurately detected using gaze or both gaze/keystroke features during the identification phase.

For the *interface-independent classifier*, Table 6 shows the overall performance of the classifiers for all interfaces across the features. The highest AUC is achieved for gaze features and both features when combining all phases (71.87%).

6.5.2 Multiple-Phase Classification. This model accumulates all information available on users' behavior, from the beginning of the registration process to a particular phase. The aim of this model is to understand which features are best for classification. We ran the classifier on gaze features only, keystroke features only, and both. Random forest and SVM classifiers resulted in a similar AUC score. However, SVM resulted in a better AUC score in most cases. Hence, in the following we will focus on and report the SVM results.

For the *interface-dependent classifier*, Table 7 shows the overall performance for the classification for each interface across all classifiers for the accumulated phases. For webmail, the AUC is best, when all phases are combined. The highest AUC is 87.73% for gaze features. However, the model shows a decrease of only 2% for considering only the *O + ID phase* as well as when the *O + ID + P phases* are considered. This means that our model can predict password reuse in the identification phase *before* the user start typing the actual password reasonably well. For the keystroke features, the best AUC is still the same as the phase-based classification. However, looking at the accuracy after each phases along the registration process, we found a difference in accuracy of 6% across the grouped phases. This means that by using the keystroke features only, the best accuracy is achieved when the user has clicked 'register'. Finally, for both features combined, the picture was diverse.

Table 5: Interface-dependent Classifier: Classification Performance per Phase for the Different Features (best AUC bold)

Email Web-client		Orientation Phase (O Phase)		Identification Phase (ID Phase)		Password Phase (P Phase)		Confirmation Phase (C Phase)		All Phases	
		AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
Gaze Features	**SVM**	64.79 ± 9.50%	55.63 ± 1.51%	**74.35 ± 2.12%**	62.77 ± 9.46%	70.22 ± 4.78%	48.61 ± 1.39%	80.81 ± 0.67%	61.46 ± 5.21%	**87.73 ± 0.23%**	77.08 ± 14.58%
	Random Forest	**72.18 ± 0.76%**	55.87 ± 1.75%	67.62 ± 0.03%	56.94 ± 6.94%	**81.67 ± 8.14%**	61.29 ± 10.93%	**81.92 ± 0.44%**	55.90 ± 0.35%	83.44 ± 0.35%	61.46 ± 5.21%
	Decision Tree	49.05 ± 0.95%	61.54 ± 10.36%	60.33 ± 0.78%	58.33 ± 8.33%	55.56 ± 5.56%	65.75 ± 17.59%	56.83 ± 0.58%	60.33 ± 0.78%	75.84 ± 1.94%	72.22 ± 2.78%
Keystroke Features	**SVM**	-	-	53.40 ± 2.48%	52.78 ± 2.78%	66.23 ± 1.42%	49.14 ± 7.48%	54.89 ± 0.26%	50.00 ± 0.00%	63.58 ± 4.02%	68.06 ± 6.94%
	Random Forest	-	-	**61.04 ± 7.34%**	50.27 ± 3.04%	**69.23 ± 2.10%**	66.24 ± 0.43%	**75.54 ± 2.40%**	50.18 ± 0.18%	**75.83 ± 0.10%**	68.75 ± 6.25%
	Decision Tree	-	-	47.64 ± 0.89%	42.91 ± 4.31%	69.23 ± 2.10%	70.75 ± 1.31%	61.83 ± 6.69%	50.90 ± 6.45%	72.16 ± 5.62%	63.11 ± 3.55%
Both Features	**SVM**	-	-	**76.85 ± 1.85%**	62.77 ± 9.46%	71.51 ± 5.34%	48.61 ± 1.39%	**81.37 ± 1.96%**	61.46 ± 5.21%	87.73 ± 0.23%	78.47 ± 15.97%
	Random Forest	-	-	70.11 ± 0.26%	67.97 ± 4.08%	**80.47 ± 8.42%**	56.70 ± 9.48%	75.83 ± 0.10%	67.36 ± 4.86%	**88.75 ± 0.14%**	62.77 ± 9.46%
	Decision Tree	-	-	55.82 ± 2.51%	58.60 ± 5.29%	56.93 ± 3.25%	57.39 ± 3.72%	54.51 ± 1.74%	57.29 ± 1.04%	74.92 ± 2.86%	72.22 ± 2.78%
News Website		Orientation Phase (O Phase)		Identification Phase (ID Phase)		Password Phase		Confirmation Phase (C Phase)		All Phases	
		AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
Gaze Features	**SVM**	**67.35 ± 1.97%**	37.36 ± 9.80%	**84.56 ± 2.74%**	74.56 ± 0.44%	**77.82 ± 3.69%**	67.85 ± 7.36%	60.37 ± 2.33%	51.98 ± 1.98%	**77.49 ± 2.67%**	60.32 ± 10.32%
	Random Forest	48.00 ± 3.13%	52.56 ± 2.56%	78.82 ± 0.15%	63.28 ± 4.18%	75.23 ± 0.40%	60.54 ± 4.59%	**63.28 ± 4.55%**	61.09 ± 2.00%	73.94 ± 0.53%	55.16 ± 5.16%
	Decision Tree	43.07 ± 0.12%	44.99 ± 1.81%	60.85 ± 1.06%	60.05 ± 0.26%	62.99 ± 6.34%	69.53 ± 6.94%	48.77 ± 9.96%	49.14 ± 4.03%	63.19 ± 0.10%	55.16 ± 5.16%
Keystroke Features	**SVM**	-	-	**73.85 ± 3.92%**	73.92 ± 5.04%	54.36 ± 19.07%	76.35 ± 10.62%	**72.94 ± 4.68%**	56.24 ± 4.25%	74.65 ± 4.72%	66.16 ± 5.67%
	Random Forest	-	-	73.22 ± 0.21%	64.16 ± 0.52%	**67.53 ± 0.30%**	71.14 ± 9.95%	72.87 ± 0.14%	59.61 ± 5.07%	**80.97 ± 3.99%**	62.77 ± 0.87%
	Decision Tree	-	-	70.22 ± 3.55%	63.51 ± 6.77%	60.92 ± 0.43%	59.59 ± 1.60%	58.91 ± 0.18%	65.81 ± 6.02%	62.68 ± 2.36%	57.28 ± 2.51%
Both Features	**SVM**	-	-	**84.56 ± 2.74%**	74.56 ± 0.44%	**77.82 ± 3.69%**	66.38 ± 5.89%	61.61 ± 4.47%	51.98 ± 1.98%	76.70 ± 1.87%	60.32 ± 10.32%
	Random Forest	-	-	80.77 ± 1.40%	67.82 ± 0.36%	76.73 ± 2.26%	65.75 ± 5.26%	**78.21 ± 2.34%**	65.55 ± 1.91%	**77.96 ± 1.76%**	72.19 ± 4.00%
	Decision Tree	-	-	64.32 ± 3.14%	61.44 ± 1.65%	63.71 ± 2.37%	68.83 ± 7.64%	73.18 ± 3.74%	58.91 ± 0.18%	63.19 ± 0.10%	55.16 ± 5.16%

Table 6: Interface-independent Classifier: Classification Performance Per Phase for the Different Features (best AUC bold).

		Orientation Phase (O Phase)		Identification Phase (ID Phase)		Password Phase (P Phase)		Confirmation Phase (C Phase)		All Phases	
		AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
Gaze Features	**SVM**	**66.27 ± 0.95%**	46.91 ± 1.91%	**58.12 ± 2.58%**	49.59 ± 1.51%	59.28 ± 5.78%	63.95 ± 3.62%	65.64 ± 3.56%	51.19 ± 4.69%	**71.87 ± 2.82%**	51.84 ± 1.84%
	Random Forest	62.40 ± 1.00%	48.65 ± 2.42%	56.81 ± 0.07%	61.28 ± 1.49%	**76.91 ± 1.77%**	63.42 ± 15.34%	**68.37 ± 1.18%**	56.94 ± 0.12%	68.22 ± 0.06%	58.50 ± 0.07%
	Decision Tree	52.22 ± 1.04%	52.15 ± 1.27%	52.79 ± 5.73%	56.26 ± 2.25%	59.81 ± 2.58%	56.00 ± 6.09%	52.61 ± 0.42%	53.74 ± 7.62%	61.21 ± 1.53%	57.20 ± 2.48%
Keystroke Features	**SVM**	-	-	56.70 ± 0.85%	53.49 ± 3.49%	54.05 ± 0.69%	52.78 ± 2.78%	**64.43 ± 1.90%**	52.16 ± 1.98%	60.34 ± 0.43%	53.40 ± 3.40%
	Random Forest	-	-	**62.66 ± 1.63%**	57.40 ± 2.16%	**57.68 ± 1.29%**	61.10 ± 2.76%	59.91 ± 1.82%	49.94 ± 0.06%	**69.22 ± 0.49%**	61.88 ± 4.35%
	Decision Tree	-	-	61.77 ± 3.87%	54.23 ± 13.79%	55.75 ± 2.25%	56.06 ± 4.40%	55.90 ± 4.19%	61.93 ± 4.99%	66.28 ± 1.47%	57.80 ± 2.34%
Both Features	**SVM**	-	-	58.30 ± 2.41%	51.10 ± 3.02%	59.74 ± 6.27%	63.95 ± 3.62%	65.96 ± 2.84%	51.19 ± 4.69%	**71.87 ± 2.82%**	51.10 ± 1.10%
	Random Forest	-	-	**61.15 ± 0.94%**	59.88 ± 7.25%	**66.89 ± 1.65%**	58.77 ± 3.40%	**69.13 ± 0.26%**	57.46 ± 3.52%	70.73 ± 0.08%	64.03 ± 3.54%
	Decision Tree	-	-	57.99 ± 4.67%	56.37 ± 4.28%	52.39 ± 1.89%	59.01 ± 5.54%	60.51 ± 5.02%	56.65 ± 8.88%	62.41 ± 3.63%	57.20 ± 2.48%

For webmail, accuracy continuously increased. Yet, for the news website, the highest accuracy was achieved in the identification phase. In subsequent phases, accuracy differed minimally.

For the *interface-independent classifier*, combining the phases did not yield a better accuracy compared to phase-based classification. This indicates that for the interface independent classifier any model will lead to a similar accuracy.

6.5.3 True Positive and True Negative Values. As multiple phase classification did not affect the true positive and true negative rate, we only report values for the phase-based classification for the gaze features models. The data set was unbalanced. The guessing baseline (i.e. trivial classifier always guessing majority class) is 71% for webmail and 63% for the news website. Our classifiers outperform the baseline (81.6% for webmail, 74.6% for news website).

For *webmail* we found that 32 out of 35 new passwords were correctly classified as new. For the reused passwords, 8 out of the 14 reuse passwords were correctly classified. For the *news website* we found that out of the 31 newly generated passwords, 21 passwords were correctly classified as new. Out of the 18 reused passwords, 15 were correctly classified as reused. For the *interface independent classifier*, out of the 66 newly generated passwords, 56 were correctly classified as new. Out of the 32 reuse passwords, 12 were correctly classified as reuse. We reflect on these results in the discussion.

6.5.4 Feature Importance. We investigated which features mostly contribute to the accuracy of the classifiers. We found only small differences between both interfaces and here show the features for webmail only. We used SHAP [34], a tool that explain the output of a machine learning model by computing the contribution of each feature to its prediction. Figure 7 shows the feature importance.

We observed that for the gaze features, the fixation and registration duration are mostly contributing (.23 and .14 respectively). For the keystroke features, we observed that the overall registration duration and flight time contributed most to prediction of the password category (.09 and .06 respectively). For both features, we found that the gaze features have a stronger influence on the model's accuracy than the keystroke features.

6.5.5 Prediction Over Time. Figure 8 visualizes the *AUC* over time for the investigated conditions. Between interfaces, we can see that gaze leads to a higher accuracy much faster for webmail, i.e. when passwords are created to protect more sensitive data. The prediction accuracy for keystrokes is plateauing in the identification phase (i.e. after about 13 seconds for the news website and 22 seconds for

Table 7: Classification performance for interface-dependent classifier (multiple phases): Phases represented by O (orientation), ID (identification), and P (password entry). Best AUC in bold.

Email Web-client		O Phase		O + ID Phases		O + ID + P Phases		All Phases	
		AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
Gaze Features	**SVM**	64.79 ± 9.50%	52.06 ± 2.06%	77.11 ± 3.04%	73.61 ± 1.39%	85.04 ± 5.41%	54.17 ± 4.17%	**87.73 ± 0.23%**	71.53 ± 9.03%
	Random Forest	**72.18 ± 0.76%**	55.87 ± 1.75%	**85.68 ± 5.13%**	61.81 ± 0.69%	**85.16 ± 2.81%**	65.97 ± 3.47%	83.44 ± 0.35%	63.46 ± 2.35%
	Decision Tree	49.05 ± 0.95%	50.00 ± 0.00%	65.22 ± 0.52%	49.92 ± 2.86%	66.07 ± 6.15%	66.07 ± 6.15%	75.84 ± 1.94%	70.32 ± 7.46%
Keystroke Features	**SVM**	-	-	53.40 ± 2.48%	45.85 ± 1.37%	67.35 ± 1.17%	63.11 ± 3.55%	63.58 ± 4.02%	48.53 ± 1.47%
	Random Forest	-	-	**61.04 ± 7.34%**	52.78 ± 2.78%	65.36 ± 0.08%	54.17 ± 4.17%	**75.83 ± 0.10%**	61.81 ± 0.69%
	Decision Tree	-	-	47.64 ± 0.89%	40.30 ± 4.19%	**69.96 ± 7.82%**	68.85 ± 8.93%	72.16 ± 5.62%	72.16 ± 5.62%
Both Features	**SVM**	-	-	**77.60 ± 4.81%**	70.85 ± 1.37%	**85.04 ± 5.41%**	54.17 ± 4.17%	87.73 ± 0.23%	71.53 ± 9.03%
	Random Forest	-	-	77.40 ± 0.54%	61.38 ± 8.07%	84.50 ± 0.68%	65.62 ± 9.38%	**88.75 ± 0.14%**	63.11 ± 3.55%
	Decision Tree	-	-	65.22 ± 0.52%	49.92 ± 2.86%	66.07 ± 6.15%	66.07 ± 6.15%	74.92 ± 2.86%	70.32 ± 7.46%
News Website		O Phase		O + ID Phases		O + ID + P Phases		All Phases	
		AUC	Accuracy	AUC	Accuracy	AUC	Accuracy	AUC	Accuracy
Gaze Features	**SVM**	**67.35 ± 1.97%**	46.45 ± 0.99%	**83.62 ± 0.29%**	72.98 ± 4.80%	**81.43 ± 0.31%**	69.76 ± 0.88%	**77.49 ± 2.67%**	67.30 ± 6.11%
	Random Forest	48.00 ± 3.13%	52.88 ± 2.88%	76.20 ± 2.77%	68.33 ± 4.69%	77.61 ± 1.42%	75.05 ± 2.33%	73.94 ± 0.53%	61.80 ± 7.25%
	Decision Tree	43.07 ± 0.12%	43.07 ± 0.12%	60.05 ± 0.26%	60.05 ± 0.26%	70.46 ± 0.18%	70.46 ± 0.18%	63.19 ± 0.10%	56.55 ± 6.55%
Keystroke Features	**SVM**	-	-	**73.85 ± 3.92%**	56.15 ± 6.15%	71.12 ± 1.89%	60.51 ± 4.57%	74.65 ± 4.72%	66.07 ± 10.12%
	Random Forest	-	-	73.22 ± 0.21%	59.61 ± 5.07%	**75.11 ± 0.29%**	63.28 ± 4.18%	**80.97 ± 3.99%**	67.14 ± 3.50%
	Decision Tree	-	-	70.22 ± 3.55%	67.48 ± 2.80%	68.69 ± 1.55%	65.36 ± 4.87%	62.68 ± 2.36%	62.68 ± 2.36%
Both Features	**SVM**	-	-	**84.42 ± 0.50%**	73.68 ± 4.10%	**81.43 ± 0.31%**	72.73 ± 2.10%	76.70 ± 1.87%	61.11 ± 11.11%
	Random Forest	-	-	77.00 ± 0.78%	63.28 ± 4.18%	76.21 ± 0.02%	62.39 ± 7.85%	**77.96 ± 1.76%**	58.82 ± 4.27%
	Decision Tree	-	-	58.91 ± 0.18%	58.91 ± 0.18%	71.65 ± 1.37%	71.65 ± 1.37%	63.19 ± 0.10%	56.55 ± 6.55%

Table 8: Comparison of eye movements for the webmail client / news website (only factors with statistically significant effects).

Gaze Features	Saccadic Duration			Avg. Fixation Duration			Saccadic Length			Keyboard Fixations		
	Email Rank	News Rank	Wilcoxon	Webmail Rank	News Rank	Wilcoxon	Webmail Rank	News Rank	Wilcoxon	Webmail Rank	News Rank	Wilcoxon
Reuse Passwords	4.25	8.80	Z=-2.22, P=.026	5.25	8.40	Z=-1.97, P=.048	4.75	8.60	Z=-2.10, P=.035	7.50	7.50	Z=-2.35, P=.019

Table 9: Comparison of keystroke dynamics for the webmail client / news website (only factors with statistical significance).

Keystroke Features	Typing Duration			Keystrokes Count			Thinking Time		
	Email Rank	News Rank	Wilcoxon	Email Rank	News Rank	Wilcoxon	Email Rank	News Rank	Wilcoxon
Reuse Passwords	4.50	8.70	Z=2.17, P=.03	6.67	7.73	Z=-2.04, P=.041	4	7.90	Z=-2.34, P=.019

webmail). Gaze enables predictions are possible from the beginning of the identification phase, providing a time advantage.

6.6 Effect of Data Sensitivity on User Behavior

To study the effect of content sensitivity on user behavior, we ran a Wilcoxon signed-rank test on users' gaze features and keystroke features. We didn't find a statistically significant effect of data sensitivity, neither on gaze behavior nor on keystroke dynamics. However, for reused passwords, we found significant effects of data sensitivity on behavior.

Table 8 and 9 show the statistical significant features. For users' *gaze behavior*, we found significant differences for the saccadic duration, average fixation duration, saccadic length, and number of keyboard fixations between the webmail client (more sensitive) and the news website (less sensitive). For users' *keystroke dynamics*, we found statistical differences for users' typing duration, keystrokes count, and thinking time. The results show differences in users' behavior between interfaces protecting data with different sensitivity, but only when registering reused passwords.

7 DISCUSSION

We presented an investigation of eye movement behaviour and keystroke dynamics to identify whether people reuse passwords, specifically during the password registration phase. In the following, we discuss several insights gained from our study before discussing practical implications for authentication systems in the next section.

7.1 Gaze is More Informative than Typing

We found that a classifier based on gaze-related features (88% AUC for the interface-dependent classifier) outperforms a classifiers based on typing behavior only (80% AUC). Note that the results for typing behavior are in line with prior work [28]. Furthermore, the accuracy can be improved by *combining* typing and gaze features in some cases. Prediction accuracy for keystroke features is higher only at a later stage – namely after users have typed the password.

These findings answer *RQ1*. More specifically they show that it is not only possible to detect password reuse from these features but to also obtain rich additional information.

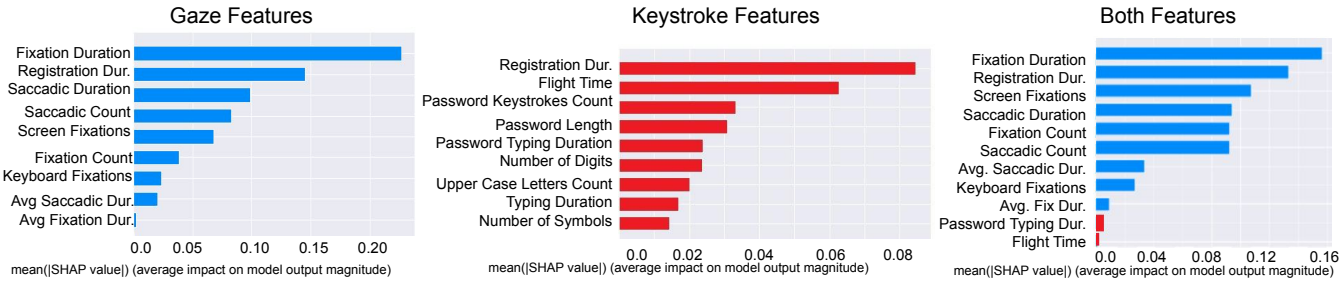
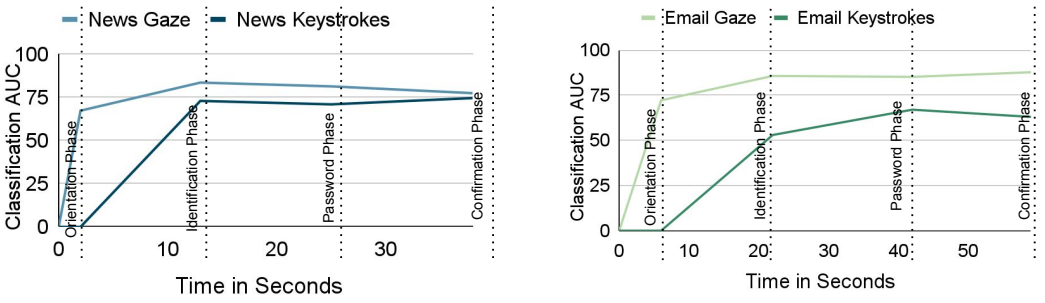


Figure 7: Results of the feature importance analysis across the tested feature groups for the email client.



(a)News Gaze and Keystroke AUC for cumulative multiple phases

(b)Email Gaze and Keystroke AUC for cumulative multiple phases

Figure 8: AUC comparison for multiple the phases classifier between gaze and keystroke features for new and reuse passwords across interfaces. It shows that our addition of using gaze outperformed using keystrokes.

7.2 Data Sensitivity Influences Accuracy of Password Reuse Prediction

We found that the sensitivity of the protected data affects the characteristics of the chosen password and whether it is a new one or a reused one is reflected in the user's gaze data. More participants reused passwords for the news website than for the webmail client. This suggests that the more sensitive the protected information is, the more effort people put into their password and the less often they reuse passwords. This also leads to users' behavior getting more distinguishable. This is revealed by the statistical analysis where, for the webmail client, most features (gaze and typing) were significantly different between reused and new passwords. In contrast for the news website we could not find significant difference in our collected data.

7.3 Dissecting Password Registration Process Enriches Modeling and Prediction

Contributing to the literature, we dissected the observation of password creation behavior into multiples phases.

For the webmail client, we found that considering users' behavior during the whole password generation (all phases combined) to detect password reuse leads to the best accuracy. In contrast, for the news website, we found that the identification phase better reflects users' behavior to detect password reuse. This suggests that people think about passwords during different phases of the registration process and that this thinking takes longer when protecting more sensitive data. We ran a Wilcoxon test to see whether the duration of the identification phase differed for new ($MeanRank = 10.27$) and reused passwords ($MeanRank = 8.29$) for the news website. We did not find statistical significant differences ($Z = -1.98$, $p > .05$). This motivates a future study, striving to obtain a deeper understanding of when and how much people 'think ahead' when creating passwords.

8 PRACTICAL IMPLICATIONS FOR THE DESIGN OF PASSWORD SYSTEMS

Being able to identify password reuse before the end of the registration process, we envision interfaces to implement interventions ultimately leading to better passwords. We reflect on the role of eye tracking, the design of interventions, the implications of user and interface characteristics on modeling, and on privacy implications.

8.1 Ubiquitous Eye Tracking

We believe the vision sketched in this paper to be timely as eye tracking is about to become ubiquitously available and to, in particular, gain relevance in usable security [29]. Access to gaze data is possible today in different ways. Firstly, there is laptop and desktop computers being equipped with dedicated eye tracking hardware. The fact that Apple bought SMI, one of the world's leading manufacturers of eye tracking hardware suggests, that one of the next generations of Macbooks might come with integrated eye tracking. Secondly, advances in computer vision made it possible to perform appearance-based gaze estimation simply by means of analyzing the video feed of a web cam or smartphone cam [32]. Thirdly, eye wear (such as augmented reality glasses and head-worn devices) are envisioned to use gaze as a communication medium for everyday interactions [40], and thus could open doors for security use cases.

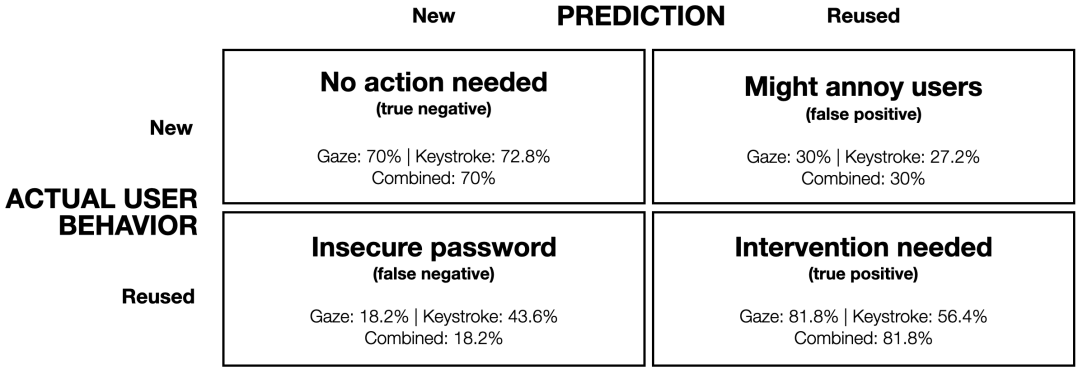


Figure 9: Interplay of Actual User Behavior and System Prediction (normalized confusion matrix). Optimally, a system would correctly predict whether a password is new or reused. In the first case, no action would be needed – our approach predicts this case with around 70% accuracy. In the second case an intervention should be shown – we predict this case with around 80% accuracy. Interestingly, gaze is particularly powerful for reused passwords, where a prediction based on keystrokes is only successful in about 56% of cases.

Our approach could be implemented in various forms. Providers wishing to support users in choosing better passwords could integrate the approach with their password registration interface (e.g., by accessing the webcam on a PC, by a smartphone app accessing the front-facing camera, or the built-in eye tracker of head-worn devices). A provider-independent solution would be a browser plugin that accesses the camera and assesses users' gaze data as they enter a website requiring the registration of a password. Finally, the approach could run as a service in eye wear, that activates when users are about to register a password and then assesses their physiological data.

8.2 Creating Design Interventions

A system that integrates the predictive model can provide several interventions based on the outcome of the prediction. Figure 9 depicts four different cases based on two dimensions. The first dimension is the actual behavior of the user, i.e. whether they used a new password or reused an old one. The second dimension is the prediction of the system, i.e. whether the system thinks the user created a new password or reused an old one.

New user password + system predicts new password No action is needed as this is the optimal behavior.

Reused user password + system predicts reused password In this case, the system presents an intervention that optimally motivates the user to rethink their choice.

New user password + system predicts reuse Interventions by the system may lead to potential adverse effect to the users and should be avoided. This should carefully weigh off potential factors, e.g., the more invasive the intervention is (e.g., forcing the user to enter a new password), the more negative it can influence user perception. Providing options to easily cancel this will become handy to the user.

Reused user password + system predicts new Here, a system would not intervene. Hence, the user would not be bothered, but potentially use an insecure password. This should be minimized for cases requiring high security.

Based on the accuracy of the trained model, designers could verify how likely the above-mentioned cases are and decide, which interventions are suitable, regarding their level of invasiveness. Other factors could influence this choice, e.g., how important it is that users do not reuse a password. Interventions could take various forms, as proposed in the literature: warnings, i.e. reminding users about security risks resulting from their behavior [33], attractors, i.e., modifications in the UI that draw user's attention to important information for decision making [12], or nudges, i.e. interventions that guide users to make beneficial suggestion [5, 19, 42].

8.3 Modeling

Different factors can influence the classification modeling.

8.3.1 Ground Truth: Determining Password Reuse. The first step to building predictive models is to *collect behavioral data* during the authentication process. The challenge during this data collection is to *obtain a ground truth*, i.e. whether or not users are creating new password or reusing an existing ones. Several alternatives exist. Firstly, users could be asked to provide this information. Yet, this creates an overhead for the user. Secondly, the created password could be compared to (the hashes of) passwords other users created for the data or service the mechanism is protecting. Third, the created password could be compared to databases of leaked passwords. Afterwards, a model can be trained based on the labeled set of behavioral data, following the approach outlined in this paper.

8.3.2 Influence of Typing Proficiency. In our study, we sampled among a University population where people were likely to have a rather high typing proficiency. However, this might be different for other samples. Typing behavior is mainly a result of how long people type daily. In addition, typing and keystroke dynamics are influenced by cognition, which differs when typing routine words (i.e. password reuse) as opposed to non-routine words (i.e. new passwords) [28]. Dhakal et al. [18] analyzed typing behavior in an online survey and they clustered typists into eight groups based on their typing performance, accuracy, rollover, and hand usage.

Given all this, we learn that user's typing proficiency plays a role to affect keystroke behavior and, hence, the accuracy of a classifier predicting password reuse. A user-dependent model is more suitable to capture individual characteristics and can enhance accuracy.

8.3.3 Influence of Screen Properties. Users might access the same password registration interface on devices with different screen properties (e.g., a laptop vs. a large external monitor). While we maintained the same screen in our study for data consistency, other display types might be worth considering. In our analysis, we inspected the degree of influence the features have on prediction accuracy. Fixation and registration durations are among the most prominent features. We expect the influence of the screen properties on such relative features to be low. However, to further enhance the classification accuracy and take into account device-dependent features such as saccadic duration and path, it might be useful to consider screen-optimised classifiers.

8.3.4 Influence of Layout. Ideally, a model would make highly accurate predictions independent of the password registration interface layout. In our study, we investigated two examples from the real world that we believe are representative for many of the layouts in use. However, other registration interfaces might look different and ask the user, for example, to provide information beyond credentials on the same page, such as an address or payment information. One might speculate whether users already display behavior related to password composition before working on the respective part of the form. If so, this would be interesting, as it gives a system employing our concept more time for an intervention and also more typing and gaze data. At the same time this would require a new model to be trained.

Future work could investigate, how exactly the registration interface, in particular, the requested information and the layout (e.g., at which part of the registration interface the password is composed) influence prediction accuracy.

8.3.5 Influence of Interaction Modality. We hypothesize that different interaction modalities will likely affect typing behavior, because input devices vary across systems (e.g., using a mechanical vs. a soft keyboard). The same is potentially true for gaze as different forms of eye trackers might be employed with different systems and typing behavior might influence gaze behavior in a different way. At the same time, it is plausible that the implicit nature of eye movements could represent a more constant predictor of password reuse across systems. This should be pursued by future research.

8.4 User Privacy

Note that it is important to consider the potential privacy implications of using gaze data. There is an ongoing discussion on the need to use gaze data carefully. From gaze, information beyond password reuse can be inferred, including but not limited to the users' interest, attention, fatigue, or sexual orientation (see Steil et al. [47] for an in-depth assessment of this topic). One could assume that users might be willing to share gaze data if it was to their benefit, in particular, in a security context. Yet, consent to collect and assess gaze data should not only be obtained by the provider of a password reuse identification system but be limited to this authentication procedure.

9 FUTURE WORK

Our work opens up many avenues for future research. Firstly, as mentioned above, one interesting direction is to investigate the *influence of the interface properties on the concept*, in particular, the integration of password registration with the assessment of other information. Secondly, we plan create novel *interventions* that prevent password reuse or that nudge users towards rethinking their strategy. The choice for the intervention might be based on the prediction and could also take the likeliness for password reuse into account. We are also interested in understanding during which phases of the password registration process this is most effective. Thirdly, we plan to explore how concepts that are *independent of the input device* can be realized – for example, password reuse is detected through a mobile eye tracker and interventions are then provided as AR overlay or on a smart watch. A final direction for future research might be *investigating additional types of user behavior and physiological states* to predict password reuse.

10 CONCLUSION

We presented a novel approach for predicting password reuse. We separated password registration into different phases, namely the 1) orientation phase, 2) identification phase, 3) password typing phase, and 4) confirmation phase. We then looked at how well password reuse can be detected in the different phases (separately and accumulated) based on gaze, keystroke dynamics and both. In addition, we compared two interfaces, meant to protect more and less sensitive data.Beyond showing that our approach improves the accuracy of prior work, we additionally demonstrated that prediction becomes now feasible throughout the entire password registration process. In addition, we provide insights how gaze and typing feature contribute to detecting password reuse and reflect on the practical implications of our findings. We hope to have provided a powerful approach for researchers and practitioners based on which novel interventions mitigating password reuse can be built.

ACKNOWLEDGMENTS

This work was supported by the Royal Society of Edinburgh (RSE award no. 65040 and 1931), the PETRAS National Centre of Excellence for IoT Systems Cybersecurity, which has been funded by the UK EPSRC under grant number EP/S035362/1, EPSRC New Investigator Award (EP/V008870/1), DFG grant no. 316457582 and 425869382, dtec.bw-Digitalization and Technology Research Center of the Bundeswehr (Voice of Wisdom), and the Studienstiftung des deutschen Volkes. This project was also partly funded by the Bavarian State Ministry of Science and the Arts and coordinated by the Bavarian Research Institute for Digital Transformation (bidt).

REFERENCES

[1] Jacob Abbott, Daniel Calarco, and L Jean Camp. 2018. Factors influencing password reuse: A case study. In *Telecommunications Policy Research Conference on Communications, Information and Internet Policy (TPRC 46). DOI: http://dx. doi. org/10.2139/ssrn*, Vol. 3142270.

[2] Yasmeen Abdrabou, Yomna Abdelrahman, Mohamed Khamis, and Florian Alt. 2021. *Think Harder! Investigating the Effect of Password Strength on Cognitive Load during Password Creation.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3411763.3451636

[3] Yasmeen Abdrabou, Ahmed Shams, Mohamed Omar Mantawy, Anam Ahmad Khan, Mohamed Khamis, Florian Alt, and Yomna Abdelrahman. 2021. *GazeMeter: Exploring the Usage of Gaze Behaviour to Enhance Password Assessments.* Association for Computing Machinery, New York, NY, USA. https://doi.org/10.1145/3448017.3457384

[4] E. R. Abdulin and O. V. Komogortsev. 2015. Person verification via eye movement-driven text reading model. In *2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS).* 1–8.

[5] Alessandro Acquisti, Idris Adjerid, Rebecca Balebako, Laura Brandimarte, Lorrie Faith Cranor, Saranga Komanduri, Pedro Giovanni Leon, Norman Sadeh, Florian Schaub, Manya Sleeper, Yang Wang, and Shomir Wilson. 2017. Nudges for Privacy and Security: Understanding and Assisting Users' Choices Online. *ACM Comput. Surv.* 50, 3, Article 44 (Aug. 2017), 41 pages. https://doi.org/10.1145/3054926

[6] Suliman A Alsuhibany, Muna Almushyti, Noorah Alghasham, and Fatimah Alkhudhayr. 2019. The impact of using different keyboards on free-text keystroke dynamics authentication for Arabic language. *Information & Computer Security* (2019).

[7] Suliman A Alsuhibany, Muna Almushyti, Noorah Alghasham, and Fatimah Alkhudier. 2016. Analysis of free-text keystroke dynamics for Arabic language using Euclidean distance. In *2016 12th International Conference on Innovations in Information Technology (IIT).* IEEE, 1–6.

[8] Florian Alt and Stefan Schneegass. 2021. Beyond Passwords—Challenges and Opportunities of Future Authentication. *IEEE Security & Privacy* (2021).

[9] Akram Bayat and Marc Pomplun. 2018. Biometric Identification Through Eye-Movement Patterns. In *Advances in Human Factors in Simulation and Modeling*, Daniel N. Cassenti (Ed.). Springer International Publishing, Cham, 583–594.

[10] Sruti Bhagavatula, Lujo Bauer, and Apu Kapadia. 2020. (How) Do people change their passwords after a breach? *arXiv preprint arXiv:2010.09853* (2020).

[11] Joseph Bonneau, Cormac Herley, Paul C. van Oorschot, and Frank Stajano. 2015. Passwords and the Evolution of Imperfect Authentication. *Commun. ACM* 58, 7 (jun 2015), 78–87. https://doi.org/10.1145/2699390

[12] Cristian Bravo-Lillo, Saranga Komanduri, Lorrie Faith Cranor, Robert W. Reeder, Manya Sleeper, Julie Downs, and Stuart Schechter. 2013. Your Attention Please: Designing Security-Decision UIs to Make Genuine Risks Harder to Ignore. In *Proceedings of the Ninth Symposium on Usable Privacy and Security* (Newcastle, United Kingdom) *(SOUPS '13).* Association for Computing Machinery, New York, NY, USA, Article 6, 12 pages. https://doi.org/10.1145/2501604.2501610

[13] Tarjani Buch, Andreea Cotoranu, Eric Jeskey, Florin Tihon, and Mary Villani. 2008. An enhanced keystroke biometric system and associated studies. *Proc. CSIS Research Day, Pace Univ* (2008).

[14] John Campbell, Wanli Ma, and Dale Kleeman. 2011. Impact of restrictive composition policy on user password choices. *Behaviour & Information Technology* 30, 3 (2011), 379–388.

[15] Virginio Cantoni, Chiara Galdi, Michele Nappi, Marco Porta, and Daniel Riccio. 2015. GANT: Gaze analysis technique for human identification. *Pattern Recognition* 48, 4 (2015), 1027–1038.

[16] Virginio Cantoni, Tomas Lacovara, Marco Porta, and Haochen Wang. 2018. A Study on Gaze-Controlled PIN Input with Biometric Data Analysis. In *Proceedings of the 19th International Conference on Computer Systems and Technologies* (Ruse, Bulgaria) *(CompSysTech'18).* Association for Computing Machinery, New York, NY, USA, 99–103. https://doi.org/10.1145/3274005.3274029

[17] Anupam Das, Joseph Bonneau, Matthew Caesar, Nikita Borisov, and XiaoFeng Wang. 2014. The tangled web of password reuse.. In *NDSS*, Vol. 14. 23–26.

[18] Vivek Dhakal, Anna Maria Feit, Per Ola Kristensson, and Antti Oulasvirta. 2018. *Observations on Typing from 136 Million Keystrokes.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174220

[19] Serge Egelman, Andreas Sotirakopoulos, Ildar Muslukhov, Konstantin Beznosov, and Cormac Herley. 2013. *Does My Password Go up to Eleven? The Impact of Password Meters on Password Selection.* Association for Computing Machinery, New York, NY, USA, 2379–2388. https://doi.org/10.1145/2470654.2481329

[20] Dinei Florencio and Cormac Herley. 2007. A Large-Scale Study of Web Password Habits. In *Proceedings of the 16th International Conference on World Wide Web* (Banff, Alberta, Canada) *(WWW '07).* Association for Computing Machinery, New York, NY, USA, 657–666. https://doi.org/10.1145/1242572.1242661

[21] Donald R. Gentner. 1983. *Keystroke Timing in Transcription Typing.* Springer New York, New York, NY, 95–120. https://doi.org/10.1007/978-1-4612-5470-6_5

[22] Daniele Gunetti, Claudia Picardi, and Giancarlo Ruffo. 2005. Keystroke Analysis of Different Languages: A Case Study. In *Advances in Intelligent Data Analysis VI*, A. Fazel Famili, Joost N. Kok, José M. Peña, Arno Siebes, and Ad Feelders (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 133–144.

[23] Ameya Hanamsagar, Simon S. Woo, Chris Kanich, and Jelena Mirkovic. 2018. *Leveraging Semantic Transformation to Investigate Password Habits and Their Causes.* Association for Computing Machinery, New York, NY, USA, 1–12. https://doi.org/10.1145/3173574.3174144

[24] S.M. Taiabul Haque, Matthew Wright, and Shannon Scielzo. 2013. A Study of User Password Strategy for Multiple Accounts. In *Proceedings of the Third ACM Conference on Data and Application Security and Privacy* (San Antonio, Texas, USA) *(CODASPY '13).* Association for Computing Machinery, New York, NY, USA, 173–176. https://doi.org/10.1145/2435349.2435373

[25] John M Henderson, Svetlana V Shinkareva, Jing Wang, Steven G Luke, and Jenn Olejarczyk. 2013. Predicting cognitive state from eye movements. *PloS one* 8, 5 (2013), e64937.

[26] Sabrina Hoppe, Tobias Loetscher, Stephanie A. Morey, and Andreas Bulling. 2018. Eye Movements During Everyday Behavior Predict Personality Traits. *Frontiers in Human Neuroscience* 12 (2018), 105. https://doi.org/10.3389/fnhum.2018.00105

[27] Robert JK Jacob and Keith S Karn. 2003. Eye tracking in human-computer interaction and usability research: Ready to deliver the promises. In *The mind's eye.* Elsevier, 573–605.

[28] Jeffrey L. Jenkins, Mark Grimes, Jeffrey Gainer Proudfoot, and Paul Benjamin Lowry. 2014. Improving Password Cybersecurity Through Inexpensive and Minimally Invasive Means: Detecting and Deterring Password Reuse Through Keystroke-Dynamics Monitoring and Just-in-Time Fear Appeals. *Information Technology for Development* 20, 2 (2014), 196–213. https://doi.org/10.1080/02681102.2013.814040 arXiv:https://doi.org/10.1080/02681102.2013.814040

[29] Christina Katsini, Yasmeen Abdrabou, George Raptis, Mohamed Khamis, and Florian Alt. 2020. The Role of Eye Gaze in Security and Privacy Applications: Survey and Future HCI Research Directions.. In *Proceedings of the 38th Annual ACM Conference on Human Factors in Computing Systems* (Honolulu, Hawaii, USA) *(CHI '20).* ACM, New York, NY, USA, 21 pages. https://doi.org/10.1145/3313831.3376840

[30] Christina Katsini, Christos Fidas, George E. Raptis, Marios Belk, George Samaras, and Nikolaos Avouris. 2018. Influences of Human Cognition and Visual Behavior on Password Strength during Picture Password Composition. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18).* Association for Computing Machinery, New York, NY, USA, 1–14. https://doi.org/10.1145/3173574.3173661

[31] Christina Katsini, George E. Raptis, Christos Fidas, and Nikolaos Avouris. 2018. Towards Gaze-Based Quantification of the Security of Graphical Authentication Schemes. In *Proceedings of the 2018 ACM Symposium on Eye Tracking Research & Applications* (Warsaw, Poland) *(ETRA '18).* Association for Computing Machinery, New York, NY, USA, Article 17, 5 pages. https://doi.org/10.1145/3204493.3204589

[32] Mohamed Khamis, Florian Alt, and Andreas Bulling. 2018. The Past, Present, and Future of Gaze-Enabled Handheld Mobile Devices: Survey and Lessons Learned. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services* (Barcelona, Spain) *(MobileHCI '18).* Association for Computing Machinery, New York, NY, USA, Article 38, 17 pages. https://doi.org/10.1145/3229434.3229452

[33] Nina Kolb, Steffen Bartsch, Melanie Volkamer, and Joachim Vogt. 2014. Capturing attention for warnings about insecure password fields–systematic development of a passive security intervention. In *International Conference on Human Aspects of Information Security, Privacy, and Trust.* Springer, 172–182.

[34] Scott Lundberg and Su-In Lee. 2017. A unified approach to interpreting model predictions. *arXiv preprint arXiv:1705.07874* (2017).

[35] Yoshitomo Matsubara, Toshiharu Samura, Haruhiko Nishimura, et al. 2015. Keyboard dependency of personal identification performance by keystroke dynamics in free text typing. *Journal of Information Security* 6, 03 (2015), 229.

[36] Fabian Monrose and Aviel D Rubin. 2000. Keystroke dynamics as a biometric for authentication. *Future Generation computer systems* 16, 4 (2000), 351–359.

[37] Robert Moskovitch, Clint Feher, Arik Messerman, Niklas Kirschnick, Tarik Mustafic, Ahmet Camtepe, Bernhard Löhlein, Ulrich Heister, Sebastian Möller, Lior Rokach, and Yuval Elovici. 2009. Identity Theft, Computers and Behavioral Biometrics. In *Proceedings of the 2009 IEEE International Conference on Intelligence and Security Informatics* (Richardson, Texas, USA) *(ISI'09).* IEEE Press, 155–160.

[38] Sarah Pearman, Jeremy Thomas, Pardis Emami Naeini, Hana Habib, Lujo Bauer, Nicolas Christin, Lorrie Faith Cranor, Serge Egelman, and Alain Forget. 2017. Let's Go in for a Closer Look: Observing Passwords in Their Natural Habitat. In *Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security* (Dallas, Texas, USA) *(CCS '17).* Association for Computing Machinery, New York, NY, USA, 295–310. https://doi.org/10.1145/3133956.3133973

[39] Sarah Pearman, Shikun Aerin Zhang, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2019. Why People (Don't) Use Password Managers Effectively. In *Proceedings of the Fifteenth USENIX Conference on Usable Privacy and Security* (Santa Clara, CA, USA) *(SOUPS'19).* USENIX Association, USA, 319–338.

[40] Ken Pfeuffer, Yasmeen Abdrabou, Augusto Esteves, Radiah Rivu, Yomna Abdelrahman, Stefanie Meitner, Amr Saadi, and Florian Alt. 2021. ARtention: A design space for gaze-adaptive user interfaces in augmented reality. *Computers & Graphics* 95 (2021), 1–12.

[41] George E. Raptis, Christina Katsini, Marios Belk, Christos Fidas, George Samaras, and Nikolaos Avouris. 2017. Using Eye Gaze Data and Visual Activities to Infer Human Cognitive Styles: Method and Feasibility Studies. In *Proceedings of the 25th Conference on User Modeling, Adaptation and Personalization* (Bratislava, Slovakia) *(UMAP '17).* Association for Computing Machinery, New York, NY, USA, 164–173. https://doi.org/10.1145/3079628.3079690

[42] Karen Renaud, Verena Zimmerman, Joseph Maguire, and Steve Draper. 2017. Lessons learned from evaluating eight password nudges in the wild. In *The {LASER} Workshop: Learning from Authoritative Security Experiment Results ({LASER} 2017)*. 25–37.

[43] Toshiharu Samura and Haruhiko Nishimura. 2012. Influence of Keyboard Difference on Personal Identification by Keystroke Dynamics in Japanese Free Text Typing. In *2012 Fifth International Conference on Emerging Trends in Engineering and Technology*. 30–35. https://doi.org/10.1109/ICETET.2012.24

[44] Tobias Seitz, Manuel Hartmann, Jakob Pfab, and Samuel Souque. 2017. Do Differences in Password Policies Prevent Password Reuse?. In *Proceedings of the 2017 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI EA '17)*. Association for Computing Machinery, New York, NY, USA, 2056–2063. https://doi.org/10.1145/3027063.3053100

[45] Richard Shay, Saranga Komanduri, Adam L. Durity, Phillip (Seyoung) Huh, Michelle L. Mazurek, Sean M. Segreti, Blase Ur, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2016. Designing Password Policies for Strength and Usability. *ACM Trans. Inf. Syst. Secur.* 18, 4, Article 13 (may 2016), 34 pages. https://doi.org/10.1145/2891411

[46] Richard Shay, Saranga Komanduri, Patrick Gage Kelley, Pedro Giovanni Leon, Michelle L. Mazurek, Lujo Bauer, Nicolas Christin, and Lorrie Faith Cranor. 2010. Encountering Stronger Password Requirements: User Attitudes and Behaviors. In *Proceedings of the Sixth Symposium on Usable Privacy and Security* (Redmond, Washington, USA) *(SOUPS '10)*. Association for Computing Machinery, New York, NY, USA, Article 2, 20 pages. https://doi.org/10.1145/1837110.1837113

[47] Julian Steil, Marion Koelle, Wilko Heuten, Susanne Boll, and Andreas Bulling. 2019. PrivacEye: Privacy-Preserving Head-Mounted Eye Tracking Using Egocentric Scene Image and Eye Movement Features. In *Proceedings of the 11th ACM Symposium on Eye Tracking Research & Applications* (Denver, Colorado) *(ETRA '19)*. Association for Computing Machinery, New York, NY, USA, Article 26, 10 pages. https://doi.org/10.1145/3314111.3319913

[48] Charles C Tappert, Mary Villani, and Sung-Hyuk Cha. 2010. Keystroke biometric identification and authentication on long-text input. In *Behavioral biometrics for human identification: Intelligent applications*. IGI global, 342–367.

[49] Kurt Thomas, Jennifer Pullman, Kevin Yeo, Ananth Raghunathan, Patrick Gage Kelley, Luca Invernizzi, Borbala Benko, Tadek Pietraszek, Sarvar Patel, Dan Boneh, et al. 2019. Protecting accounts from credential stuffing with password breach alerting. In *28th {USENIX} Security Symposium ({USENIX} Security 19)*. 1556–1571.

[50] D. Vitonis and D. W. Hansen. 2014. Person Identification Using Eye Movements and Post Saccadic Oscillations. In *2014 Tenth International Conference on Signal-Image Technology and Internet-Based Systems*. 580–583.

[51] Emanuel von Zezschwitz, Alexander De Luca, and Heinrich Hussmann. 2013. Survival of the Shortest: A Retrospective Analysis of Influencing Factors on Password Composition. In *Human-Computer Interaction – INTERACT 2013*, Paula Kotzé, Gary Marsden, Gitte Lindgaard, Janet Wesson, and Marco Winckler (Eds.). Springer Berlin Heidelberg, Berlin, Heidelberg, 460–467.

[52] Rick Wash, Emilee Rader, Ruthie Berman, and Zac Wellmer. 2016. Understanding Password Choices: How Frequently Entered Passwords Are Re-used across Websites. In *Twelfth Symposium on Usable Privacy and Security (SOUPS 2016)*. USENIX Association, Denver, CO, 175–188. https://www.usenix.org/conference/soups2016/technical-sessions/presentation/wash

[53] Yongtuo Zhang, Wen Hu, Weitao Xu, Chun Tung Chou, and Jiankun Hu. 2018. Continuous Authentication Using Eye Movement Response of Implicit Visual Stimuli. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.* 1, 4, Article 177 (Jan. 2018), 22 pages. https://doi.org/10.1145/3161410