Fingerprinting and Personal Information Leakage from Touchscreen Interactions

Martin Georgiev
University of Oxford
Oxford, UK
martin.georgiev@cs.ox.ac.uk

Simon Eberz
University of Oxford
Oxford, UK
simon.eberz@cs.ox.ac.uk

Ivan Martinovic
University of Oxford
Oxford, UK
ivan.martinovic@cs.ox.ac.uk

ABSTRACT

The study aims to understand and quantify the privacy threat landscape of touch-based biometrics. Touch interactions from mobile devices are ubiquitous and do not require additional permissions to collect. Two privacy threats were examined - user tracking and personal information leakage. First, we designed a practical fingerprinting simulation experiment and executed it on a large publicly available touch interactions dataset. We found that touch-based strokes can be used to fingerprint users with high accuracy and performance can be further increased by adding only a single extra feature. The system can distinguish between new and returning users with up to 75% accuracy and match a new session to the user it originated from with up to 74% accuracy. In the second part of the study, we investigated the possibility of predicting personal information attributes through the use of touch interaction behavior. The attributes we investigated were age, gender, dominant hand, country of origin, height, and weight. We found that our model can predict the age group and gender of users with up to 66% and 62% accuracy respectively. Finally, we discuss countermeasures, limitations and provide suggestions for future work in the field.

1 INTRODUCTION

Smartphones have become an integral part of our daily lives. Both mobile websites and dedicated applications are used for a variety of tasks, including entertainment, banking, and shopping. The prevailing method for interacting with smartphones is by using their touchscreen displays. Touch interactions, such as swiping and scrolling on mobile devices are necessary to navigate websites and applications, and as such, data related to our touch behavior can be collected trivially in both settings. Furthermore, it is possible to achieve that without the explicit permission or knowledge of the user. Touch-based interactions have been actively studied for their potential in enhancing the security of users. One such application of the technology is continuous authentication [15, 23].

However, as the technology matures, touch-based biometrics have the potential to become a privacy threat that can be used with malicious intentions. In this study, we explore how touchscreen interactions data can be used to invade the privacy of users without their explicit knowledge. To this end, we make use of a two-fold approach. First, we investigate the feasibility of fingerprinting users based on the way they interact with their mobile devices. Fingerprinting, also known as stateless tracking, can be a major concern for users as it can be used for a number of malicious purposes, such as discrimination and surveillance [8]. However, it can also be beneficial in cases such as personalizing the user experience in mobile websites and applications. In the second part of the study, we examine the possibility to extract personal information from mobile users through the use of touch-based interactions. By personal information, we refer to the physical and intrinsic characteristics of humans. We evaluate the potential to reveal the following six attributes - age, gender, dominant hand, country of origin, height, and weight.

We perform a series of experiments on a large publicly available touch interactions dataset to quantify the feasibility of the two privacy threats and give directions for future work. In this paper we make the following contributions:

- Investigate a practical user fingerprinting approach using touch-based interactions. Introduce a realistic evaluation method and test two different approaches for discrimination of new and returning users and re-identification of returning users.
- Explore the feasibility of touchscreen interaction models to reveal the age, gender, dominant hand, country of origin, height, and weight of users. We test our results using three different data processing approaches and three machine learning algorithms.
- We discuss our findings, countermeasures, and avenues for future research in the area.

2 RELATED WORK

Touchscreen interactions have been studied in the context of continuous authentication since the early 2010s. Several papers survey the development of touch-based authentication [31, 33] and investigate the approaches used in the field [16, 32]. However, studying the privacy implications of touch-based biometrics has been limited. Our work examines and quantifies the practical touch-based biometric implications on privacy with a focus on fingerprinting mobile users and revealing their personal information using touchscreen interactions.

2.1 Desktop and mobile fingerprinting

Tracking user behavior through fingerprinting can be used for a variety of purposes - some commercial (e.g. online advertisement, user-oriented search) and others much more harmful (e.g. price discrimination, government surveillance). The majority of scientific work on the topic has been done in the context of desktop web interfaces. The field is mature and many detailed surveys have explored its development [8, 21].

More recently, with the rapid adoption of smartphones, new methods for fingerprinting have been proposed that focus entirely on mobile devices. Hupperich et al. [18] conducted a comprehensive and large-scale study of 900 participants on fingerprinting of mobile devices using 45 features such as user-agent, operating system,

screen height, and width. Similarly, Kurtz et al. [20] used personalized configurations (e.g. device language, installed apps) of over 8,000 mobile devices to fingerprint them, achieving an accuracy of 97%.

The built-in sensors of mobile phones have been proposed as another method to fingerprint users and their devices. One of the earliest works by Bojinov et al. [6] recorded speaker-microphone and accelerometer sensor data from 10,000 mobile devices. They showed that it is possible to uniquely identify a device among thousands with a low chance for collision. Several studies have used motion sensors such as accelerometers and gyroscopes for mobile fingerprinting in various conditions [3, 35, 37]. Das et al. [12, 13] give practical advice for countermeasures concerning mobile motion sensor attacks. Unconventional sensors such as magnetometers have also been shown to work for device fingerprinting [26].

Masood et al. [25] investigated the uniqueness of touchscreen interactions, thus showing that touch gesture features carry highly identifiable information about users and have a potential for fingerprinting. The study has similar goals to the first part of our paper, however, we employ different evaluation methods with a focus on practicality and understanding.

2.2 Personal information leakage

Personally identifiable information such as age, gender, and height of mobile users can be revealed by a variety of side-channel methods. Malmi et al. [24] used the apps installed on a device to predict the gender, age, race, children count, marriage status, and income of mobile users. They achieved results between 60% and 80% accuracy in these categories. Frias-Martinez et al. focused on gender classification using calling patterns history and achieved 80% accuracy. The way people type on the smartphone (keystrokes) has also been found to reveal age, gender, and operating hand with accuracies varying between 82% and 95% [9]. Other studies examine models which take advantage of the general patterns of usage in applications, browsers, WiFi and Bluetooth [28, 29]. Such features can reveal a multitude of personal information attributes such as age, gender, education level, marital status, employment, and others with high accuracy.

Predicting personal information using touch interactions has also received attention from researchers in the field. Bevan et al. [5] conducted a study of 178 users and showed that the differences in some features of strokes can reveal information about the handedness, thumb length, and gender of users. Antal et al. [4] created models to predict the gender and phone experience level of users. They report high accuracy of between 88-100% for gender and 80%-100% for phone experience requiring up to 20 strokes to make a decision. Miguel-Hurtado et al. [27] focused exclusively on gender prediction achieving up to 78% accuracy. Similarly, Jain et al. [19] focused specifically on gender prediction improving on previous work and achieving ~93% accuracy. Acien et al. [2], Nguyen et al. [30] and Cheng et al. [10] used touch interaction data of children and adults to differentiate between the two groups. All three studies report accuracies of above 96%.

Davis et al. [14] used swipe data to predict both the gender of users and their age. In the gender predicting scenario, they report a more modest 70% accuracy average across the different classifiers used. Predicting age groups above or below 40 years resulted in 80% accuracy. More recently B. Wiliams [34] examined the feasibility of predicting location (the state within the United States), gender, race, and education level with 46%, 73.3%, 73.3%, and 26.7% accuracy respectively.

A variety of machine learning models have been used in these studies, including Support Vector Machine (SVM), Neural Network (NN), Random Forest (RF), Naive Bayes, Decision Tree, Logistic Regression, Nearest Neighbor, and others.

We differentiate the second part of our study from previous work by conducting a comprehensive evaluation of 6 personal traits (age, gender, handedness, height, weight, and location) using a constrained (9 iOS models) but large open-source dataset in the context of privacy threats in touch-based biometrics.

3 DATASET AND FEATURES

In order to evaluate the privacy concerns stemming from the use of touch interactions, we conducted our experiments on a large open-source touch-based dataset [17]. The dataset consists of 470 users with up to 31 and an average of 13 sessions per user. The participants were required to perform two sessions daily, consisting of tasks aiming to mimic natural scrolling and swiping behavior with each session lasting slightly less than 2 minutes on average. The social media task required users to scroll up and down through a randomly generated feed of posts and find an article related to a particular question. The image gallery task required users to swipe left and right through a series of images and count the number of specified objects. The data is limited to 9 iOS device models and there are a total of 6,006 usable sessions for each of the tasks. The dataset has been collected over a relatively long period of time for such study, meaning that the effects on the stability of the fingerprinting and personal information leakage methods are taken into account.

For our experiments, we used the set of 149 features introduced by Georgiev et al. [16]. These are geometric features describing the properties of a particular touchscreen gesture done by the user. The features are extracted from a series of points consisting of X, Y, pressure, and area values which describe a complete swipe. Since the authors found that this set of features works best for authentication, we hypothesized that it would also perform well on our tasks.

While we complete our investigation on an app-based dataset, we believe the findings in our study are relevant to websites as well because the recording of touchscreen strokes is also feasible on a mobile browser. For instance, it is possible to collect touch interaction data using JavaScript and the TouchEvents API [1]. This API is available on all modern mobile web browsers and extra permissions from the user are not required to access it. We do however acknowledge that there might be some limitations in place. For example, the API is clear that it cannot guarantee a specific touch sampling rate - *"The rate at which touchmove events are sent is browser-specific, and may also vary depending on the capability of the user's hardware. You must not rely on a specific granularity of these events."*. Furthermore, in practice, there might be differences in the reported values for pressure and X, Y coordinates between APIs.

[1] https://developer.mozilla.org/en-US/docs/Web/API/Touch_events

Despite that, we believe that our findings are relevant for mobile web browsing as well. Similarly, our experiments are done on an iOS-exclusive dataset, however, there is no reason to believe there would be major differences with the Android operating system as the same methods for data collection are available.

4 FINGERPRINTING

Touch-based interactions can be used for tracking users on web and mobile applications due to the unique differences in the way people interact with their touchscreens as illustrated by continuous authentication research. In this section, we investigate the feasibility of tracking users by the way they interact with their smartphone screens. Unlike fingerprinting with sensors such as accelerometer, gyroscope, and magnetometer, touch-screen-based tracking is invisible to the user and does not require extra permission requests. We briefly describe the evaluation methods used in fingerprinting systems, formalize our approach and present our findings.

4.1 Evaluation approaches

We categorize and briefly describe each of the evaluation methods for fingerprinting systems. Furthermore, we clarify why some approaches are preferable for touch-based tracking.

Identification / Multi-class classification. This approach assumes a closed set of users where a model is trained to predict the class a session belongs to. A class in this case is a user or a device. Whenever a new session is performed, the model can be used to match it to the correct class. This method however assumes that we have full knowledge about the number of users in the system and hence the classes a session can belong to. In other words, the evaluation does not take into account that a new session could be coming from a user that has never used the system before. This becomes an identification challenge, which is closer to authentication in nature than fingerprinting. However, the method can still be applicable in some narrow cases. For instance, it can be useful in the detection of multiple user accounts in a closed set of registered users. Das et al. [12] use this approach to evaluate their web tracking paper based on mobile motion sensors.

Entropy and anonymity set. The use of entropy as a measure of identifying information in a fingerprint has been widely adopted in the field, particularly in desktop settings [22, 36]. However, the notion of Shanon Entropy loses its usefulness when making decisions based on similarity rather than equality which is the case in fingerprinting systems using continuous feature values (e.g. motion sensors and touchscreen interactions). In other words, the minimum number of bits required to distinguish a user is not an important metric in our investigation, as nearly all of our sessions will be unique. However, a fingerprinting system needs to also recognize when sessions are performed by the same user. Similarly, the anonymity set, which describes the size of groups with identical fingerprints, is also irrelevant in this case as there are practically no anonymity sets larger than one. It is possible to use bins to create categorical data from continuous in order to use this evaluation method. However, losing the granularity of the data might lead to poorer performance. That is the approach taken by Masood et al. [25].

Simulation. In this approach, a simulation of user sessions visiting a website (or application) is modeled. The evaluation is performed in two steps - discrimination and re-identification. In the first step (discrimination), the system decides whether a session is coming from a new user who has never visited the website or a returning one that the system has observed already. The next step (re-identification) is performed only when the user is a returning one. At this stage, the system matches the session with the correct existing user. Both steps are typically done by measuring the distance between the features of the new session and the existing ones on record. Simulation fingerprinting systems have been used for motion sensor mobile fingerprinting by Hupperich et al. [18]. While this approach is suitable for continuous data, it can also be used for categorical data. For instance, Kurtz et al. [20] use the Jaccard similarity coefficient to measure the distance between sessions on categorical data.

4.2 Formalizing our approach

In order to evaluate the fingerprinting system based on touch interactions, we chose to use the simulation approach. We believe a simulation fits the continuous nature of our data and provides a better examination of how the technology can be used in practice. An example of how this technology can be applied is a web store where a number of users have placed an order in the past. The users consent to being fingerprinted and do not register an account. The goal of the system then is to identify returning users in order to upsell or advertise based on previous behavior. In this case, the discrimination step is in place to avoid treating new users as returning ones and the re-identification step to avoid advertising to the wrong person. Although we use this as an example of how the technology can be used, the system has much wider applicability, both for benevolent and malicious purposes. We formally define our simulation as follows:

There is a set of users $G = \{U_1, U_2 ... U_n\}, n \in \mathbb{N}$. Each user in turn is a set of sessions $U_n = \{S_1, S_2 ... S_n\}, n \in \mathbb{N}$ and each session is a set of features $S_n = \{F_1, F_2 ... F_n\}, n, i \in \mathbb{N}$. For instance, S_n^i can be the set of features described in Section 3.

The goal of our system is, then, to correctly classify a new session S_u which either belongs to an existing user U_n or a new user U_{n+1}. In order to make this decision, we use a dissimilarity function $D(S_1, S_2)$ where S_1 and S_2 are generic feature sets of the same size. In other words, D measures how much two vectors differ from each other. Many functions can be used as a dissimilarity measure, including Euclidean distance, Manhattan distance, and machine learning algorithms such as Support Vector Machines and Neural Networks. Using the dissimilarity function D, we define d_{min} to be the distance to the session most similar to S_u.

$$d_{min} = _{S \in U, U \in G} \; D(S, S_u)$$

We then define a threshold δ and if $d_{min} \leq \delta$ we classify S_u as belonging to a new user U_{n+1}. Else if $d_{min} > \delta$ we classify the session as a returning one and mark the session as belonging to an existing user $U_i \in G$ which contains the session closest to S_u. Correct classification of a new session S_u is then either of the following:

- S_u is performed by a new user U_{n+1} and we classify it as such
- S_u is performed by a returning user and we match it to the correct user $U_i \in G$

It is worth noting that we do not update the sets G and U_n while running this simulation. This is because we carefully balance the training and testing sets in our experiments for fair evaluation and updates to these sets will make results difficult to interpret. However, in practice, adding new users and sessions to the system is technically trivial.

4.3 Method

In order to evaluate the performance of the proposed fingerprinting approach, we executed the simulation described in the previous section on both the image gallery and social media tasks provided in our dataset. We investigated the performance of our system on the following three tasks:

- **Discrimination** - In this case, we choose a threshold that optimizes the accuracy of the discrimination model which decides whether a session is coming from a new or returning user. This is a binary problem with a baseline accuracy of 50%.
- **Re-identification** - This task measures the accuracy of the model to match new sessions to users already known by the system. In this case, we only consider the returning sessions, hence the threshold here is not relevant. The baseline accuracy of this model is $1/n$, where n is the number of users who are already known by the system.
- **Combined** - This task evaluates the correct classification of the system exactly as described in the simulation. In this case, both discrimination and re-identification steps need to be correct to mark a decision as accurate. The baseline accuracy is 50% since we can mark all sessions as new which is exactly half of our testing set. However, this baseline can be misleading as it does not measure the real purpose of the system which also includes matching returning sessions to the original user.

In order to evaluate the system correctly, we split the session data such that there is an equal number of sessions coming from new and returning users. That is done by selecting n number of users and setting aside half of their sessions for training and the other half as returning sessions for testing. The remaining users' sessions are labeled as new and also used for testing. Users who have performed only a single session are always treated as a new user because they do not have a second session that can be used for testing. There are a total of 61 users with only one session. The parameter n is selected such that the amount of returning and new sessions used for testing the system is as close as possible. Then sessions are dropped to ensure the sets are exactly equal in size.

In our experiments, we started with the set of 149 features for each stroke as described in Section 3. Each value of the final feature vector is acquired by calculating the mean of each stroke feature throughout a whole session. In addition, we included an extra feature - the total number of swipes in a session, which results in a total of 150 features for each session. Then we applied a feature selection algorithms to reduce the overall dimensionality of our data and achieve better computational performance and results. We applied the Analysis of Variance (ANOVA) feature selection algorithm using the F-value between features and labels. The label in this case is the user id of the participant performing the session. We fit the algorithm only on training users and transform both the training and testing sets. We conducted preliminary experiments with varying the number of features k and found that the best-performing k for our purposes was 100.

We also performed a separate experiment where the phone model of the device performing the session (e.g. iPhone 7) is known and we use it as an extra feature in our analysis. This is a realistic scenario as the phone model of a device is easily accessible by standard mobile application and web APIs.

For the dissimilarity function, we experimented with two approaches - one based on a vector distance metric and one based on a machine learning algorithm. The first approach uses the cosine distance (D_c) which is $1 - S_c$ where S_c is the cosine similarity. It belongs to the interval $[0, 2]$. Given two feature vectors S_1, S_2, the cosine distance is defined as follows:

$$D_c = 1 - S_c = 1 - \frac{S_1 \cdot S_2}{||S_1|| \cdot ||S_2||}$$

The second approach we tested is based on a machine learning algorithm rather than a vector distance function. First, we trained an individual SVM classifier for each of the users in the training set. The positive class consists of the sessions of a particular user and the negative class consists of sessions from the other users in the simulation training set. We keep the positive and negative classes balanced. These are small classifiers as there are typically only a few positive examples per user. In the discrimination scenario, whenever we want to classify a session as new or returning, we make a prediction from each of the classes and record the maximum distance to the SVM hyperplane. If it is above the threshold (δ) we mark it as a returning session, otherwise, as a new user session. Similarly, in the re-identification scenario, we classify the session to belong to the user with the SVM model producing the largest distance to the hyperplane. This is in contrast to the vector distance approach, where we choose the user with the lowest distance value. That is because the distance to the hyperplane represents how confident the model is about a prediction. The larger it is, the more likely the session is originating from the user the model has been trained on.

In practice, when a new user visits a website, it is unrealistic for the system to wait until the whole data for a particular session is available to make a decision. It should be possible to do that with a fraction of the data available. For this reason, we also repeated our experiments with a portion of the whole session data. We tested the system by only using the first 10% to 90% of the session data at 10 percentage point intervals.

For each of our experiments described in this section, we executed the simulation a total of 10 times where each time the group of n users is randomized.

4.4 Results

The results for the discrimination, re-identification, and combined scenarios on both image gallery and social media tasks are shown

Table 1: Performance of fingerprinting users with all strokes from a session. The cosine distance and SVM approaches are compared. The discrimination model differentiates between a new or returning user, the re-identification model matches a session to the user it originated from and the combined model fulfills both conditions. The image gallery task predominantly consists of swiping (left/right) behavior and the social media task mainly consists of scrolling (up/down) behavior. The threshold (δ) is given in parentheses where relevant.

		Image Gallery Task			Social Media Task		
		Accuracy (δ)	Phone Model Feature (δ)	Baseline	Accuracy (δ)	Phone Model Feature (δ)	Baseline
Distance	Discrimination	60.6% (0.25)	64.2% (0.30)	50%	59.7% (0.28)	65.8% (0.34)	50%
	Re-identification	40.8% (-)	53.2% (-)	0.36%	41.9% (-)	59.2% (-)	0.36%
	Combined	54.0% (0.17)	58.1% (0.24)	50%*	53.4% (0.19)	60.2% (0.28)	50%*
SVM	Discrimination	69.3% (0.85)	72.3% (0.77)	50%	69.3% (0.84)	75.0% (0.72)	50%
	Re-identification	58.3% (-)	68.4% (-)	0.36%	61.5% (-)	74.0% (-)	0.36%
	Combined	62.7% (0.94)	67.3% (0.84)	50%*	64.2% (0.90)	71.4% (0.78)	50%*

*Can be achieved by predicting all testing sessions as coming from a new user, however that does not represent the real purpose of the system.

in Table 1. In addition, we highlighted the baseline performance of each scenario. The accuracy of our system with the additional phone model feature is also given and the distance and machine learning-based approaches are compared.

Overall, the SVM-based distance approach performed significantly better than the cosine distance measure with 62.7% and 64.2% accuracy compared to 54.0% and 53.4% for the two tasks in the combined scenario without the phone model feature. We discuss the results on the SVM approach for the rest of this section.

The touch-based fingerprinting method performs well in the re-identification scenario where returning sessions are matched to their original users. It achieves 58.3% and 61.5% accuracy on the image gallery and social media tasks respectively. That is significantly higher than the baseline of 0.36%. The good performance of touch-based interactions for re-identification purposes is supported by the research in continuous touch-based authentication which reports strong results in tasks with similar goals. In the discrimination step, where a session is labeled as originating from a new or returning user, the fingerprinting model also performs well. It achieves an accuracy of 69.3% on both tasks compared to the baseline of 50%. The discrimination scenario represents an upper bound to the performance of the combined scenario where we achieved 62.7% and 64.2% accuracy for the image gallery and social media tasks respectively.

The performance of our system on both image gallery and social media tasks is equal in the discrimination case and differentiates only by 3.2% and 1.5% for the re-identification and combined scenarios respectively. The image gallery task results are slightly better in each case. This similarity in performance is encouraging for the applicability of the system for a variety of purposes and settings both on websites and mobile applications.

Including the phone model of the device as a single extra feature in our system resulted in better accuracy in all tasks. The increase in performance is by 5.7%, 12.5%, and 7.2% for the discrimination, re-identification, and combined scenarios in the social media task respectively. The increase is similar (3%, 10.1%, and 4.6%) for the social media task. This is encouraging for the practical use of the system as it suggests that including a few conventional fingerprinting features in conjunction with the touch-based approach can result in a highly effective system.

Due to space constraints, we present the rest of the results in this section for the image gallery task only and the social media task results are shown in Appendix A.

The results of varying the session size used in our experiments are shown in Figure 1. The accuracy of the model in all scenarios is lower when less data is used. However, the increase in performance when more data is used is marginal, particularly with session sizes above ~30%. This means that users can be fingerprinted during the early stages of their session without sacrificing much accuracy.

We show the performance of the discrimination and combined scenarios at varying thresholds (δ) at three session sizes (10%, 30% and 100%) in Figure 2 and Figure 3 respectively. These figures highlight the importance of selecting the correct threshold to achieve optimal performance. In practice, selecting a well-performing threshold is a difficult task as access to the testing data is not available. It is possible to make approximate decisions about the threshold using training data only. Another possible solution is to use well-performing thresholds in similar domains, considering that the social media and image gallery tasks have comparable threshold values. However, investigating the optimal threshold selection method is beyond the scope of this study.

We believe this practical investigation of fingerprinting using touch interactions is an important step in uncovering the privacy

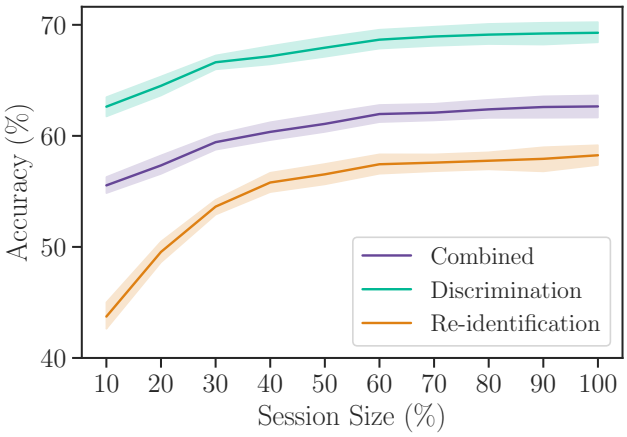


Figure 1: Performance of the fingerprinting simulation when using a fraction of the whole session data. The results are on the image gallery task without the additional phone model feature and using the SVM fingerprinting approach. The shaded area represents a 95% confidence interval.

implications of touch-based biometrics. The bottleneck of our models is in the discrimination scenario where user sessions are marked as new or returning. Further improvements, such as more sophisticated machine learning models or dynamic threshold selection might result in much better performance. The touch interaction distance we described in this study can be used in conjunction with other, easily accessible features to achieve a more complete fingerprinting system. For instance, Bojinov et al. [6] achieve 8% accuracy using their fingerprinting method but the performance increases to 53% by including only a single extra feature to the model - the user agent.

5 PERSONAL INFORMATION LEAKAGE

The way individuals interact with touchscreen devices has been extensively studied for authentication purposes. Similarly, we believe that there are variations in the way groups of people behave on touchscreen devices. For instance, left-handed people hold the phone slightly differently from the way right-handed people do. This could result in scrolls occurring on one side of the screen more often than on the other. A machine learning algorithm can establish these differences and predict whether a session belongs to users from one group or another. For this reason, we conducted a series of experiments to predict personal information attributes based on these unique interaction characteristics.

Effective personal information leakage systems can lead to major issues in terms of censorship, tracking, and discrimination of people. However, in some cases, they might have beneficial uses such as restricting certain content from children or improving the mobile user experience. For example, one positive use case of this technology is to request additional age verification to visitors of a gambling website where the touch-based model predicts that the user is younger than 18 years with a high probability. We decided to explore a number of different personal information attributes that might be revealed from the way users interact with their touchscreens - age, gender, dominant hand, country of origin, height, and weight. Furthermore, we evaluated a number of data selection methods and machine learning classifiers to establish the best-performing ones for this purpose.

Table 2: Personal information attributes considered for touch interactions inference. The binary classes and the number of users in each class are given. The number of features used in the image gallery task models is shown on the left and the social media task models on the right.

Attribute	Class 0	Class 1	# Features
Gender	Male (187)	Female (146)	130 | 130
Handedness	Right (295)	Left (38)	30 | 12
Country	USA (234)	India (19)	100 | 10
Age	$\leq$ 25 years (73)	$\geq$ 45 years (57)	10 | 35
Height	$\leq$ 159cm (30)	$\geq$ 183cm (40)	27 | 50
Weight	$\leq$ 50kg (57)	$\geq$ 91kg (43)	10 | 18

5.1 Method

The dataset we used to evaluate the personal information leakage potential has been collected remotely and the demographic information shared by participants has been self-reported. Due to this, we performed a pre-processing step to clean the data from outliers. Firstly, we removed all users which have reported an unreasonable height (less than 100cm or above 250cm), weight (less than 20kg and above 250kg), and age (less than 18 years and above 90 years). Users below 18 years of age are not allowed on the platform used for remote collection. Furthermore, for the weight and height attributes, we only considered users within 2 standard deviations of the average. In terms of gender, we only investigated people identifying as males or females and removed the other 4 users. Although it is certainly possible for some of the users we pruned to have reported true values, we decided to minimize the risk of polluting the prediction data. This pre-processing step reduced the total number of users we investigate in this section from 470 to 333.

For ease of comparison and a better understanding of the potential for leakage of personal information from touch interactions, we treated each of the predictions as a binary classification problem. Using the data we have, predicting the gender and dominant hand of participants is already a binary classification problem. However, for the country of origin attribute, we divided the dataset into participants originating from the USA and India which were the first and second largest country groups respectively. Furthermore, we converted the continuous age, weight, and height attributes into binary classes. In order to achieve this split (e.g. younger and older users) we separated the classes into groups above and below one standard deviation of the mean in each category. The binary classes for each of the six personal information attributes are shown in

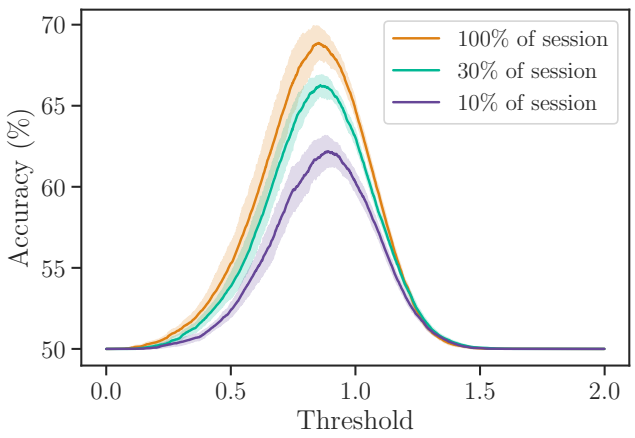


Figure 2: Discrimination model accuracy line plot at varying thresholds (δ). Results are shown on the image gallery task and presented with session sizes of 10%, 30%, and 100%. The SVM fingerprinting approach without the additional phone model is used. The shaded area represents a 95% confidence interval.

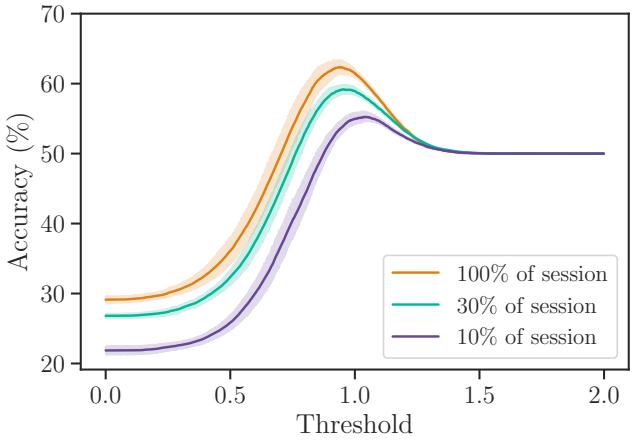


Figure 3: Combined model accuracy line plot at varying thresholds (δ). Results are shown on the image gallery task and presented with session sizes of 10%, 30%, and 100%. The SVM fingerprinting approach without the additional phone model is used. The shaded area represents a 95% confidence interval.

Table 2. In addition, we show the number of users in each of the categories.

Since some of the 150 features we have extracted are likely not relevant for each of the attributes examined, we applied a feature selection algorithm to reduce the dimensionality of our data. This approach ensures better computational performance but can also improve the overall accuracy of the model by removing irrelevant features. Similar to the fingerprinting scenario, we used the ANOVA feature selection algorithm. After some preliminary experiments, we chose different values for the number of k features selected in each attribute prediction scenario. These are shown in Table 2 for both social media and image gallery tasks. We fit the feature selection algorithm on training data only. It is worth noting that in the privacy leakage experiments we do not use the additional phone model feature mentioned in Section 4.3. However, it has been shown that phone models can be identified by the touch-based interactions performed on them [17]. In that sense, part of that phone model data is intrinsic to the other features in the dataset.

In order to fairly compare the performance of our models, we balance the two attribute classes such that there is an equal number of users in both. The users are split into 80% training and 20% testing groups where again we ensure the two classes are equally balanced. We only include user sessions in the testing set if sessions from the same user have not been used for training. This is critical since the model can learn the identities of users and correctly identify the particular group of users (e.g. males in our dataset) a session belongs to instead of learning the attribute in question (e.g. gender). We also normalize the features by subtracting the mean and dividing by the standard deviation of the training data.

We investigated three data selection approaches for predicting personal information based on touch interactions:

- **Single-stroke** - The personal information predictions are made using the features from a single stroke. This is the most challenging scenario where the least amount of information is available and there is likely a large deviation in prediction performance across individual strokes. However, that is also the fastest way to make a decision about personal information attributes during a session.
- **Multi-stroke** - In this scenario, we use multiple single-stroke predictions to make a final decision. Similar to an aggregation step in touch-based authentication systems [16], we use the mean of the individual stroke predictions to form a prediction. In these experiments, we use 10 consecutive strokes to make a decision. It takes an average of 7 seconds of swiping to collect this number of strokes and thus make a prediction.
- **Session** - This approach is similar to the fingerprinting feature extraction. We use the data from a whole session in order to reach a conclusion and the features are averaged across the whole session as described in Section 4.3. Naturally, this is the method that requires the most amount of time before a final decision can be made. It is possible to use a portion of the session to shorten the time needed, however, we do not examine this scenario as it will likely reduce the accuracy as is the case with fingerprinting.

Furthermore, we examined the performance of our models using a number of machine learning classifiers - Support Vector Machine, Random Forest, and Neural Network. The three classifiers were chosen as they are the best performing for touch-based authentication [16] and are commonly found in related work [4, 10, 14, 30, 34]:

- **Support Vector Machine (SVM)** - We use an SVM with an RBF kernel with a 'scale' coefficient and probability estimations enabled.

- **Random Forest (RF)** - Our implementation uses 100 estimators and has a maximum depth of 20. Probability estimations are enabled.
- **Neural Network (NN)** - The feed-forward network we use consists of two layers of sizes 150 and 75 respectively. We use a 'ReLU' activation function for the hidden layers and a 'Sigmoid' activation function for the output which predicts probability between 0 and 1 for each of the binary outputs. There is batch-normalization at each layer and a dropout (0.3) between the hidden layers. The optimizer is 'adam' and the loss function is a 'binary cross-entropy'. The network is trained with a batch size of 32 over 20 epochs.

The Support Vector Machine and Random Forest classifiers were implemented using the `scikit-learn` [7] machine learning library and the Neural Networks were implemented using `Tensorflow` [1] with the `Keras` [11] API. All the experiments are repeated 10 times while randomizing the groups of testing and training users at each iteration. The mean accuracy across the repetitions is reported.

5.2 Results

We introduce our results for personal information prediction based on touch interactions in Table 3. The results are presented across the machine learning and data selection approaches on the image gallery and social media tasks. The baseline performance for each of these experiments is 50% as the output is binary and this accuracy is achievable by guessing that all examples belong to one of the classes.

The models predicting the gender of users performed well on the image gallery task and consistently achieved more than 60% accuracy with a maximum of 62% using the multi-stroke approach. The same experiments resulted in ~3 percentage points lower on the social media task. The model predicting the dominant hand of a user performed well on the image gallery task, achieving over 60% accuracy consistently. However, it does not perform comparably on the social media task, suggesting swiping behavior is more distinguishable between left and right-handed users. The model predicting the country of origin attribute fails to achieve any reasonable performance on both tasks with high standard deviation across iterations. This could be due to the small number of samples for one of the classes but it is also possible that there are no intrinsic differences in touch behavior between countries and cultures. The age group prediction models achieved reasonably high results on the social media task with up to 65% accuracy but results are closer to ~55% on the image gallery task. That is the opposite of the height prediction model, which performed well on image gallery task but poorly on social media task The weight model is consistent across the two tasks but only performed slightly above the baseline at ~55% on both tasks.

It is worth noting that some of the attributes we are trying to predict are related to each other. For instance, the height and weight of users are likely correlated and both are likely correlated with the gender since females tend to be shorter and therefore lighter on average.

Overall, the single-stroke approach performed the worst with an average accuracy of 56.7% across all modalities. We excluded the country of origin in this analysis as the results on this task were not meaningful. Averaging out each feature across the whole session, performed better at 58.1%. We found that the best performing method was the multi-stroke with 58.6% accuracy on average. In general, the SVM model performs marginally better than the rest of the classifiers with an average of 58.2% accuracy across all modalities (excluding the country of origin). The second best is the Neural Network with 57.8% accuracy and then the Random Forest with 57.4%. The differences are not large enough to strongly recommend using one model over another. However, the SVM model struggles with scaling to a large number of examples and might be undesirable for practical use. The attribute prediction models performed slightly better on the image gallery task which mainly consists of swiping (left/right) behavior with 58.7% accuracy on average across all modalities tested (excluding the country of origin). The social media task achieved 56.8% accuracy on average on the same tasks.

Since the multi-stroke scenario was the best performing method in our experiments, we decided to test the approach at varying window sizes of strokes. The group sizes we considered were 5, 10, 15, 20, 30, and one where we use the whole session data available. We use the Neural Network classifier for this comparison. The results of this experiment are shown in Figrue 4. Overall, using more swipes resulted in better performance and less variation of performance across multiple iterations. However, differences in performance are small and inconsistent across the modalities.

6 DISCUSSION

Our results show the feasibility of touch interactions to be used as a method for tracking users and revealing their personal information. While some of the results in both the fingerprinting and personal information leakage scenarios are good, the immediate threat to the privacy of mobile users is limited. However, considering no permissions are needed to collect touch data, we believe the technology can be applied in conjunction with other methods to achieve much better performance. For instance, we can use additional available data such as system and hardware attributes for the fingerprinting scenario [18] and keystroke behavior for the personal information leakage scenario [9].

As mentioned, the technology we describe in this paper can be used for malicious and undesirable goals such as discrimination, surveillance, and even identity theft. These can be manifested in many ways - banks assessing your creditworthiness, employers discriminating based on gender, and governments oppressing dissidents. However, it can also be used beneficially, to personalize the experience of users, particularly if they knowingly consent to such use.

We found that in general, personal information prediction studies in related work report much higher accuracy than the ones we achieved. This could be due to a variety of reasons including dataset quality, experimental protocols, and machine learning processing. However, we want to highlight, once again, the issue of using data from the same user in both training and testing sets. This is problematic even if the test data is coming from different sessions than the training one. Often, in related work, it is not clear whether this division of data is maintained [4, 14, 27, 30]. In order to exemplify this pitfall, we conducted the experiments described in Section 5.1 without fulfilling the condition of separating users into training

Table 3: Results for personal information leakage from touch interactions. The attributes investigated are Gender (male or female), Dominant hand (left or right), Country of origin (USA or India), Age (≤25 or ≥45), Height (≤159cm or ≥183cm) and Weight (≤50kg or ≥91kg) Each experiment is repeated 10 times and the average of all iterations is given. The standard deviation is shown in parentheses.

		SVM			Random Forest			Neural Network		
		Single	Multi	Session	Single	Multi	Session	Single	Multi	Session
Image Gallery Task	Gender	59% (±3)	62% (±4)	62% (±6)	59% (±3)	61% (±4)	63% (±5)	59% (±3)	61% (±3)	63% (±5)
	Hand	60% (±5)	63% (±6)	61% (±7)	61% (±8)	62% (±10)	57% (±8)	59% (±7)	64% (±7)	61% (±8)
	Country	53% (±14)	53% (±17)	55% (±21)	52% (±16)	52% (±19)	50% (±23)	54% (±14)	56% (±15)	52% (±19)
	Age	54% (±3)	56% (±4)	58% (±6)	53% (±4)	54% (±6)	58% (±4)	55% (±4)	56% (±5)	59% (±6)
	Height	60% (±2)	62% (±3)	60% (±5)	59% (±2)	61% (±4)	61% (±4)	59% (±2)	61% (±4)	58% (±4)
	Weight	56% (±6)	57% (±6)	56% (±5)	54% (±3)	56% (±6)	55% (±5)	55% (±4)	56% (±5)	57% (±5)
Social Medial Task	Gender	57% (±3)	59% (±4)	58% (±4)	56% (±4)	58% (±5)	59% (±6)	56% (±3)	59% (±4)	58% (±4)
	Hand	54% (±7)	55% (±9)	57% (±10)	54% (±9)	56% (±11)	48% (±10)	52% (±7)	55% (±9)	54% (±11)
	Country	50% (±10)	47% (±17)	53% (±17)	49% (±10)	46% (±16)	57% (±18)	48% (±11)	45% (±18)	53% (±13)
	Age	62% (±2)	65% (±3)	65% (±4)	60% (±3)	63% (±4)	65% (±4)	62% (±2)	65% (±3)	64% (±4)
	Height	54% (±2)	54% (±3)	53% (±4)	53% (±2)	53% (±3)	54% (±5)	53% (±2)	53% (±3)	52% (±5)
	Weight	55% (±3)	58% (±4)	55% (±5)	54% (±3)	56% (±4)	56% (±4)	55% (±3)	58% (±4)	55% (±4)

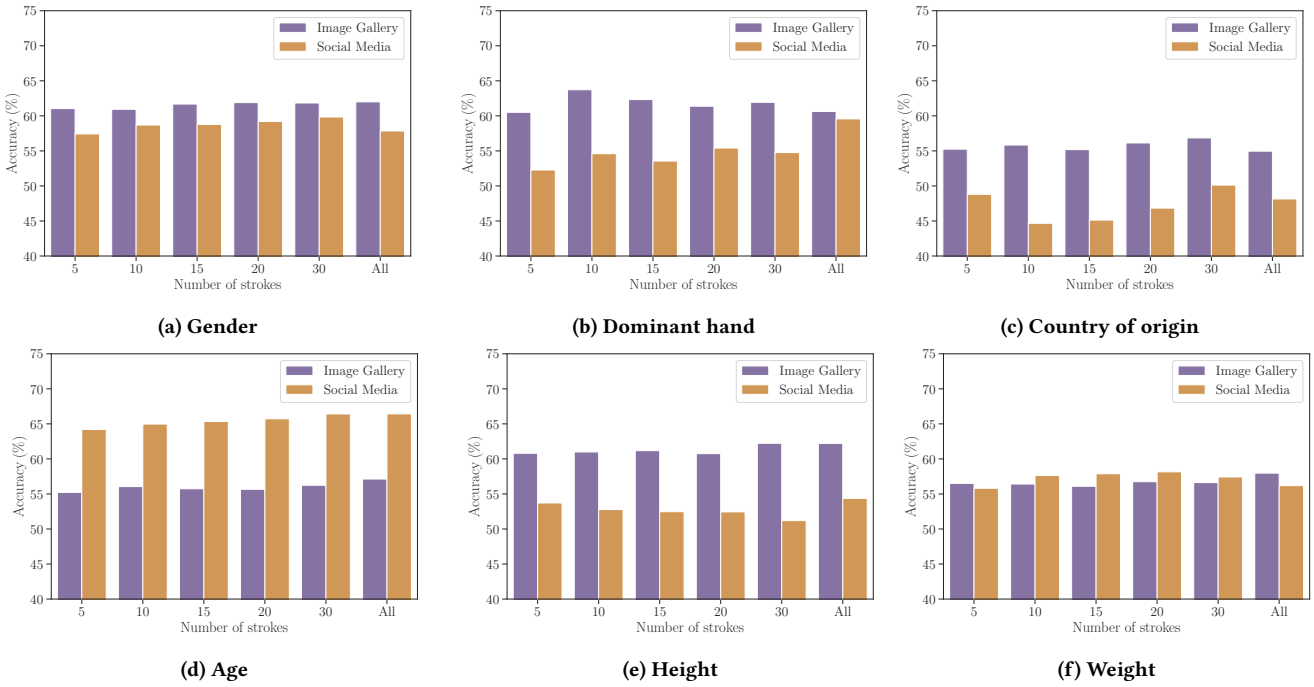
and testing groups. We used a Neural Network classifier with the multi-stroke approach on the whole session data. The results of the experiments are shown in Table 4 and are compared to the realistic evaluation method. The unrealistic approach produces higher results in all of the cases we tested with an increase of between 7.2% and 23.6%.

The SVM approach to the fingerprinting problem resulted in much better performance than the vector distance approach. There might be computational performance concerns over creating single models for each user. However, since each model is relatively small, we found that the SVM approach was not computationally heavy when applied to the dataset we used. In fact, without formally analyzing performance, we recorded the following times for model training and decision-making on a commercial off-the-shelf computer. Training all of the SVM user models took 1.1 seconds in total and a decision about a session was made on average in 52ms. These are similar to the decision function approach where there is no training, however, a decision is made on average in 60ms.

It is also worth noting that the simulation framework described in Section 4.2 can be used in other fingerprinting approaches beyond touch interactions. It can be particularly useful for continuous feature values such as the ones coming from sensors.

Table 4: Performance of personal information leakage models including data from the same user in the training and testing sets (i.e. no user separation between sets). The performance of a realistic evaluation is also given and the comparison is done on the Neural Network model using the multi-swipe approach on the whole session data.

	Image Gallery Task		Social Media Task	
Attribute	No user separation	Realistic	No user separation	Realistic
Gender	85.3%	62.0%	81.4%	57.8%
Handedness	81.9%	60.6%	70.4%	59.6%
Country	81.0%	55.0%	63.6%	48.1%
Age	64.3%	57.1%	75.3%	66.4%
Height	74.6%	62.2%	75.7%	54.4%
Weight	65.2%	58.0%	69.9%	56.2%



(a) Gender **(b) Dominant hand** **(c) Country of origin**

(d) Age **(e) Height** **(f) Weight**

Figure 4: Performance of multi-stroke models on personal information leakage predictions when varying the number of strokes considered. The results are shown on the image gallery task with a Neural Network classifier.

6.1 Countermeasures

Countermeasures to the privacy issues described in this paper are difficult to implement considering the pervasiveness of the technology. We believe at the current stage of this investigation, the ways in which users can be protected from fingerprinting and personal information leakage is limited. We briefly discuss some of the approaches taken in the field and the issues relating to their use:

- Permission requests - This approach requires users to accept permission requests before allowing the use of specific sensors. It is often proposed in related work based on mobile device sensors [26, 37]. However, asking for permissions is not possible in the touch-based case as interacting with the touchscreen is a necessity for using the mobile device in the first place.
- Limiting sampling rate - Another approach could be to limit the touch sampling rate of the browser/application APIs. This might be able to reduce the accuracy of our models, however, it is also likely to reduce the smoothness of operation of the mobile device and hence have an impact on the user experience.
- Disabling JavaScript - It might not be possible to collect touch interactions on a mobile browser without JavaScript. However, removing this core functionality of many websites would impact the user experience and completely prevent the use of some functionality. Furthermore, it is not possible to disable collection on mobile application APIs.
- Software prevention - The browser, operating system or third-party extensions can check for malicious patterns of use in the touch APIs. However, data and behavior can be obfuscated by website and application developers. This approach can become a race between security specialists and malicious actors.

6.2 Limitations and Further work

While we present strong results in using touch interactions as a method for privacy invasion, there are a number of possible improvements to our work.

First of all, our investigation could benefit from collecting a larger dataset to ensure the validity of our results. This is particularly important for the fingerprinting section, where in practice, the number of users is an order of magnitude larger, and scalability might be an issue. Furthermore, it would be more realistic to also collect data on mobile browsers, which can be slightly different from the native touch API data we use. In fact, the limitations and differences in sampling imposed by the browser APIs themselves could be used for device fingerprinting. We believe that the fingerprinting method we propose can also be used for cross-browser and cross-device tracking as the behavior of the users themselves become the fingerprint. This has been suggested in related work [25], however, further experimentation is needed to determine its validity.

The features we use in the fingerprinting scenario are based on averaging out individual stroke features over the whole session. This might not be the optimal approach for feature extraction and better feature engineering might increase the performance of the model. For instance, that can include methods suitable for reducing

the dimensionality of the feature data such as autoencoders or time-series analysis. Furthermore, using more sophisticated distance measures or machine learning algorithms as a dissimilarity function could yield better overall results. In particular, ones tailored for high dimensional vector processing.

We have shown that threshold selection is an important part of the fingerprinting system and finding an optimal threshold can be difficult and imprecise. We believe that further research and quantitative results using different approaches are needed.

The dataset we use has been collected remotely and the personal information associated with the users has been self-reported as mentioned in Section 5.1. Some of the personal information such as date of birth and country of origin can be considered too sensitive and personally identifiable. It is not unreasonable to assume that some of the participants have opted out to obscure their real personal information. However, users do not have an incentive to 'fake' their behavior while using the application. That is why we excluded users from our experiments only in the personal information leakage section and not in the fingerprinting section of our study.

The usefulness of splitting the age, height, and weight attributes into binary classes is limited. Creating regression models or bucketing the values for multi-class classification would be preferable. Furthermore, the number of features selected for each attribute (i.e. gender, age, dominant hand) is based on preliminary experiments and can be somewhat arbitrary. The optimal feature count might differ between the single stroke and whole session scenarios. A more thorough analysis of feature selection in the personal information leakage experiments is needed.

Finally, the countermeasures we propose are superficial and impractical for implementation without understanding their impact. Further work is needed to develop better countermeasures to combat the issues described in this study and quantify their effectiveness. It is also unclear whether our results are applicable to larger mobile devices such as tablets and further work is required to establish that.

7 CONCLUSION

In this paper, we illustrated how the privacy of users can be compromised without their explicit knowledge by using touchscreen interactions. We introduced a simulation system to measure the extent to which swipes and scrolls can be used to track mobile users online. Our results demonstrated that a user can be fingerprinted using touch interactions with high accuracy and we showed that the technology can be used in conjunction with other features for additional performance. In the second part of the paper, we investigated how touchscreen interactions can be used to reveal personal information of users such as their age, gender, dominant hand, country of origin, height, and weight. Our findings suggest that age, gender, dominant hand, and height can be consistently predicted with accuracies of over 60% on certain tasks. Furthermore, we showed that imprecise evaluation methods can lead to an artificial increase in the performance of models by up to 23.6%. Finally, we briefly discussed our findings and potential countermeasures to the threats and described the limitations of the current study whilst giving directions for further work.

REFERENCES

[1] Martín Abadi, Ashish Agarwal, Paul Barham, Eugene Brevdo, Zhifeng Chen, Craig Citro, Greg S. Corrado, Andy Davis, et al. 2015. TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems. https://www.tensorflow.org/ Software available from tensorflow.org.

[2] Alejandro Acien, Aythami Morales, Julian Fierrez, Ruben Vera-Rodriguez, and Javier Hernandez-Ortega. 2019. Active detection of age groups based on touch interaction. *IET Biometrics* 8, 1 (2019), 101–108.

[3] Sara Amini, Vahid Noroozi, Sara Bahaadini, S Yu Philip, and Chris Kanich. 2018. Deepfp: A deep learning framework for user fingerprinting via mobile motion sensors. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 84–91.

[4] Margit Antal, Zsolt Bokor, and László Zsolt Szabó. 2015. Information revealed from scrolling interactions on mobile devices. *Pattern Recognition Letters* 56 (2015), 7–13.

[5] Chris Bevan and Danaë Stanton Fraser. 2016. Different strokes for different folks? Revealing the physical characteristics of smartphone users from their swipe gestures. *International Journal of Human-Computer Studies* 88 (2016), 51–61.

[6] Hristo Bojinov, Yan Michalevsky, Gabi Nakibly, and Dan Boneh. 2014. Mobile device identification via sensor fingerprinting. *arXiv preprint arXiv:1408.1416* (2014).

[7] Lars Buitinck, Gilles Louppe, Mathieu Blondel, Fabian Pedregosa, Andreas Mueller, Olivier Grisel, Vlad Niculae, Peter Prettenhofer, Alexandre Gramfort, Jaques Grobler, Robert Layton, Jake VanderPlas, Arnaud Joly, Brian Holt, and Gaël Varoquaux. 2013. API design for machine learning software: experiences from the scikit-learn project. In *ECML PKDD Workshop: Languages for Data Mining and Machine Learning*. 108–122.

[8] Tomasz Bujlow, Valentín Carela-Español, Josep Sole-Pareta, and Pere Barlet-Ros. 2017. A survey on web tracking: Mechanisms, implications, and defenses. *Proc. IEEE* 105, 8 (2017), 1476–1510.

[9] Attaullah Buriro, Zahid Akhtar, Bruno Crispo, and Filippo Del Frari. 2016. Age, gender and operating-hand estimation on smart mobile devices. In *2016 International Conference of the Biometrics Special Interest Group (BIOSIG)*. IEEE, 1–5.

[10] Yushi Cheng, Xiaoyu Ji, Xiaopeng Li, Tianchen Zhang, Sharaf Malebary, Xianshan Qu, and Wenyuan Xu. 2020. Identifying child users via touchscreen interactions. *ACM Transactions on Sensor Networks (TOSN)* 16, 4 (2020), 1–25.

[11] François Chollet et al. 2015. Keras. https://keras.io.

[12] Anupam Das, Nikita Borisov, and Matthew Caesar. 2016. Tracking Mobile Web Users Through Motion Sensors: Attacks and Defenses.. In *NDSS*.

[13] Anupam Das, Nikita Borisov, and Edward Chou. 2018. Every Move You Make: Exploring Practical Issues in Smartphone Motion Sensor Fingerprinting and Countermeasures. *Proc. Priv. Enhancing Technol.* 2018, 1 (2018), 88–108.

[14] Storm P Davis, Alireza Ashayer, and Nasseh Tabrizi. 2020. Predicting Sex and Age using Swipe-Gesture Data from a Mobile Device. In *2020 IEEE Intl Conf on Parallel & Distributed Processing with Applications, Big Data & Cloud Computing, Sustainable Computing & Communications, Social Computing & Networking (ISPA/BDCloud/SocialCom/SustainCom)*. IEEE, 1136–1143.

[15] M. Frank, R. Biedert, E. Ma, I. Martinovic, and D. Song. 2013. Touchalytics: On the Applicability of Touchscreen Input as a Behavioral Biometric for Continuous Authentication. *IEEE Transactions on Information Forensics and Security* 8, 1 (2013), 136–148.

[16] Martin Georgiev, Simon Eberz, and Ivan Martinovic. 2022. Techniques for Continuous Touch-Based Authentication Modeling. https://doi.org/10.48550/ARXIV.2207.12140

[17] Martin Georgiev, Simon Eberz, Henry Turner, Giulio Lovisotto, and Ivan Martinovic. 2022. Common evaluation pitfalls in touch-based authentication systems. In *Proceedings of the 2022 ACM on Asia Conference on Computer and Communications Security*. 1049–1063.

[18] Thomas Hupperich, Davide Maiorca, Marc Kührer, Thorsten Holz, and Giorgio Giacinto. 2015. On the robustness of mobile device fingerprinting: Can mobile users escape modern web-tracking mechanisms?. In *Proceedings of the 31st Annual Computer Security Applications Conference*. 191–200.

[19] Ankita Jain and Vivek Kanhangad. 2019. Gender recognition in smartphones using touchscreen gestures. *Pattern Recognition Letters* 125 (2019), 604–611.

[20] Andreas Kurtz, Hugo Gascon, Tobias Becker, Konrad Rieck, and Felix C Freiling. 2016. Fingerprinting Mobile Devices Using Personalized Configurations. *Proc. Priv. Enhancing Technol.* 2016, 1 (2016), 4–19.

[21] Pierre Laperdrix, Nataliia Bielova, Benoit Baudry, and Gildas Avoine. 2020. Browser fingerprinting: A survey. *ACM Transactions on the Web (TWEB)* 14, 2 (2020), 1–33.

[22] Pierre Laperdrix, Walter Rudametkin, and Benoit Baudry. 2016. Beauty and the beast: Diverting modern web browsers to build unique browser fingerprints. In *2016 IEEE Symposium on Security and Privacy (SP)*. IEEE, 878–894.

[23] Lingjun Li, Xinxin Zhao, and Guoliang Xue. 2013. Unobservable Re-authentication for Smartphones. In *20th Annual Network and Distributed System Security Symposium, NDSS 2013, San Diego, California, USA, February 24-27, 2013*.

The Internet Society. https://www.ndss-symposium.org/ndss2013/unobservable-re-authentication-smartphones
[24] Eric Malmi and Ingmar Weber. 2016. You are what apps you use: Demographic prediction based on user's apps. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 10. 635–638.
[25] Rahat Masood, Benjamin Zi Hao Zhao, Hassan Jameel Asghar, and Mohamed Ali Kaafar. 2018. Touch and You're Trapp (ck) ed: Quantifying the Uniqueness of Touch Gestures for Tracking. *Proc. Priv. Enhancing Technol.* 2018, 2 (2018), 122–142.
[26] Nikolay Matyunin, Yujue Wang, Tolga Arul, Kristian Kullmann, Jakub Szefer, and Stefan Katzenbeisser. 2019. Magneticspy: Exploiting magnetometer in mobile devices for website and application fingerprinting. In *Proceedings of the 18th ACM Workshop on Privacy in the Electronic Society*. 135–149.
[27] Oscar Miguel-Hurtado, Sarah V Stevenage, Chris Bevan, and Richard Guest. 2016. Predicting sex as a soft-biometrics from device interaction swipe gestures. *Pattern Recognition Letters* 79 (2016), 44–51.
[28] Tempestt J Neal and Damon L Woodard. 2018. A gender-specific behavioral analysis of mobile device usage data. In *2018 IEEE 4th International Conference on Identity, Security, and Behavior Analysis (ISBA)*. IEEE, 1–8.
[29] Tempestt J Neal and Damon L Woodard. 2019. You are not acting like yourself: A study on soft biometric classification, person identification, and mobile device use. *IEEE Transactions on Biometrics, Behavior, and Identity Science* 1, 2 (2019), 109–122.
[30] Toan Nguyen, Aditi Roy, and Nasir Memon. 2019. Kid on the phone! Toward automatic detection of children on mobile devices. *Computers & Security* 84 (2019), 334–348.
[31] Vishal M. Patel, Rama Chellappa, Deepak Chandra, and Brandon Barbello. 2016. Continuous User Authentication on Mobile Devices: Recent progress and remaining challenges. *IEEE Signal Processing Magazine* 33, 4 (2016), 49–61. https://doi.org/10.1109/MSP.2016.2555335
[32] A. Serwadda, V. V. Phoha, and Z. Wang. 2013. Which verifiers work?: A benchmark evaluation of touch-based authentication algorithms. In *2013 IEEE Sixth International Conference on Biometrics: Theory, Applications and Systems (BTAS)*. 1–8. https://doi.org/10.1109/BTAS.2013.6712758
[33] Pin Shen Teh, Ning Zhang, Andrew Beng Jin Teoh, and Ke Chen. 2016. A survey on touch dynamics authentication in mobile devices. *Computers Security* 59 (2016), 210–235. https://doi.org/10.1016/j.cose.2016.03.003
[34] Baylea Denise Williams. 2021. *Tactile Demographics: Predicting Demographic Information Using Touch Data from Mobile Devices*. East Carolina University.
[35] Zhiju Yang, Rui Zhao, and Chuan Yue. 2018. Effective Mobile Web User Fingerprinting via Motion Sensors. In *2018 17th IEEE International Conference On Trust, Security And Privacy In Computing And Communications/12th IEEE International Conference On Big Data Science And Engineering (TrustCom/BigDataSE)*. IEEE, 1398–1405.
[36] Ting-Fang Yen, Yinglian Xie, Fang Yu, Roger Peng Yu, and Martin Abadi. 2012. Host Fingerprinting and Tracking on the Web: Privacy and Security Implications.. In *NDSS*, Vol. 62. 66.
[37] Jiexin Zhang, Alastair R Beresford, and Ian Sheret. 2019. Sensorid: Sensor calibration fingerprinting for smartphones. In *2019 IEEE Symposium on Security and Privacy (SP)*. IEEE, 638–655.

A FINGERPRINTING RESULTS ON SOCIAL MEDIA TASK

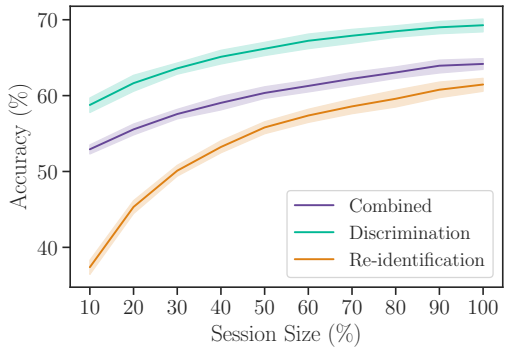


Figure 5: Performance of the fingerprinting simulation when using a fraction of the whole session data.

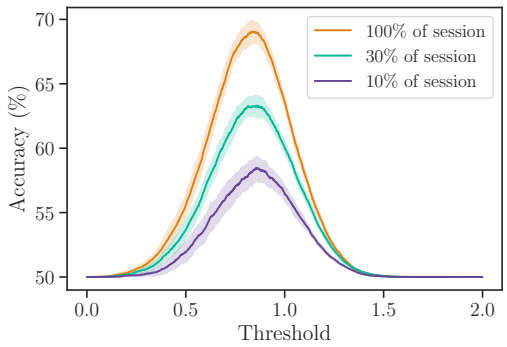


Figure 6: Discrimination model accuracy line plot at varying the thresholds (δ). Results are shown on the social media task and presented with session sizes of 10%, 30%, and 100%. The SVM fingerprinting approach without the additional phone model is used. The shaded area represents a 95% confidence interval.

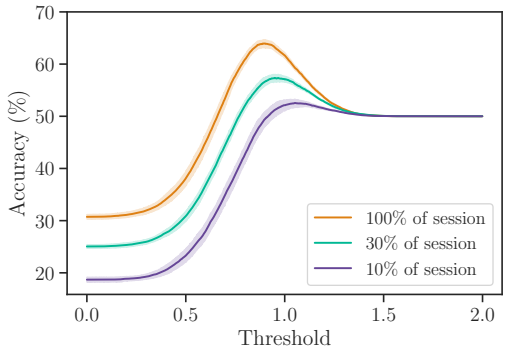


Figure 7: Combined model accuracy line plot at varying the thresholds (δ). Results are shown on the social media task and presented with session sizes of 10%, 30%, and 100%. The SVM fingerprinting approach without the additional phone model is used. The shaded area represents a 95% confidence interval.