Technische Universität Berlin

Institut fü Softwaretechnik und Theoretische Informatik
Quality and Usability Lab

Fakultaet IV
Franklinstrasse 28-29
10587 Berlin

Master Thesis

Predicting tap locations on touch screens in the field using accelerometer and gyroscope sensor readings

Emanuel Schmitt

Matriculation Number: 333772

Examined by:
Prof. Dr.-Ing. Sebastian Möller
Prof. Dr.-Ing. Axel Kuüper

Supervised by:
Dr.-Ing. Jan-Niklas Antons

Thanks to all dem brothas

Hereby I declare that I wrote this thesis myself with the help of no more than the mentioned literature and auxiliary means.

Berlin, 01.01.2050

. . . . . .

(Signature [your name])

Abstract

This template is intended to give an introduction of how to write diploma and master thesis at the chair 'Architektur der Vermittlungsknoten' of the Technische Universitï£¡t Berlin. Please don't use the term 'Technical University' in your thesis because this is a proper name.

On the one hand this PDF should give a guidance to people who will soon start to write their thesis. The overall structure is explained by examples. On the other hand this text is provided as a collection of LaTeX files that can be used as a template for a new thesis. Feel free to edit the design.

It is highly recommended to write your thesis with LaTeX. I prefer to use Miktex in combination with TeXnicCenter (both freeware) but you can use any other LaTeX software as well. For managing the references I use the open-source tool jabref. For diagrams and graphs I tend to use MS Visio with PDF plugin. Images look much better when saved as vector images. For logos and 'external' images use JPG or PNG. In your thesis you should try to explain as much as possible with the help of images.

The abstract is the most important part of your thesis. Take your time to write it as good as possible. Abstract should have no more than one page. It is normal to rewrite the abstract again and again, so probaly you won't write the final abstract before the last week of due-date. Before submitting your thesis you should give at least the abstract, the introduction and the conclusion to a native english speaker. It is likely that almost no one will read your thesis as a whole but most people will read the abstract, the introduction and the conclusion.

Start with some introductionary lines, followed by some words why your topic is relevant and why your solution is needed concluding with 'what I have done'. Don't use too many buzzwords. The abstract may also be read by people who are not familiar with your topic.

Zusammenfassung

Contents

List of Figures

List of Tables

1 Introduction

The utilization of smartphones has become an integral part of our everyday life. We use them to perform various tasks ranging from highly privacy-sensitive tasks such as for bank transactions or personal communication to more casual tasks such as setting an alarm clock or checking the weather. This universal applicability is one important factor that has contributed to the widely success of the smartphone. Another factor is the rich set of embodies sensor, such as an accelerometer, digital compass, gyroscope, GPS, microphone and camera [27] which have enabled developers to introduce highly interactive applications that provide valuable services to the ever growing smartphone user base.

Location based services, for instance, utilizing the GPS sensor [31] can lead users on the fastest route to their desired destination while health tracking applications [12], enabled by the motion sensors, can recommend health beneficial behavior based on the amount of physical activity sensed. Furthermore, newly introduced augmented reality applications utilize the camera and the motion sensors to enhance our perception of our immediate surroundings resulting in a whole new interactive experience. As these are positive example for sensor usage, there have also been reports of sensor utilization with a more malicious intent.

The motion sensors, gyroscope and accelerometer, which are typically used for detecting the device orientation and for gaming applications [16], can be used to infer the locations of touch-screen taps. As the striking force of a tapping finger creates an identifiable signature on the 3-axis motion sensors, previous research has shown that the granularity of inference is adequate to obtain PINs and passwords [11, 35, 43] or at least significantly reduce the search space to do so. The situation is furthermore reinforced by the fact that motion sensor do not require special privileges or access rights on operation system level. Do to this purpose, the motion sensors expose a vulnerable side-channel for potential attackers to eavesdrop on user interactions.

However, as the motion data used to train the inference systems in the previous research was acquired from users in a controlled setting [35, 11, 43], the feasibility of tap location inference has not been shown for a more realistic data set that is capable of modeling natural user behavior as well as their changing environments. It is plausible that when

a user interacts with the touch-screen, for instance, while walking in the park or during a public transportation ride, the sensory data will be effected by this activity potentially mitigating an eavesdropping attack.

In order to address this issue, I would like to propose *TapSensing*. TapSensing is a data acquisition system designed to acquire tap information with corresponding accelerometer and gyroscope readings. After having conducted a laboratory and field study, I have collected over 45,000 taps from 27 different subjects to investigate if the proposed security threat posed by leaking motion sensor also applies to an uncontrolled environment.

1.1 Motivation

In recent years, the number of smartphone users has rapidly increased. According to a report published by Smart Insights[1], the number of smartphone users grew from 400 million users in 2007, to more than 1,800 million in 2015. In addition, the report claims that at the end of 2017, 97% of adults, aged 18 to 34, in the US were mobile device users. Due to this rapid adaption and the smartphones rich set of sensors, smartphone have increasingly become targets for various attacks [13, 3, 11].

As gyroscope and accelerometer are commonly used for all genres of applications ranging from gaming [16] to productivity apps [45], the mobile operating system providers offer easy to use access via standard APIs[2]. Consequently, a smartphone application designed for a real-world attack could sense the user's motion in the background and send the information to a server-side application for machine learning analysis. This is possible due to the fact that background tasks are fully supported on the Android platform while on iOS, methods have been developed to run initially prohibited background tasks[3] for long durations than usually supported. Notably, all this can be achieved without the user's consent due to the lack of access restriction for motion sensors.

A further point emphasizes the security threat is that in 2017 the World Wide Web Consortium, W3C, has released the so-called Device Orientation specification [1] for JavaScript. This specification allows browser application to access the motion sensor hardware of smartphones opening further possibilities for an attacker. A potentially harmless website could therefore obtain the motion sensor data and reveal private information of the user.

[1] http://www.smartinsights.com/mobile-marketing/mobile-marketing-analytics/mobile-marketing-statistics/
[2] Application Programming Interface
[3] https://github.com/yarodevuci/backgroundTask

Besides all the privacy issues motion sensors raise, tap location inference could also be used for usability research purposes. Assuming that the inference provides high enough accuracies, it could be used as a in-app behavior tracking for applications where the source code is publicly not available. The inference system could run in a background task to evaluate the user behavior in a target application.

1.2 Outline

This work is structured as follows:

- **Related Work**: In chapter 2, previous research concerning side-channel attacks is presented. As this work is based on similar studies, they will be outlined in this chapter.
- **Machine Learning Fundamentals**: In chapter 3, I will review the machine learning fundamentals required to understand the technical aspects of this thesis. Since learning the relation between the sensory data and the tap location is a supervised learning task, the classification algorithms used in this thesis will be introduced.
- **Data Aquisition System**: In chapter 4, I will introduce *TapSensing* which is the data acquisition system used for obtaining taps and sensor readings for both the laboratory and the field environment.
- **Methodology**: In chapter 5, the methodology of the data acquisition, data pre-processing and machine learning classification is explained.
- **Results**: In chapter 6, meaningful results are presented from the data acquired study and the tap inference.
- **Conclusion**: In the final chapter 7, I will conclude the thesis with an overview of the results and ideas for future studies.

2 Related Work

The focus of this thesis lies on the practice of utilizing motion information in order to reconstruct user interactions. As this practice is a form of eavesdropping, this chapter will shed light into similar approaches of side-channel attacks that have been revealed by researchers in the past. The first section deals with different device emanations that have been utilized in various forms to obtain confidential information. The second section covers the foundations of this work as it discusses previous similar attempts predict tap locations on smartphone screens.

2.1 Eavesdropping on Emanations

Eavesdropping is defined as the the practice of secretly listening to private conversations of others without their consent [9]. As this definition originally refers to conversations between humans, eavesdropping can also be seen in terms of human-computer-interaction. In this context, the computer is seen as one conversational partner whereas the user interacting with the device is seen as the other. As the channel of communication in a human conversation is the acoustic channel, the channel in which human-computer interaction takes place has various forms. Typically, a human may enter information on a peripheral device while the computer gives feedback through an image representation. However, speech interfaces and other forms of interaction are also possible. In order to spy on the human-computer conversation, a third party with malicious intent must apply different techniques in order to spy on these interactions. A subset of these techniques which involve the utilization of device emanations will be discussed in this chapter.

In this context, a frequently discussed practice throughout academia involves the use of leaking emanations for eavesdropping purposes. These emanations can be monitored in order to carefully reconstruct the contained information. As these emanations occur in various formats, a categorization has been done based on the sensory channel they transpire. Therefore, the following sections will cover eavesdropping techniques based on acoustic, optical, electromagnetic and motion emanations.

2.1.1 Acoustic Emanations

One way of spying on electrical devices is by utilizing the acoustic channel [6, 2, 54]. Many electronic devices deploy tiny mechanics that generate sounds as a byproduct during interactions or during operation. These distinct sounds can differ in their characteristics making them adequate to identify the original information currently being processed by the machine. In these scenarios an eavesdropper targets a microphone in near proximity of the target device to capture the audio signals the device is exposing. A learning algorithm is then applied to the audio signals to reconstruct information.

In 2010, Backes et al. examined the problem of acoustic emanations of dot matrix printers, which where, at that time, still commonly used in banks and medical offices. By using a simple consumer-grade microphone, the researchers were able to recover whole sentences the printer was printing based. Backes et al. processed the audio samples in order to extracted frequency-domain features. These features worked as input for a hidden markov model, a technology commonly used in audio speech recognition. As a result, the recognition system was able to reconstruct individual characters based on the sound inputs. To demonstrate a potential attack, the researchers deployed the system in a medical office being able to obtain up to 72% of the sentences being printed on medical subscriptions [6].

Being inspired by the findings concerning the dot matrix printer, Asonov and Agrawal investigated acoustic emanations produced by hitting keystrokes on a desktop and a notebook keyboard. Following their hypothesis claiming that each keystroke has a macroscopic difference in it's construction mechanics, as well as a distinct reverberation caused by the position in the board, individual keystrokes were recorded [2]. Researchers then extracted frequency domain features from the audio signals and passed them into a neural network. In an experiment performing 300 keystrokes, 79% of the characters could be correctly recognized [2]. As this technique required substantial training before recognition, other studies have reached similar accuracies using an unsupervised approach [54] on the one hand and by using acoustic dictionaries [8] on the other.

2.1.2 Optical Emanation

Besides acoustic emanations, optical emanations can also pose a valuable source of information for a potential eavesdropper. Most electronic devices, such as notebook, smartphones and tablet computers, provide graphical user interfaces through their own built-

in screens. Even though these screens are meant to target the human eye, they can reflect off other surfaces. The reflections can be caught by high resolution camera sensors, which can then display the image revealing secret information.

One example of the use of optical emanations has been developed by Kuhn aiming to eavesdrop on cathode-ray-tube(CRT) monitors at distance. The researcher has shown that the information displayed on the monitor can be reconstructed from its distorted or even diffusely reflected light. In an experiment, Kuhn targeted a screen displaying an image against a wall while the reflections of the screen were captured using a photomultiplier. The experiment showed that enough high-frequency content remained in the emitted light for a computer to reconstruct the original image [25]. A similar approach that comprises reflections has been shown by Backes et al., however focusing on LCD displays. In this experiment, the researchers caught reflections in various objects that are commonly to be found in close proximity to a computer screen. Such objects included eyeglasses, tea pots, spoons and even plastic bottles. This work was later extended to additionally capture screens based on the reflections on the human eye's cornea [5].

2.1.3 Electro-magnetic Emanation

As electric currents flow through computer components, they emit electromagnetic waves to their near surrounding. These electromagnetic radiations can be picked up as a side channel using sensitive equipment in order to retrieve data. Electromagnetic emanations have been a research topic that has been present for several decades while the first research conducted reaches far in time.

Back in 1943, a research group under the codename TEMPEST, a subdivision of the NSA[1], were able to infer information from the infamous Bell Telephone model 131-B2, a teletype terminal which was used for encrypting wartime communication [42]. Using an oscilloscope, researchers could capture leaking electromagnetic signals from the device and by carefully examining the peaks of the recorded signals, the plain message the device was currently processing could be reconstructed [42]. This technique was later advanced and used in the the Vietnam war. Through similar electric emanation the US military could detect approaching Viet Cong trucks giving them an immense competitive advantage [38]. Today, TEMPEST is a security standard for electronic devices ensuring that

[1] National Security Agency

certified devices do not accidently emanate confidential information [39].

A second prominent finding of eavesdropping on electromagnetic emanations was done by the researcher van Eck, who discovered that cathode-ray-tube monitors could be spied upon from a distance [51]. By using general market equipment, such as antennas van Eck and standard receivers, signals emitted from the cable connecting the computer to the monitor could be received. Since these cables only transmit the video signal for visualization, the researcher could display the visual output of the target monitor revealing a full screen cast of the original image. This attack is referred to in literature as *Van Eck Phreaking* [20, 24].

Furthermore, research has shown that side-band electromagnetic emanations are present in keyboards [52], computer screens [51, 26], printers [44], computer interfaces, such as USB 2 [41] and the parallel port [50] and in Smart Cardes[46].

2.1.4 Motion Emanation

In the past decade modern devices are increasingly equipeed with highly responsive sensors, such as the gyroscope and accelerometer enabling the devices to sense rich interactions with their environment. As user interactions, such as typing the keyboard or tapping on touchscreens, require the user to apply a certain force while entering information, this motion can be captured by motion sensors in order to be used for a side-channel attack [35, 43, 11].

Marquardt et al. conducted an experiment where an Apple iPhone that captures accelerometer motion was placed next to a dekstop keyboard. Subjects then had to enter sentences while the application was monitoring the user's motion. The researchers could decode the accelerometer signals by mapping the certain vibration caused the typing motion to their keystrokes. The decoded characters were then matched based on a dictionary containing a frequency distribution of commonly used words. As a result, words could be successfully obtained with an accuracy of up to 80% [33].

As motion emanations are highly relevant for the work in this thesis, the next section will be dedicated to further work regarding this topic.

2.2 Eavesdropping on touch screen user interactions

As we have seen in the previous section, user interactions with peripheral devices, such as the keyboard or PIN pads, can be obtained by either the acoustic channel, by electromagnetical leakage or by capturing the motion of the user. However, with the rise in soft keyboard usage, the same discussed methods that extract information from keyboard do not apply to tap interactions on a touchscreen surface. Since a touchscreen does not embody fine mechanics producing sounds nor does it have emanating cables, the research community has developed nouvelle methods to spy on user inputs based on the smartphone embodied motion sensors.

The general idea behind the three approaches that are going to be discuss in the following is that a tap, or to be more precise the magnitude of the force of a tap, on a specific touch screen location creates an identifiable pattern on the motion sensors that can be sufficient to infer the initial tap location. This is particularly interesting since motion sensors are not considered as being privacy-sensitive and therefore lack access restrictions by the operating systems of the devices.

Touchlogger

The first paper regarding this security threat was published by Cai and Chen. In their proof-of-concept study they created an Android[2] application which displays a 10-digit PIN-pad. During interactions with the PIN-pad, the accelerometer signals were monitored and used for later data analysis. Having observed that a tap movement affects the rotation angle of the screen, the researchers handcrafted features based on the path of the *pitch*[3] and the *roll*[4] angles of the accelerometer. These were intersected to find a dominating edge on where the tap had presumably taken place. By using a probability density function for a Gaussian distribution the researchers were able to achieve an average accuracy of 70% for interred PIN-pad digits. The training set size involved 449 pin strokes [11]. Even though*Touchlogger* was a promising first step, due to it's low granularity of only 10 distinguishable large screen areas, it remained unclear if the attack can be carried over to a full software keyboard. Furthermore, since the inference was performed on only a single smartphone model, the question is left open whether other smartphones or tablet computers are similarly vulnerable.

[2] Android operating system for smartphones by Google Inc.
[3] The pitch-angle corresponds to the x-axis of the accelerometer.
[4] The roll angle corresponds to the y-axis of the accelerometer.

ACCessory

In order to show the feasibility for a full software keyboard, Owusu et al. performed a second attempt to the problem by creating *ACCessory*. *ACCessory* is an Android application with functionalities similar to the previously mentioned *Touchlogger*. However, the application significantly differ in it's tap area granularity providing two separate modes for tap inputs: *area mode* and *character mode*. *area mode* consists of tap areas arranged in a 60-cell grid, whereas a QWERTY keyboard within landscape orientation was displayed in *character mode*. Having extracted features mainly from the time-domain, a classification using the Random Forests algorithm reached an accuracy of 24.5% for the 60-cell grid. Here, the corresponding dataset consisted of 1300 keystrokes collected from 4 participants. As the *area mode* experiment focused on recognizing individual keystrokes, the *character mode* experiment focused on cracking passwords. By combining the keystrokes into a sequence and assuming recognition errors in individual characters, the researchers could create a ranked list of candidate passwords by running a maximum likelihood search for the most probable classification errors for an obtained password. Here, 6 out of 99 password could be inferred under 4.5 median trials given that one trails refers to traversing down one item of the candidate list. Furthermore, the majority of 59 out of 99 passwords could be inferred within 2^{15} median trials. As general result, even though the overall accuracy of the learning system scored low, the researchers could significantly reduce the search space for reconstructing a password indicating that accelerometer readings can in deed yield confidential information.

TapPrints

The most comprehensive study to date regarding the topic was conducted by Miluzzo et al. and differs from *ACCessory* and *Touchlogger* in many important ways. While both previous studies are both evaluated on the Android smartphones, *TapPrints* investigates the tap inference on both iOS and Android operating systems including tablets and smartphones alike. Another important point of differentiation is the used learning system. In order to raise the level of entropy, *TapPrints* combines readings from the accelerometer and gyroscope for a more sophisticated feature extraction. Here, time-domain and frequency-domain features, as well as the correlation and angles between individual sensor components are considered. For classification purposes, the researchers use an ensemble method combining decision tress, support vector machines, k-nearest neighbors and multinomial logistic regression in a winner-takes-it-all[5] voting fashion. The dataset

[5] This implies that all classifiers classify separately and the classifier with the highest prediction score wins.

collected in this experiments contains over 40.000 individual taps collected from 10 different users. In addition, the researchers also requested user to use different input modalities while typing including the usage of the index finger and thumb.

The *TapPrints* undertaking consists of two separate experiments: The first is a icon tapping experiment where icons are arranged in a 20 cell grid and the second one being a letter tapping experiment involving the standard software keyboard offered by the operating system. In the first experiment, an average accuracy of 78% was achieved for the iPhone whereas 67% of icon taps could be correctly inferred on the Android device. In the letter tapping experiment users were asked to enter pangram[6] sentences on the OS soft-keyboard. Results showed that on both iPhone and Android an average of 43% of the letters could be correctly classified. Even though the average accuracy for individual letters were not particularly high, Miluzzo et al. could show that when pangrams were repeatedly entered, a majority vote could be applies to individual character recognitions allowing to recover the whole pangram in approx. 15 trials. To conclude, *TapPrints* could demonstrate that motion sensor can be used to obtain passwords on multiple platforms and formats and with different input modalities.

Comparison to similar studies

Since the data used in *TapPrints* and in the other related work was collected in a controlled environment [43, 11, 35], it is not possible to tell if the feasibility of tap inference will also apply to data collected in a field environment. As this has not been investigated, this will form the central research question in this thesis. Additional questions will be discussed in the hypothesis section.

To draw the border between this study and the previous mentioned ones, this study will differ or relate as follows:

- **User interface**: For data acquisition, the user interface will cover tap area grids, as we have seen in *ACCessory* and *Touchlogger*, with 4, 12 and 20 distinguishable classes.
- **Devices**: Unlike *Touchlogger* and *ACCessory*, the devices used for this study are on the iOS Platform. Apple iPhone 6, 6s and 7 will be considered due to their mutual screen size.

[6] A pangram is a sentence using every letter of a given alphabet at least once.

- **Motion sensors**: *TapPrints* has shown that the gyroscope yields more information than the accelerometer [35], therefore both sensors will be monitored. Furthermore, sensors will be read at a frequency of 100Hz, since this had been proven to deliver best results [35, 43].
- **Dataset**: In comparison to related studies, a dataset containing data points collected in a laboratory environment and the field environment will be acquired from a total of 27 users. This data will contain the index finder and thumb as input modalities and standing and sitting as body postures while input.
- **Feature extraction**: *TapPrints* shows a deliberate list of features [35] that will be partially adopted. The features extracted will be discussed in section TODO.
- **Classification**: A feed-forward neural network, as well as a support vector machine with radial kernel will be used for classification. More will be covered in the classifcation section.

3 Machine Learning Fundamentals

As machine learning techniques will be used for the classification of individual tap locations on a smartphone touchscreen, the following chapter will give a brief overview of the fundamental concepts evolving around statistical learning. Individual categories of learning algorithms will be discussed followed by two supervised-learning algorithms, namly Support Vector Machines and Neural Networks.

3.1 Overview and Definition

Ever since computers were invented, there has been a desire to enable them to learn [48]. This desire has grown into the field of machine learning which seeks to answer questions on how to build systems that automatically improve with experience. Machine learning covers a set of methods and algorithms designed to accomplish tasks where conventional hard-coded routines have brought insufficient results [36].

The goal of a machine learning algorithm is to learn a function f that is able to predict sensible output values $y \in Y$ give input values $x \in X$:

$$f : X \to Y \tag{3.1}$$

Solving this problem is hard as the amount of input values used for learning this function is typically smaller in size than the unseen input values on to which f is applied to. Therefore, the challenge lies in finding a function that generalizes to unseen input values without simply remembering the seen inputs.

Mathematically, machine learning problems are formalized as optimization problems of an objective function which indicates the quality of the functional mapping between X and Y. This problem can either be a maximization or minimization problem. If we speak of the latter, the objective function is referred to as the *error function*.

There are four main categories of machine learning methods to be found in literature which all differ in their approach to learn the function f and in respect to the amount of

training samples available [14, 34]. These categories consist of *supervised learning*, *unsupervised learning*, *reinforcement learning* and *evolutionary learning*. As the task at hand refers to a supervised learning problem, supervised learning will be outlined in the following.

3.2 Supervised Learning

Supervised learning refers to the case where N samples are given from $X \times Y$, called the training data set $T = \{x^{(i)}, y^{(i)} | i \in \{i = 1 \dots N\}\}$. The training data is assumed to consist of approximate samples of a target function $F : X \to Y$, that is be learned by the learning algorithm. The data samples $x^{(i)} \in X$ are called input *features* whereas $y^{(i)} \in Y$ correspond to the so-called *labels*. If Y consists of discrete labels the learning problem corresponds to a classification task whereas if Y is on a continuous scale the problem refers to a regression task (see [34]).

Practical applications of supervised learning are image recognition [49, 28], e-mail spam filtering [18] or network anomaly detection [30]. However, these are just a small subset of what can be accomplished so far.

Presumably the most widely known machine learning techniques belong to this category, such as the Support Vector Machines (SVMs), Artificial Neural Networks, Bayesian Statistics, Random Forests and Decision Trees [14].

3.3 Bias-Variance Trade-off

The error of a supervised learning algorithm can be decomposed into three components being bias, variance and noise (see [15]):

$$\text{error} = (\text{bias})^2 + \text{variance} + \text{noise} \tag{3.2}$$

The *Variance* refers to the amount of which f adapts to the variations in the training data set. Models with a high variance have a tendency to learn minor relations irrespective of the real signal of the input values and are therefore prone to *overfitting* the training examples. This phenomenon applies to very flexible models such as a complex artificial neural network. On the other hand, *bias* is the model's tendency to consistently learn the wrong things from the training examples as it can not take all the information into

account. This refers to a too simple model *underfitting* the training examples. An example of high bias would be a linear method such as linear regression trying to map a non-linear function. Finally, the *noise* is the irreducible error of the data distribution (see [22].

As the goal is to minimize the error function, there is always a trade-off between bias and variance with very flexible models having low bias and high variance, and rigid models having high bias and low variance. Therefore, the model with the ideal predictive capability is the one that leads to the best balance between bias and variance [14].

3.4 Support Vector Machines

The SVM is a non-linear kernel based extension of the so-called maximum margin classifier. Originating from binary classification problems, were $y \in \{1, -1\}$, the general idea of a maximum margin classifier is to find a separating hyperplane in the p-dimensional feature space.

This hyperplane separates the training examples leading to a maximum distance between the observations of the two classes. This distance is referred to as margin M measuring the smallest distance of a training observation towards the defined hyperplane. Mathematically, the support vector classifier can be described as following optimization problem [22]:

$$\max_{\beta,\epsilon} M \tag{3.3}$$

$$\text{subject to} \sum_{s=1}^{p} \beta_s^2 = 1 \tag{3.4}$$

$$g(x_i) \quad = y_i(\beta_0, \beta_1 x_1, \ldots \beta_p x_p) \geq M(1-\epsilon) \tag{3.5}$$

$$\epsilon \geq 0, \sum_{i=1}^{n} \epsilon_i \leq C \tag{3.6}$$

The objective of this optimization problem is to maximize the margin M while choosing appropriate vector parameters β and ϵ. In this context, the parameter vector β contains the coefficients of the hyperplane whereas the vector ϵ includes so-called slack variables that account for instances which are located on the wrong side of the margin and the hyperplane. These can be expressed as follows assuming that M is positive [22]:

$$\epsilon_i = \begin{cases} 0 & g(x_i) \geq M \\ > 0 & M < g(x_i) < 0 \\ > 1 & g(x_i) > 0 \end{cases} \tag{3.7}$$

The hyperparameter C allows for a certain sum of ϵ_i observations to be on the wrong side of the margin or hyperplane, respectively [22]. C manages the bias-variance trade-off, since a low C tries to find a maximum margin hyperplane that separates the two classes, resulting in a low bias classifier for the available data set, but in a high variance classifier for test data. Subsequently, allowing a high C results in a high bias classifier that widens the margin, introducing more violations ϵ_i and reducing the variance of the classifier. Futhermore, C also controls the number of considered support vectors in dependence of the margin width (see [15]).

Extending the support vector classifier to non-linear decision boundaries brings us to the SVM. Instead of extending the predictor space using higher order polynomials and interactions, SVM uses the so called "kernel trick" [15] resulting in the optimization problem to be rewritten as follows:

$$\hat{f}(x) = \beta_0 + \sum_{i \in S} \alpha_i \langle x, x_i \rangle \tag{3.8}$$

where $i \in S$ defines the subset of support vectors and $\langle x, x_i \rangle$ is the dot product of all pairs in the support vector. Thus, the parameters β_0 and $\sum_{i \in S}$ can be estimated with the help of least squares by computing the inner products of each pair in the support vector [15]. The expression in $\langle x, x_i \rangle$ can be generalized by a kernel function

$$K(x_i, x_{i'}) = \sum_{s=1}^{p} x_{is} x_{i's} \tag{3.9}$$

indicating the linear kernel that quantifies the distance between each pair in the data set [22]. Accordingly, the equation above can be rewritten as

$$\hat{f}(x) = \beta_0 + \sum_{i \in S} \alpha_i K(x, x_i) \tag{3.10}$$

but in spite of restricting $K(\cdot, \cdot)$ to 3.9, an arbitrary kernel function can be chosen mapping the data into a high dimensional space where it is linearly separable. The "kernel trick" allows the SVM to work in an enlarged predictor space, by computing $\binom{n}{2}$ kernel functions $K(\cdot, \cdot)$, as opposed to an explicitly augmented predictor space, which is in fact computationally intractable [22]. In this work, we will be using a radial kernel function:

$$K(x_i, x_{i'}) = \exp(-\gamma(\sum_{s=1}^{p} x_{is} x_{i's})^2), \tag{3.11}$$

where γ is a tuning parameter. After the parameters are learned on the basis of the train-

ing set, a new observation with the feature vector x_0 is classified via the following decision rule

$$\hat{f}(x) = sign(\beta_0 + \sum_{i \in S} \alpha_i K(x_i, x_{i'})). \tag{3.12}$$

In view of the task at hand, an extension to multi-class classification of the SVM is utilized via one-versus-one classification. Given n classes, $\binom{n}{2}$ binary classifiers are learned.

3.5 Artificial Neural Networks

Artificial neural networks (ANNs) are computing systems inspired by the biological neural networks found in animal brains [19]. As these systems consist of several components, the first section will cover the artificial neuron which forms the fundamental processing unit of a network. Subsequently, individual activation functions of ANNs are discussed followed by the backpropagation algorithm which is used for training.

3.5.1 Artificial Neurons

A neuron is fundamental processing unit of an artificial neural network. The diagram shows a model of a neuron which consists of following components (see [19]):

- A set of *connecting links* or *synapses* which have a certain weight defined as the vector $\vec{w}$. The signal, represented as $\vec{x}$, flows through the *synapse* and is mutipied by it's weight w_i.
- A *adder* for summing the input signals and weights of the incoming synapses. These operations constitute a linear combiner.
- An *activation function* φ for limiting the amplitude of the output signal. This function is often referred to as the *squashing function* since it squashing the possible output range [19].
- An externally applied *bias* b which has the ability to lower or rise the net input to the activation function depending if the bias is negative or positive.

Mathematically, a neuron k can be expressed with the following equation [19]

$$u_k = \sum_{j=1}^{m} w_{kj} x_j \tag{3.13}$$

$$h_k = \varphi(u_k + b_k) \tag{3.14}$$

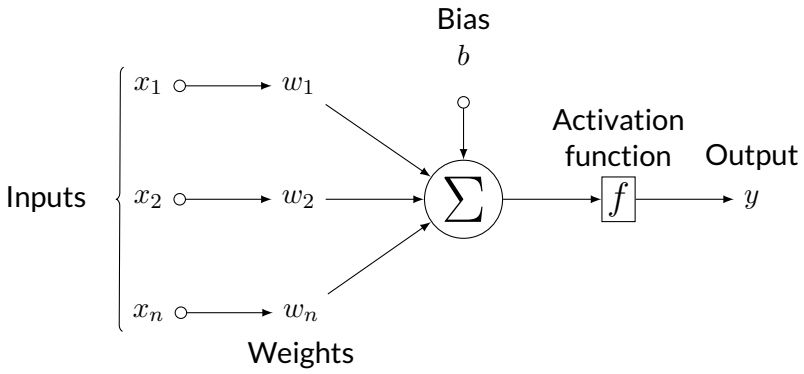

Figure 3.1: The diagram shows an artificial neuron's model with it's individual components.

where $x_1, \ldots, x_n$ are the input signals; $w_{k1}, \ldots, w_{kn}$ refer to the synaptic weights of the neuron; u_k is the linear combiner output of the summation on which the bias b is added. The output of the neuron is expressed as h_k.

3.5.2 Activation Functions

The activation function $\varphi(u_k)$ denotes the output of the neuron k and forms the junction between the neuron's input x_k and output h_k. In order for the network to learn any complex non-linear function, each neuron in the networks requires a non-linear activation function [17]. In this context, one commonly used function is the s-shaped *sigmoid function* of which the *logistic function* [19] is an example:

$$\varphi(v) = \frac{1}{1 + \exp(-v)} \tag{3.15}$$

The logistic function, as in figure 3.2, is well suited for classification as it is a non-linear function and it transforms the output values to either side of the curve. This results in a clear distinction between classes. However, the function has a compact domain range, meaning that the logistic function squashes output values into ranges $(0, 1)$. Consequently, when the inputs of a neuron become large, the function saturates at 0 or 1 with a derivative in these points being close to 0. As the derivatives are used for training the network[1], the network trains slower if the weights or biases are high. This is referred to in literature

[1] The training is performed via the backpropagation algorithm which is explained in section 3.5.4

as the *vanishing gradient problem* [37].

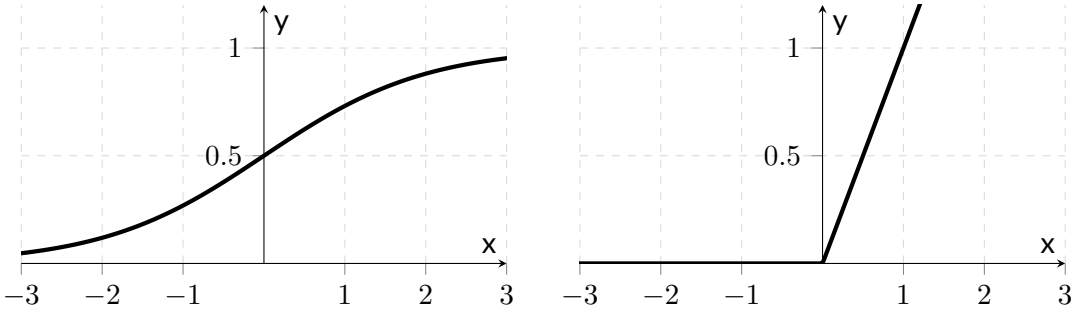

Figure 3.2: The figure shows a plot of the hyperbolic tangent activation function on the left and a plot of the ReLU activation function on the right.

In order to avoid saturation problems, a commonly used non-linear activation function is the Rectified Linear Unit (ReLU) function [37], which can be seen in figure 3.2:

$$\varphi(v) = \max(0, v) \tag{3.16}$$

ReLUs are non-saturating which results in a neuron always learning if the input is positive. Networks with ReLUs train several times faster and have become, as of 2015, the standard activation function for deep neural networks [29].

3.5.3 Feedforward neural networks

A feedforward neural network, or multilayer perceptron (MLP), is an ANN which consists of multiple layers L of artificial neurons. The goal of a feedforward network, as to other Ml algorithms, is to approximate some function $\hat{f}$ [17]. In terms of a classifier, $y = \hat{f}(x)$ maps an input x to a label y. Consequently, a neural network with m input nodes and 1 output node serves as a function with m inputs and 1 output.

The network is named feedforward as it the information flowing through the network passes with the outermost input layer and and ends at the output unit of the output layer [17]. All units in a layer are fully connected to the succeeding layer, however, there is no interconnection between units in the same layer. Figure 3.3 shows a typical feedforward architecture with three layers including a single hidden layer L_{k-1}.

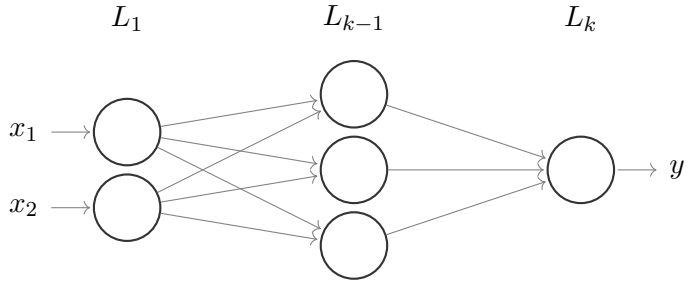

Figure 3.3: The diagram shows the typical structure of a feedforward network. The network has two input units in the input layer L_1 and one output unit in the output layer L_k. Layers $L_2 \dots L_{k-1}$ the the so-called hidden layers.

Considering equation 3.13, we can now compute the input to the output node, given by

$$\sum_{k=1}^{n} \alpha_k h_k \tag{3.17}$$

where h_k is the output of the kth hidden node and α_k being the weight from the kth hidden to the output node. The output unit's activation function is then applied to this value, transforming the output to the given equation

$$y = \varphi \left(\sum_{k=1}^{n} \alpha_k \varphi \left(\sum_{i=1}^{m} w_{ik} x_i + b \right) \right). \tag{3.18}$$

.

3.5.4 Backpropagation

The purpose of the backpropagation algorithm [47] is to adjust the synaptic weights of the network to approximate the function $\hat{f}$ mapping the inputs x of the training data to the corresponding output labels y. In an iterative process the algorithm computes the overall error of the functional mapping and adjusts the weights of each neuron according to it's individual error contribution. The result of this algorithm is a neural network configured to sensibly respond to unseen inputs for the specific supervised learning task.

As a first step, the weights of the network are initialized which is usally done in a randomly manner or based on a certain heuristic. Then, each input pattern p, with features $p = \{x_0, x_1, x_2, \dots, x_n\}$ and label y, is then sequentially processed, layer by layer, by

the network in two phases. In the first phase, the *forward phase*, the ouput of the network is computed. The square error for the output nodes j is then calculated as follows, where $\hat{y}_j$ denotes the output node's generated output and y_j is the desired ouput (see [19]):

$$E = \frac{1}{2}\sum_j (\hat{y}_j - y_j)^2. \tag{3.19}$$

Continuing in the *backward phase*, the network measures how much each neuron in the output layer L_k has contributed to each output neuron's error. Furthermore, as to measure the error contributions coming from each neuron in the previous layer, this step is repeated until the input layer is reached and all error contributions, the gradients, are computed. To put this in other words, the *backwards phase* measures the error gradients across all connection weights in the network by propagating the error gradients back into the network.

In this iterative process, the error gradients of the error function are calculated based on the partial derivative with respect to each connecting weight. If we define o_j as output of a neural unit with

$$o_j = \varphi(net_j) \tag{3.20}$$

and the units input as

$$net_j = \sum_{j=1}^{n} x_i w_{ij}, \tag{3.21}$$

the chain rule can be applied in order to compute the partial derivatives as follows:

$$\frac{\delta E}{\delta w_{ij}} = \frac{\delta E}{\delta o_j}\frac{\delta o_j}{\delta net_j}\frac{\delta net_j}{\delta w_{ij}} \tag{3.22}$$

Given the partial derivative in respect to the weight w_{ij}, the change of the weight Δw_{ij} can be determined. Here, the weight update equation (3.22) is computed depending on two cases. Either if the node j is in the hidden layer or in the output layer:

$$\Delta w_{ij} = -\eta\frac{\delta E}{\delta w_{ij}} = -\eta\delta_j o_i \tag{3.23}$$

$$\delta_j = \begin{cases} \varphi'(net_j)(o_j - \hat{y}_j) & \text{node } j \in L_k \\ \varphi'(net_j)\sum_k \delta_k w_{jk} & \text{node } j \notin L_k \end{cases} \tag{3.24}$$

In equation (3.23), η denotes the *learning rate*, which defines to which amount the reverse gradients are applied to the weight update; k refers to a node in the successor layer of the node j.

Once the weights are updated, the global error of the network is calculated and the forward and the backward phase are repeated for the rest of the training patterns. One pass through all of the training patterns is called a *training epoch*. Several strategies exist to stopping the described training process. One condition is to stop after to a fixed number of training epoches, a second can be a stop once the change in the weights reaches a certain low-end threshhold [19]. After the process, the final values of the weights are saved and can be used for prediting new incoming patterns.

4 Data Acquisition System

In this chapter, I would like to introduce TapSensing. TapSensing is a labeled data acquisition system designed to collect user generated taps with corresponding sensor readings. The system comprises of an iOS Swift and a Python Django server-side application. The chapter begins with a broader view on the system by explaining the overall system architecture and will then dive into it's individual components and implementation.

4.1 Overall System Architecture

The TapSensing application consists of two main components: the mobile and the server-side application. In brief, the mobile client provides the ability for a user to generate taps with corresponding motion sensor signals. For the data to be stored in a centralized manner, the server-side application provides HTTP endpoints as a gateway to the database. The architecture, as illustrated in figure 4.1, consists of various components which are outlined in the following.

- **Mobile application**: The iOS application provides interfaces for the user to generate taps and to label the data created. Furthermore, the application is capable of sending the acquired data to the server-side application. More features of the mobile application are presented in section 4.2.
- **NGINX**: For accepting and routing incoming HTTP requests, an NGINX reverse proxy/load balancer is used. The reverse proxy forwards the incoming requests to the server-side application and is capable of serving static content, such as images, HTML, CSS and JavaScript files of the form application.
- **Gunicorn**: Gunicorn[1] is a Python Web Server Gateway Interface (WSGI) HTTP server which runs the source code of the backend application.

[1] For more information on Gunicorn, visit `http://gunicorn.org/`.

- **Backend Application**: The backend application provides authentication and persistence functionalities which are accessible through HTTP endpoints. More information on the backend application is to be found in section 4.3
- **PostgreSQL**: TapSensing uses a PostgreSQL[2] database for storing the application state, user related information, the survey data and the retrieved tap and sensor information.
- **Form application**: The form application provides a web user interface where study participants can answer survey questions.

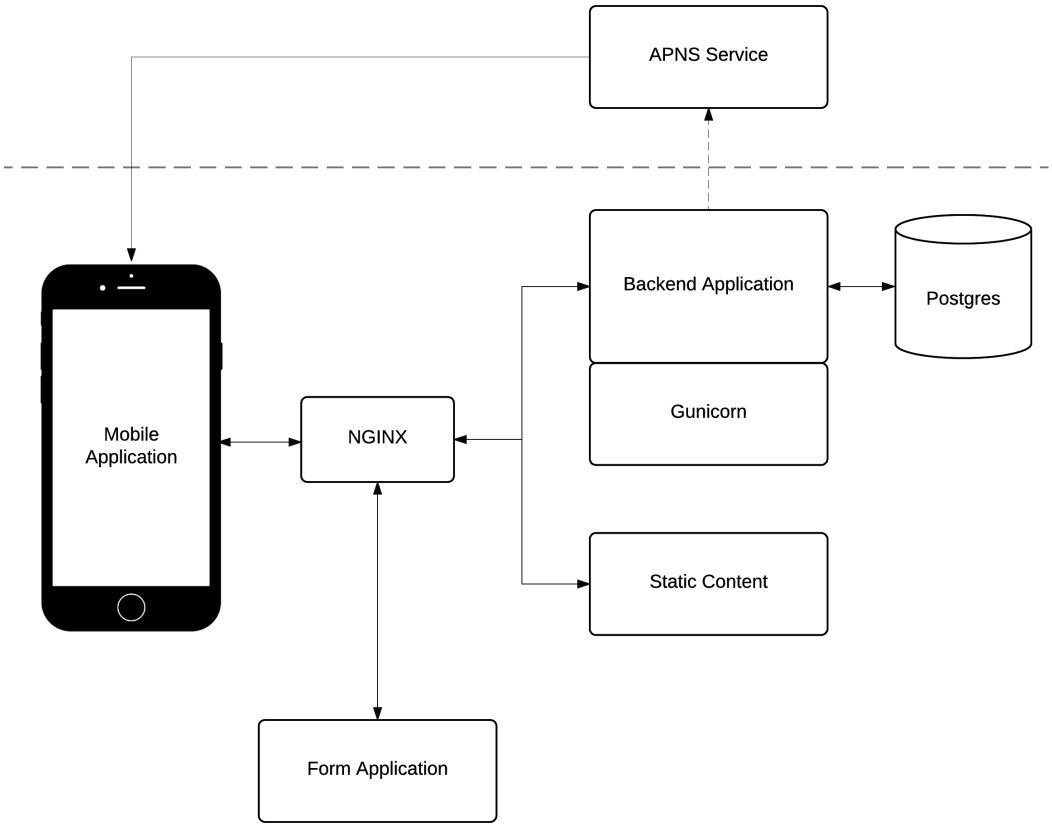

Figure 4.1: The diagram shows the overall architecture of TapSensing.

[2] PostgreSQL is a general purpose and object-relational database management system.

4.2 Mobile Application

The TapSensing mobile application is a iOS App written in the Swift programming language designed to run on devices with iOS 10 and above. The main purpose of the application is to provide a user interface for subjects to create taps with accelerometer and gyroscope readings for the labeled data acquisition phase of the experiment.

4.2.1 User Interface

Login Screen

When the application is opened for the first time, a login screen appears. As in standard login screens, the interface asks for credentials including username and password. Authenticating users, has the advantage, that the generated data can be mapped to individual users automatically.

Start Screen

During the trial the user is asked to generate taps once a day. In order to indicate if the user is eligible to perform a tap generation trial, the start screen shows a button that is either active or inactive. This switch depends on 4 distinct conditions: When data is col-

Figure 4.2: This figure shows the start screen in different configurations. On the left hand-side screenshot, the user has not performed a trial while the the middle image shows the screen where a trial has been performed. The right-hand side image shows the *lab mode* where trails can permanently be performed.

lected in the laboratory environment, the app is set to *lab mode*. In *lab mode* the button is

always active and trials can be performed. When a user has not performed a trial today, the button is inactive. Consequently, when a user has already performed a trial on a specific day, the button is inactive and a further trial can only be performed on the following day. Once all field trials are performed, the app confirms that all data is collected and the button remains inactive.

Tap Input Screen

To acquire individual user taps the mobile application offers a user interface where buttons are aligned in a grid shape structure. The structure is calculated based on a specific configuration set where the amount the vertical and horizontal buttons in the interface can be set. Figure 4.3 shows the interface, where 4, 12 and 20 distinguishable buttons are configured. For the user to tap on every location of the screen exactly once, a red button

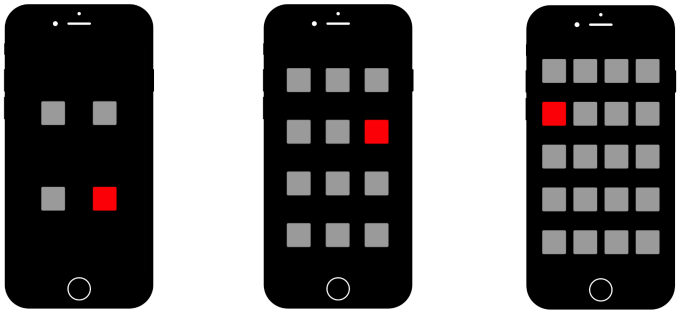

Figure 4.3: The figure shows the tap input user interface with buttons aligned in a grid shape structure. The leftmost structure offers 4 buttons, the middle offers 12 buttons whereas displays 20 distinguishable buttons.

indicating the next button to tap is highlighted guiding the user through the interaction process. While the user is tapping the grid, the gyroscope and accelerometer information is recorded. After all buttons gave been tapped, either a new grid is loaded or the tap acquisition phase ends proceeding with the question interface.

Questions Screen

To label the data acquired in the tap input interface, the application provides screens for the user to answer several questions. These question regard the body posture, input modality and the mood in which the user has generated the taps. In the table 4.1 below the questions asked with corresponding answer choices are to be found.

Figure 4.4: The figure displays question views with icons as answer possibilities.

Question	Answer Choices
Which body posture was used during the interaction ?	Standing, Sitting
Which finger did you use while tapping?	Index Finer, Thumb
Which hand did you use to tap?	Left, Right
What is your current mood?	1 - 5

Table 4.1: The Table shows the questions asked in the question view.

To enable fast interactions, each answer possibility to a question is represented in the interface with an icon. Once an icon has been pressed, the application transits to the next question until all questions have been answered.

Upload Screen

After all taps and questions are gathered, the acquired data is sent to the server. The interface at this points displays a spinning wheel for the user to acknowledge that the mobile phone is processing the data. In case an upload fails, the application provides a manual upload screen, where past sessions can be uploaded.

4.2.2 Implementation Notes

Obtaining Sensor Information

In order to access the gyroscope and the accelerometer, Apple provides a high-level API[3] for accessing the device's sensors: *Core Motion*. Core Motion provides motion and environmental related data from sensors including accelerometers, gyroscopes, pedometers, magnetometers, and barometers in easy to use manner.

Sensor values can either be accessed as proceeded including aggregations of the values or as raw version. For TapSensing, raw values are recorded to avoid any form of bias. The update interval can be configures at ranges from 10Hz - 100Hz. Higher update-rates are possible but are not ensured to be processed in real-time by the device. For TapSensing, the update rate is configured with the highest (safe) value possible. This ensures that tap patterns are captured with high resolution in order to make a classification easier.

Local Persistence

It is possible that a session upload fails due to lack of internet access or another reason. Due to this, the application stores all session information and sensory data in it's local SQLite database. For local persistence Apple provides it's own framework called *Core Data* which is extensively used in TapSensing. Once the data has been successfully received by the backend, the data is deleted from the local database.

Ensuring data consistency

Sending all collected data in a single HTTP-request results in the sent package being too large for the server-side system to process. For this reason, the data is split up into packages of 300 objects per request. To send these requests in an asynchronous manner, the *Promise*[4][32] library *Hydra*[5] is used.

During an upload phase, packages are sent in a sliding window approach. Packets are sent three at a time until all packages have been acknowledged by the server-side application.

[3] An Application Program Interface is a set of rules and subroutines provided by an application system for the developer to use. The following link leads to the Core Motion API documentation: `https://developer.apple.com/documentation/coremotion/`

[4] Promises are a software abstracting for dealing with asynchronous computation. Promises are objects that may produce a value at some point in the future.

[5] `https://github.com/malcommac/Hydra`

For each transmitted package, the server responds with the amount of data objects contained within the request. The mobile client can therefore track the amount of packages transmitted to ensure that all data has been transmitted successfully. If one packet fails to arrive, the package is resent with an exponential back-off.

Push Notifications

TapSensing is registered with the Apple Push Notification Service allowing the application to receive push notifications. Notifications are used to remind the user during the field study, that he has to take part in the study.

Distribution

Installing iOS applications that are not uploaded to the Apple App Store, requires each device to be registered in the Apple Developer Portal with the smartphone's serial number. To avoid this, the application has been uploaded to the App Store enabling an easy distribution.

4.3 Backend application

The purpose of the backend application is to provide persistence functionalities for the acquired taps generated with the mobile application. The application is written on top of the Django[6] web application framework and the Django REST framework[7].

4.3.1 HTTP Endpoints

For the purpose of interoperability the network communication from the mobile application and the forms application to the server-side application is done via HTTP Requests in the JSON[8] format. The endpoints listed below represent tiny logic components that can be called from an external client.

[6] `https://www.djangoproject.com/`
[7] `http://www.django-rest-framework.org/`
[8] JSON (JavaScript Object Notation)[10] is a lightweight data-interchange format.

HTTP Method	URL	Description
POST	/login	Provides login functionalities for the mobile client. This method returns an authentication, that is used for further requests to authenticate the user.
POST	/session	Provides upload functionality for a session data objects.
POST	/touchevent	Provides upload functionality for touchevent data objects.
POST	/sensordata	Provides upload functionality for a sensordata data objects.
POST	/apns	Retrieves the Device's APNS token. This is required to send push notifications to the users.
GET	/trial-settings	Provides the configuration for the tap input view. Here, the amount of buttons and the amount of grid repetitions that are to be performed in a single session can be defined.
POST	/survey	Provides upload functionalities for the survey form application.

Table 4.2: The table shows all HTTP endpoints of the server-side application.

4.3.2 Data Model

TapSensing's data model reflect the schemas that are used in the PostgreSQL database. As seen in figure 4.5, the data model consists of 4 data objects that I will describe in this section:

- **User**: The user model is inherited by the Django's user model[9]. The user model is used for authentication and persisting the authentication token. The other models described in this section are connected to the user model through a foreign key relation to identify the user associated with the data object.
- **Session**: The session data object holds all information associated with the user's trial. This includes the data collecting in the "labeling part"[10] of the application such as the hand used, the body posture and the typing modality. In addition, information

[9] For more information on the Django user model, the following URL leads to the model's documentation: `https://docs.djangoproject.com/en/1.11/ref/contrib/auth/`

[10] The "labeling part refers to the question views, where additional information on the data acquired is collected"'

is stored such as device infos and if the session took place in a laboratory or field environment.

- **Touchevent**: The touchevent data object stores information regarding the "ground truth" of the user generated tap including the exact x and y coordinates, timestamp and an identifier of the specific grid rectangle tapped in the tap input view. Furthermore, it is noted if the user hit a specific rectangle and if the event is a touch-down or touch-up event.
- **Sensordata**: The sensordata data object captures all information obtained by the gyroscope and accelerometer of the mobile device. To differentiate between accelerometer and gyroscope values, the data object includes a type field. In addition, the model captures the timestamp and the 3-sensor components (x, y, z) of the individual sensors.

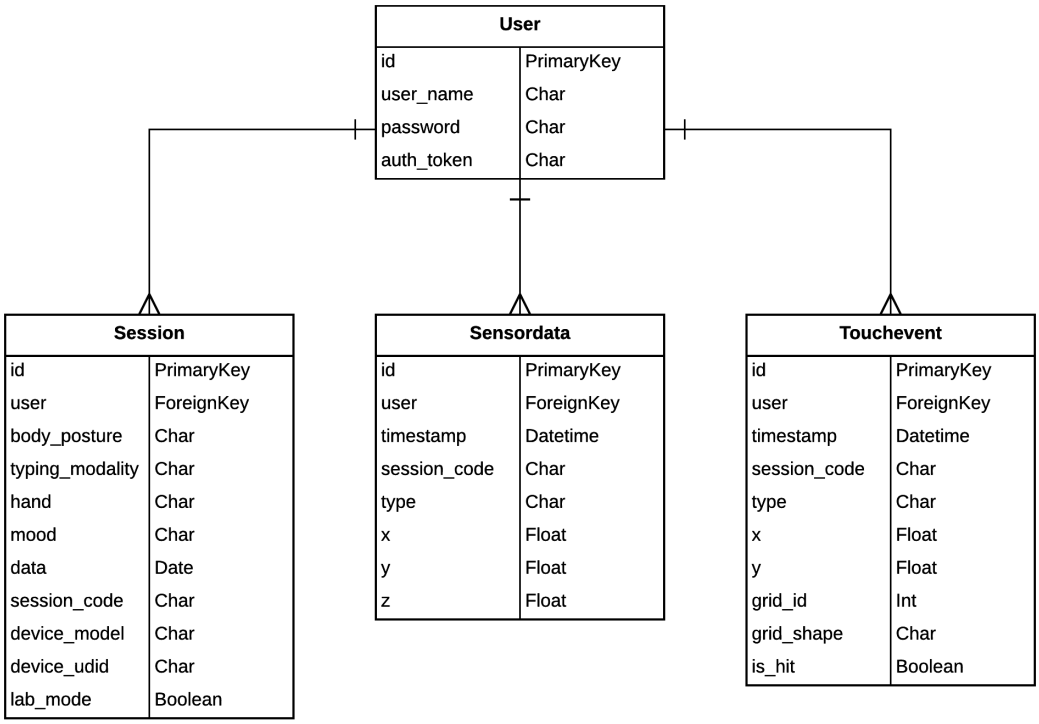

Figure 4.5: The diagram shows TapSensing's data model with the user, session, sensordata and touchevent data object.

4.3.3 Implementation Notes

Authentication

The authentication mechanism implemented in TapSensing is based on a standard token authentication scheme. This allows the association of an incoming request with a set of identifying credentials, which is - in this context - the user the request came from. In order to obtain an authentication token, the mobile application sends the login credentials of a user including username and password to the login endpoint[11]. When the credentials have been checked for validity, the endpoints returns an authentication token in the following format:

```
{ "token": "Token 3acc6c2a58723e2f1579d4526add2511f6a0a525" }
```

The token is then added to the HTTP-request in the "Authorization" header to authenticate the user.

Push Notifications

As a previous study has shown that push notifications greatly enhance user participation [7] during a mobile field study, notifications are sent on a daily basis reminding the participants to take action. In the same study[7], passive notifications - notifications without sound - have been seen to futhermore impact the participation. Following these assumptions, a push notification strategy has been designed combining notifications with and without sound. Figure **??** shows this implemented strategy with corresponding trigger times.

Time	Message	Sound
9:00 GTM+1	Good Morning. This a friendly reminder to take part in the tapsensing study today.	inactive
12:00 GTM+1	Tapping is a lot of fun. Have you tapped the buttons today?	inactive
18:00 GTM+1	I know your day is busy, but don't forget to take part in the study.	inactive
21:00 GTM+1	You have not taken part in the study today. Please do it now.	active

Table 4.3: The table shows the push notifications strategy with individual notifications sent.

[11] The login endpoint is described in section 4.3.1.

The interaction with the APNS service has been implemented using the Django package *django-push-notifications*[12]. It is used to provide mechanisms to register devices as well as to send notifications. To trigger the notifications at specific times unix cronjobs are executed.

Backups

To prevent data loss, the PostgreSQL database is backed up via a unix cronjob every evening. The database dumps are sent automatically to Amazon Web Services S3 Bucket.

Deployment

The server-side application with the NGINX reverse-proxy and a gunicorn WSGI application server has been deployed on a Ubuntu 14.04 virtual environment. The server comprises of 1GHz of shared CPU and 1 GB RAM.

[12] `https://github.com/jleclanche/django-push-notifications`

5 Method

5.1 Hypothesis

To recall, previous research has shown that is it possible to predict location on a smartphone touch screen based on acceleromter and gyroscope recordings [35, 43, 11]. However, since the data used for classification in these approaches was collected in a controlled environment, it has not been shown that the feasibility to predict tap locations also applies to a field environment. Therefore, data will be collected from both the field and the laboratory environments to investigate how classifiers, when fed with data from these environments, perform. Assumptions are made that the environment has an effect on the classifier's prediction accuracy.

H.1 The environment of recorded sensory data has an effect on the prediction accuracy.

In order to test the hypothesis stated above, a further assumption is made assuming that when a subject interacts with a smartphone screen, for instance, while walking or during a public transportation ride, the recorded sensory data will contain more noise compared to the controlled environment. This noise will, most likely, have a negative impact on the prediction accuracies.

H1.1: The prediction accuracy for a classifier trained with the data in the laboratory environment will score higher than one trained with data collected in the field.

Next, assumptions are made on the way the user interacts with the device. A user can either use the thumb to touch, which is presumably the most common way of interacting with the device, or the index finger while holding in the device in the other hand. Assuming that the input modality also has an effect on the behavior of the estimator, data sets for both hands will be evaluated.

H2: The input modality has an effect on the prediction accuracy.

It is plausible that tapping with the index finger, due to the force resistance of the supporting other hand, will cause less movement of the smartphone as typing with a thumb will. This is presumably to be reflected in the motion data.

H2.2: The prediction accuracy for a classifier trained with index finger tap data will score higher than one trained with thumb tap data.

Finally, assumptions are made based on the body posture a user has while tapping. Overall, a difference in classification results is assumed between standing and sitting.

H3: The body posture has an effect on the prediction accuracy.

To show approve or reject this hypothesis, it is assumed that a user while sitting will move less compared to who is standing which, as a result, will lead to a better predictability of the tap location.

H2.2: The prediction accuracy for a classifier trained with taps where a user sat will score higher than one trained with taps where a user stood.

5.2 Experimental Approach

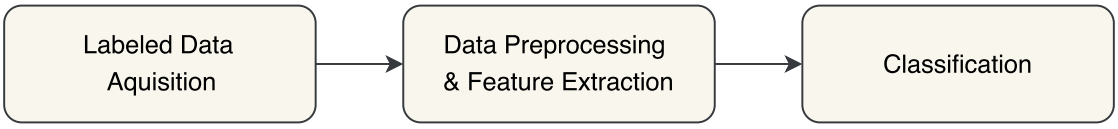

Figure 5.1: This figure shows the overall appraoch to the experiment.

The overall experiment consists of a three step process. In the first step, labeled data is acquired from subjects invited to take part in the experiment. For this purpose, the TapSensing application presented in the last chapter is used. The data acquisition part is necessary to collect sensor data with the corresponding ground truth of the tap locations for the supervised learning task.

After the data is successfully acquired, the continuos sensor recordings are preprocessed to obtain the portion of recording which represents each individual tap. To extract certain characteristics of the sensor signature created by the each tap, feature are extracted in a further step. These features are then used to train a set of classifiers.

The classification results are then discussed in the results section of this work.

5.3 Labeled Data Acquisition

5.3.1 Subjects

A total of 27 subjects were invited to participate in the study. There were no restrictions concerning the demographics of individual subjects. However, each subjects had to be in possession of one of the permitted iOS devices.

5.3.2 Devices

The devices have been restricted to the Apple iPhone 6, 6s and 7 based on their mutual screen size. Furthermore, as the screen size has a large effects on the sensor signature created by a tap, the screen size is an important factor to enable the comparison of classification results among devices.

5.3.3 Environments & Conditions

As the study aimed at collecting sensor reading both in the field as well as in a laboratory environment, the data acquisition is done in two distinct settings:

1. **Laboratory Evironment**: Subjects were invited to a laboratory room which had a standard office ergonomics setup. Subjects have been asked to either sit at the desk or stand in the room while tapping.
2. **Field Environment**: Subjects were asked to generate taps using their smartphone at any place they are currently located. For example, this could be at home, at work or during leisure activity.

Besides the environment of the recorded sensor data, the collected varies in the input modality and body posture while the tap was made. Subjects are either allowed to use the index finger (while holding the device in the other hand) or the thumb to generate taps. Sitting and Standing are allowed as body postures as these two represent the natural interactions with the smartphone.

5.3.4 Acquisition Procedure

In order for participants to generate tap data, the subjects were invited to come to the laboratory for the first part of the experiment. To avoid information asymmetry, each participant read an experiment instruction[1]. Subjects should then confirm whether they have understood the previously read information. Additional questions regarding the instructions were answered. In the next step, subjects were asked to fill out an online questionnaire regarding their persona and personal smartphone usage.

To begin with the data acquisition, participants were asked to download the application from the Apple App Store and to login using provided credentials.

Subsequently, subjects were asked to perform 6 consecutive trials in the TapSensing application, whereas one trail includes tapping each grid four times in randomized order. Recall, the grid sizes defined consist of 4, 12 and 20 distinguishable buttons. For an equal distribution of the body posture and the input modality while tapping, each trial is dependent on one of four configurations onto which each subject is assigned. These configuration are to be seen in Figure 5.2. It is important to note that each subject is not allowed to alter the body posture or input modality during a trial. After all trials are performed, subjects were asked to continue with the field study.

During the field study, subjects performed one trial daily on 6 separate days. On each day push notifications were sent as a reminder to participate. Subjects were free to decide which input modality or body posture to use as the aim of the field study is to collect data that represents regular smartphone usage.

[1] The instruction sheet is to be found in the appendix.

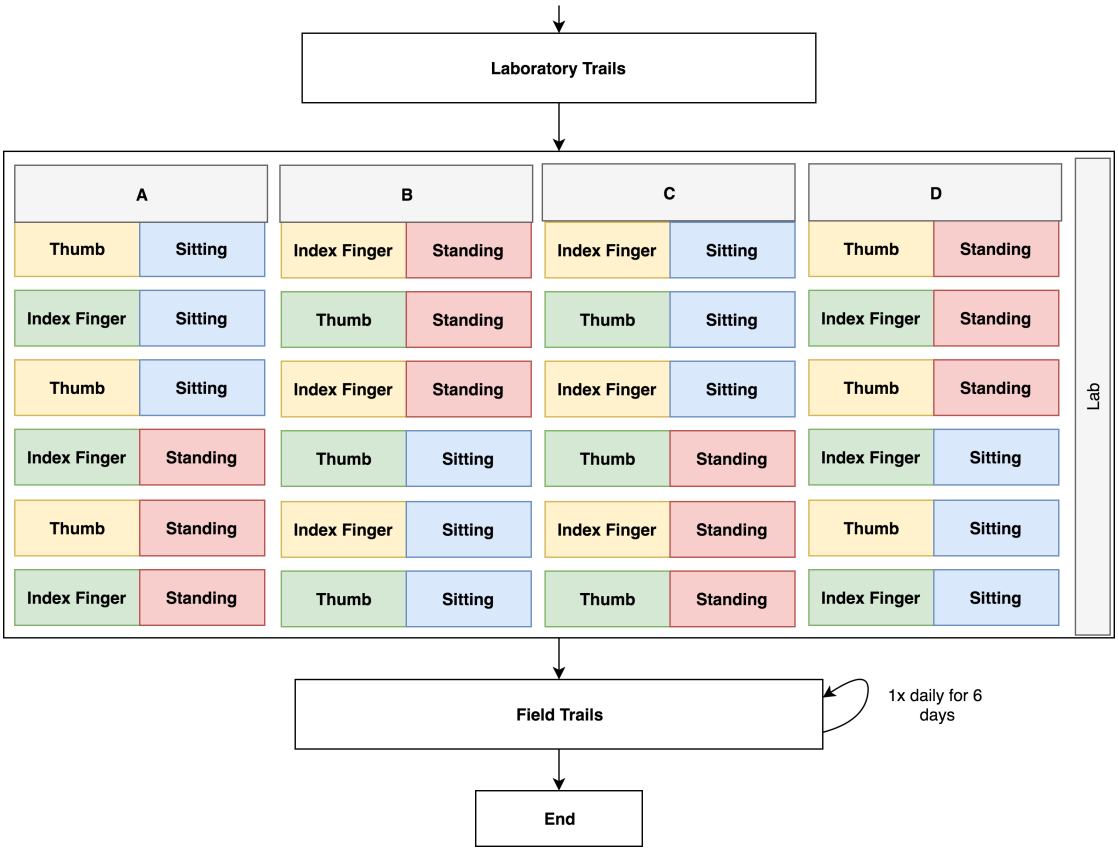

Figure 5.2: This figure shows the procedure every subject has to perform in the study.

5.4 Data Preprocessing

5.4.1 Preprocessing

Before features can be extracted from individual taps, the continuous sensor recording from each trial needs to be sliced to obtain the portion of the recordings which is relevant for the tap. The timestamp of the touchdown event is used to find an appropriate starting point. Here, 20ms are subtracted from the timestamp as due to the latency between the physical touchdown event and the recorded event. Based on the data collected in the pilot study, the average tap duration (being the duration between a touchdown and the corresponding touchup event) between 70ms - 200ms. Taking this value into account, a window size of 150ms was chosen to enrich each slice with more information. Figure 5.3 shows the sensor components of a gyroscope reading with a slicing window marked with a grey background.

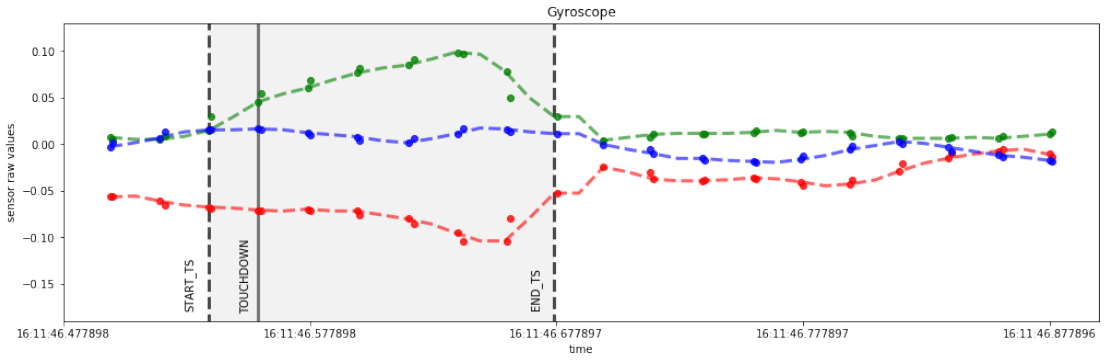

Figure 5.3: The figure shows a continuos gyroscope reading with the slicing window and corresponding timestamps. The timestamp of the touchdown event is used as an anchor point.

Unfortunately, due to different CPU loads on the mobile devices, sensor values are not pushed from the operating system to the mobile application in a constant manner leading to an uneven distribution of the sensor recordings on the time scale. To balance out each tap to a fixed amount of values, a cubic spline interpolation is used. Figure 5.3 shows the interpolated sensor components with dots representing the raw recordings. Furthermore, recordings that are below 75Hz are filtered out. After the slices are produced for each tap, features are extracted.

5.4.2 Feature Extraction

A feature is an individual measurable property or characteristic of a phenomenon being observed [14]. Therefore, choosing discriminant and independent features is a crucial step for the performance of the later classification. Fortunately, Miluzzo et al.'s preliminary work has shown a comprehensive list of possible sensor features of which the following features are partially adapted.

The features extracted are differentiated based on the set of axis where aggregational functions are applied. These can be categorized as column features, sensor features and matrix features, as illustrated in figure 5.4. Column features refer to features that are a product of a function being applied to each *x*, *y* and *z* sensor axis separately. Sensor features are defined as features where sensor axis are combined. An example for a sensor feature is the Pearson correlation which is calculated for each sensor component pair *xy*, *xz* and *yz*. With sensor features, relations between the components are captured. Lastly, matrix features are a product of a function being applied to both gyroscope and accelerometer time series.

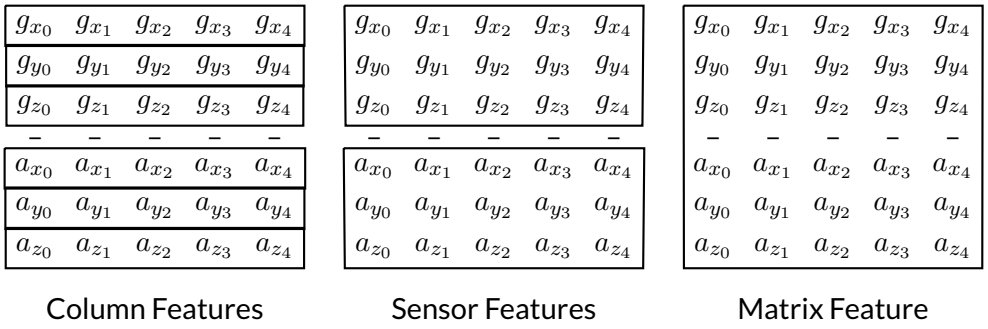

Figure 5.4: The figure shows how different features are extracted from the overall sensor matrices: Column features, sensor features and matrix features.

Since taps on different locations of the screen generate different sensor signatures we design features that are able to capture the properties of the sensor readings generated by a tap. For this purpose, a total of 230 features have been extracted for each individual tap. The table 5.1 below shows the complete list of features ordered by feature type. Features have been extracted from both the time domain and the frequency domain. In addition, to standardize the range of the features extracted, a Min-Max scaling is applied to each feature.

Name	Description	Feature Type	Amount
peak	Amount of peaks in the time series	column	6
zero_crossing	Amount of zero crossings in the signal	column	6
energy	Energy of the signal	column	6
entropy	Entropy measure of the the signal	column	6
mad	Median absolute deviation of the signal	column	6
ir	Interquartile Range	column	6
rms	Root mean square of the signal	column	6
mean	Mean of the time series	column	6
std	Standard deviation of the time series	column	6
min	Minimum of the time series	column	6
median	Median of the time series	column	6
max	Maximal value of the time series	column	6
var	Variance of the time series	column	6
skew	Skewness of the time series	column	6
kurtosis	Kurtosis of the time series	column	6
sem	Standard error of the time series	column	6
moment	Moment in the time series	column	6
spline	Spline interpolation of the signal (12 features)	column	6*12
fft	Fast Fourier Transform (5 features)	column	6*5
cos_angle	Cosine Angle of sensor component pairs	sensor	6
pears_cor	Pearson Correlation of sensor component pairs	sensor	6
fro_norm	Frobenius matrix norm	matrix	1
inf_norm	Infinity matrix norm	matrix	1
l2_norm	L2 matrix norm	matrix	1

Table 5.1: Table of features extracted from every tap.

5.5 Classification

After the features are extracted for all the tap data acquired, learning algorithms are applied in order to measure the classification accuracy. In this experiment, a SVM with radial basis kernel is used as well as a feedforward artifical neural network.

5.5.1 Evaluation

To evaluate the classifiers trained in the classification part of the experiment, a K-fold cross-validation is used. In K-fold cross-validation, the training set is divided into K subsets of equal sizes. Sequentially, every subset is tested using the classifier trained on the remaining K-1 subsets. During this process, each pattern in the dataset is predicted once. After K classifiers are trained, the mean of the accuracy scores is computed to determine how well the classifier performs.

5.5.2 Grid Search

In order to optimize the hyperparameters of each learning algorithm, a grid search is performed. A grid search is an exhausting search through a defined subset of the hyperparameter space for each algorithm. The subset of parameters are combined using the cartesian product in order to configure the classifier. To evaluate the each classifier with a set of hyperparameters, a 5-fold cross validation is performed.

The parameters C and γ of the SVM RBF kernel are tested during the grid search using exponential growing sequenced, a recommended method to find suitable parameters [21].

$$C = \{2^k | k \in \{-3, -2, ..., 14, 15\}\} \tag{5.1}$$

$$\gamma = \{2^k | k \in \{-13, -12, ..., 2, 3\}\} \tag{5.2}$$

As for the SVM, hyperparameters are defined for the ANN. Presumably the most crucial parameter in an ANN is the amount of hidden layers defined as when a network is too large it tends to overfit the training data while a too small network can lead to high bias. Therefore, networks ranging from 2 to 12 hidden units that are defined and evaluated during the grid search. To regularize the networks an L_2 penalty [40] is used. The following table 5.2 lists all configurations of the evaluated ANNs.

Hidden units	L2 Penalty	Learning Rate	Activation	Optimizer
12	0.001	0.01	RELu	Adam [23]
10	0.003	0.01		
8	0.0001	0.1		
4	0.0003			
2				

Table 5.2: ANN configurations during grid search.

After the grid search is computed, the best performing model with corresponding hyper-parameters is used for analysis.

5.5.3 Metrics

As the amount of taps are equal for each subject and grid cell, the amount of training examples for each class is balanced. Therefore, the standard accuracy score for the classifier is used.

$$A = \frac{TN + TP}{TN + FP + TP + FN} \tag{5.3}$$

, where TN is the number of true negative cases, FP is the number of false positive cases, FN is the number of false negative cases and TP is the number of true positive cases.

6 Results

6.1 Data Acquisition

The data acquisition was performed with 27 subjects in total, 12 (44%) females and 15 (56%) males. Participants had an average age of 26.4 years (17, 53, 6.39) and all 27 were right-handed (100%). 19 (70%) were students, 8 (30%) had other occupations. 18 (66%) subjects stated that their most used input modality is the their thumb while 4 (14%) preferred using their index finger and 5(18%) use both thumbs during interactions.

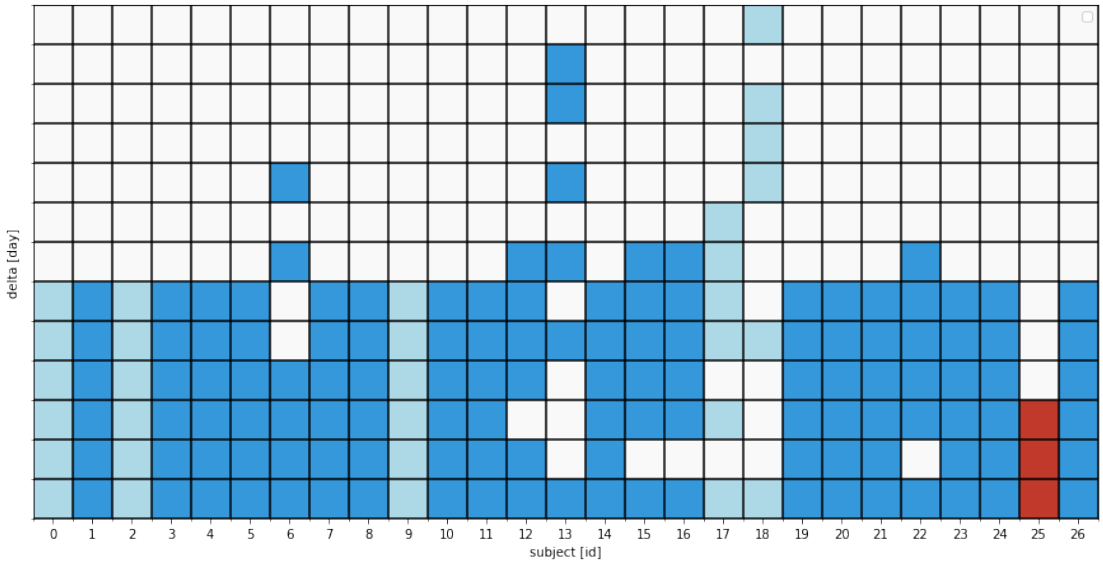

Figure 6.1: The figure shows on which days the participants took part in the field study. Light blue subjects which did not receive push notifications while subjects with dark blue marks had received push notifications. Red rectangles indicates a dropped out subject.

In regards to the participation during the study, 26 subjects managed to finish all laboratory and field study trials while 1 subject dropped out in the field study. Figure .5 shows on

which day individual participants performed trails. 4 participants did not accept push notification and were therefore not reminded on a daily basis.

In total over 46.000 taps were generated in the whole data acquisition phase. To be more precise, approx. 25.000 taps were collected on the 5x4 grid, over 15.000 taps on the 4x3 grid and more than 5.000 taps on the 2x2 grid.

The devices used to obtain the tap information were 12 (45%) Apple iPhone 6s, 10 (37%) iPhone 6 and the least common device was the iPhone 7 with 5 (18%).

In order to visualize the data collected, a t-Distributed Stochastic Neighbor Embedding (t-SNE) embedding is shown in figure 6.2 where the 230-dimensional feature vector has been reduced to 2 dimensions. Furthermore, a plot showing the interpolated gyroscope and accelerometer signals acquired during a single tap generation trial is to be found in the Appendix.

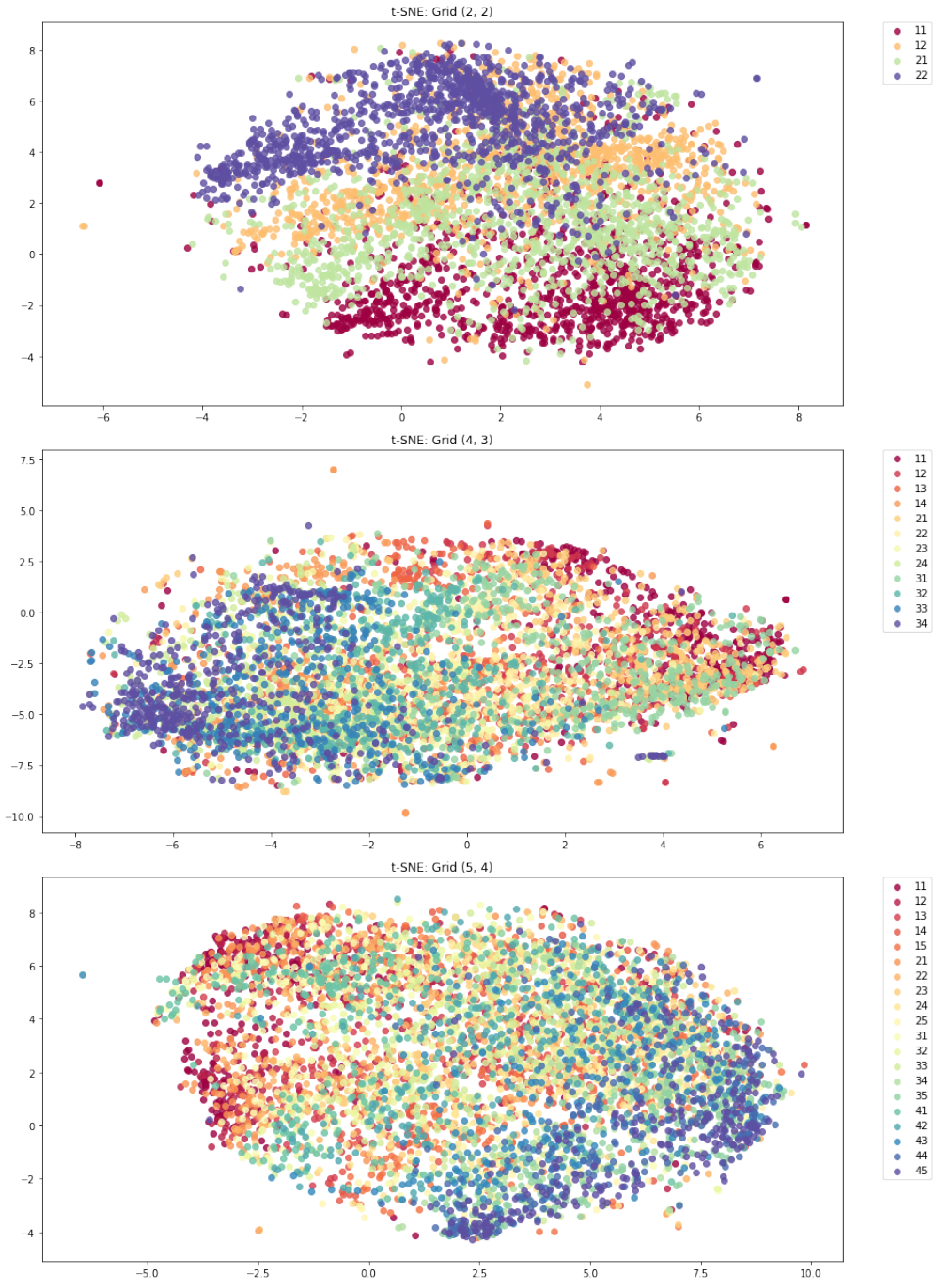

Figure 6.2: Visualization of 230-dimensional feature vectors reduced to 2 dimensions. The used t-SNE dimensionality reduction technique is unsupervised, thus it does not consider the labels during the optimization

6.2 Laboratory and Field Comparison

For the analysis between the controlled and the uncontrolled environment, a subset of the overall training material has been filtered based on the environment, grid size and the mobile device. After performing a grid search of the hyperparameter space for the SVM and, likewise, for the ANN, a 10-fold cross validation has been performed on the best of the two estimators yielding the highest mean accuracy.

2x2 Grid

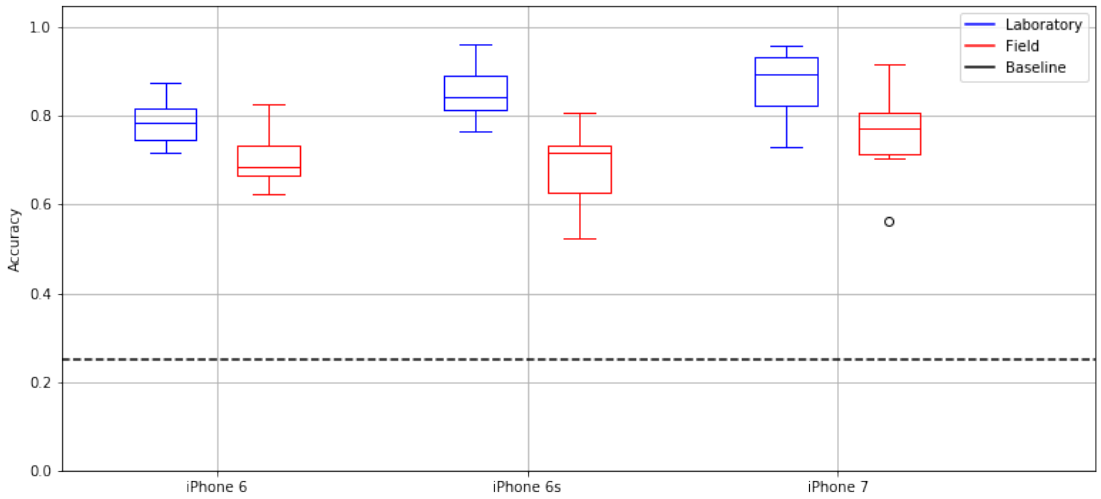

Figure 6.3: The figure shows the tap inference accuracies for the 2x2 grid of the 10-fold cross-validation. The measured classification accuracies are above the guessing probability of $\frac{1}{4} = 25\%$.

For the 2x2 grid, the results show mean accuracy measures in range 0.79 to 0.87 for the laboratory environment and ranges 0.68 to 0.76 for the field environment. Furthermore, the iPhone 7 data scores highest with a mean accuracy score of the 0.85 for this particular classification problem.

Moreover, across all devices, the results list that the mean accuracies for the field data are always lower compared to the laboratory data. The fact that the classification measures for both environments differ significantly is confirmed by a Wilcoxon signed-rank test yielding that the fold accuracies in the laboratory were significantly higher than the fold accuracies in the field environment $Z = 5$, $p < 0.05$.

		Accuracy				
Device	**Environment**	mean	min	max	std	**Classifier**
iPhone 6	Laboratory	0.79	0.72	0.87	0.05	SVM
	Field	0.71	0.62	0.83	0.06	SVM
iPhone 6s	Laboratory	0.85	0.77	0.96	0.06	SVM
	Field	0.68	0.52	0.81	0.09	SVM
iPhone 7	Laboratory	0.87	0.73	0.96	0.08	SVM
	Field	0.76	0.56	0.92	0.09	SVM

Table 6.1: Classification results for the 2x2 tapping grid. Notably, the SVM outperforms the ANN for this task.

4x3 Grid

For the 4x3 grid, the analysis shows mean accuracy scores in range 0.46 to 0.59 for the laboratory environment and ranges 0.40 to 0.47 for the field environment. Furthermore, the iPhone 7 data scores highest with a mean accuracy score of the 0.59 for this 12-class problem.

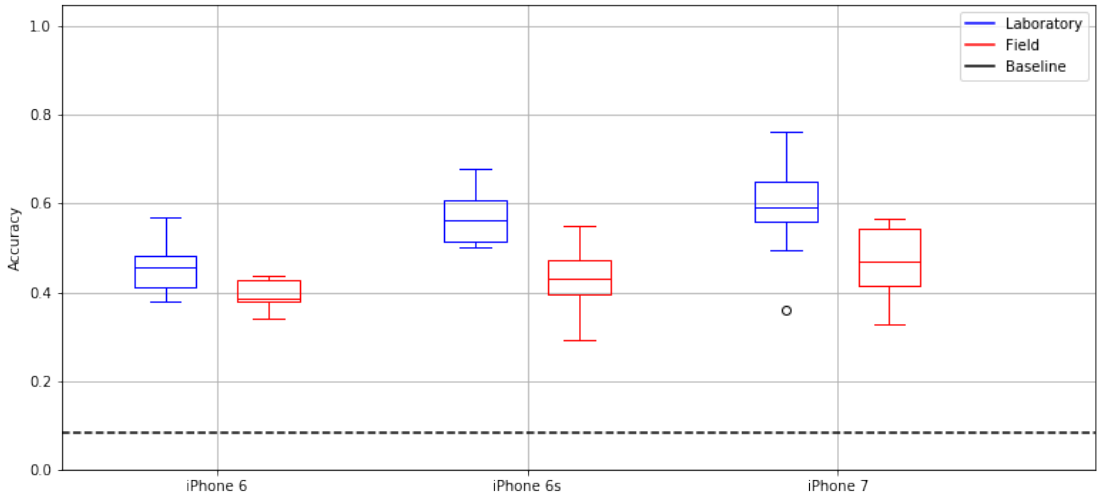

Figure 6.4: The figure shows the tap inference accuracies for the 4x3 grid of the 10-fold cross-validation. The measured inference accuracies are above the probability baseline of guessing ($\frac{1}{12} = 8.33\%$ for the 12 distinguishable buttons).

A performed Wilcoxon signed-rank test shows that the classification results for both environments differ significantly. The fold accuracies in the laboratory were statistically higher than the fold accuracies in the field environment $Z = 19$, $p < 0.05$.

		Accuracy				
Device	**Environment**	mean	min	max	std	**Classifier**
iPhone 6	Laboratory	0.46	0.38	0.57	0.06	SVM
	Field	0.40	0.34	0.44	0.03	SVM
iPhone 6s	Laboratory	0.57	0.50	0.68	0.06	SVM
	Field	0.43	0.29	0.55	0.07	ANN
iPhone 7	Laboratory	0.59	0.36	0.76	0.11	SVM
	Field	0.47	0.33	0.57	0.08	SVM

Table 6.2: Classification results for the 4x3 tapping grid.

5x4 Grid

For the grid with 20 distinguishable areas the inference accuracies measures range from 0.35 to 0.43 for the laboratory and 0.28 to 0.32 for the field data, respectively. Aligning with the previous results, the iPhone 7 shows highest mean accuracy scores of 0.43 for the data collected in the laboratory.

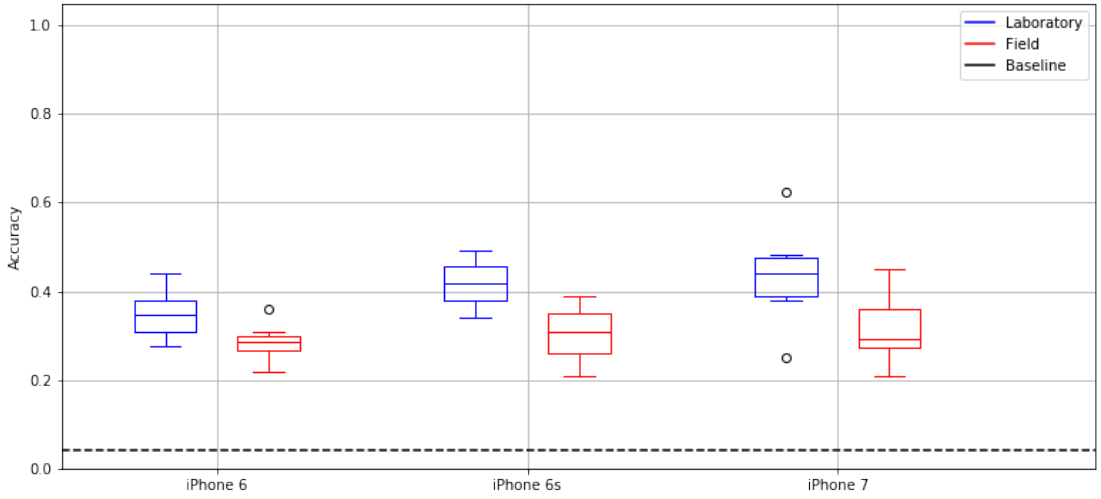

Figure 6.5: The figure shows the tap inference accuracies for the 5x4 grid of the 10-fold cross-validation. Results show that all inference accuracies are above the baseline of $\frac{1}{20} = 5\%$ for this classification problem.

Furthermore, The Wilcoxon signed-rank test shows that the classification results for both environments alter significantly. The fold accuracies in the laboratory were significantly higher than the fold accuracies in the field environment $Z = 12$, $p < 0.05$.

		Accuracy				
Device	**Environment**	mean	min	max	std	**Classifier**
iPhone 6	Lab	0.35	0.28	0.44	0.05	ANN
	Field	0.28	0.22	0.36	0.04	ANN
iPhone 6s	Lab	0.42	0.34	0.49	0.05	SVM
	Field	0.31	0.21	0.39	0.06	ANN
iPhone 7	Lab	0.43	0.25	0.62	0.09	SVM
	Field	0.32	0.21	0.45	0.08	SVM

Table 6.3: Classification results for the 5x4 tapping grid.

6.3 Input Modalities Comparison

As for the comparison between controlled and uncontrolled environments, the same classification experiment was performed to detect differences in the predictive models between the two input modalities: Index finger and thumb. As subjects were free to decide which input modality to use during the field study, the sample size has been adjusted in order to train each classifier with the same amount of training material.

Classifier trained for the individual grid sizes show similar results. For this reason, only the results for the 5x4 grid will be shown here. The results for the other grid sizes are listed in the Appendix.

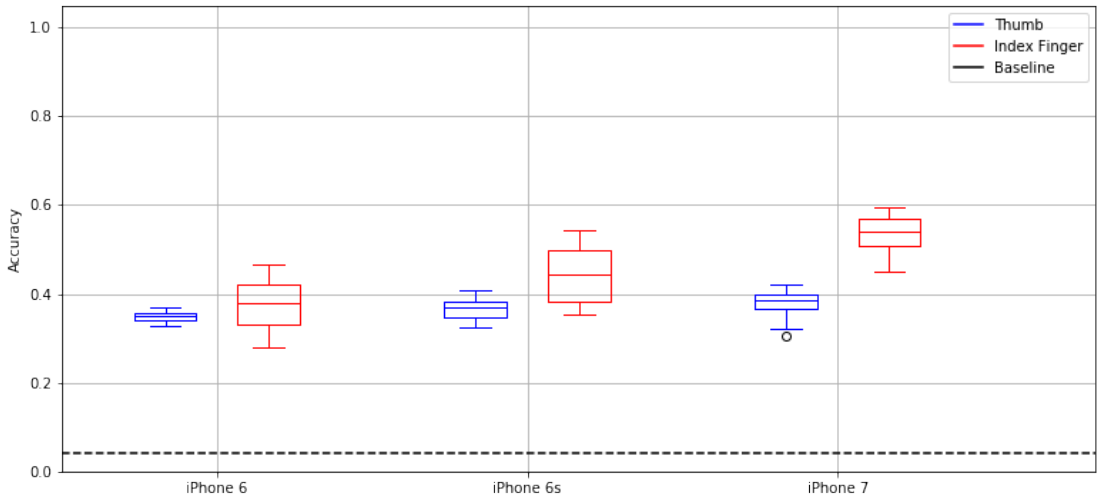

Figure 6.6: The figure shows the tap inference accuracies for the 5x4 grid of the 10-fold cross-validation.

For the 5x4 grid, the results show that across all devices the mean inference accuracies

for estimators trained with the thumb taps were lower compared to the classifiers trained with data containing index finger taps. For the iPhone 7, the estimator yields a mean accuracy of 0.54 for index finger samples compared to 0.38 for data representing the thumb as input modality. The same results apply for the other tested devices. A Wilcoxon signed-rank test shows that the classification results for both input modalities differ significantly. The fold accuracies on thumb data were statistically lower than the fold accuracies on index finger data $Z = 29$, $p < 0.05$.

		Accuracy				
Device	**Input Modality**	mean	min	max	std	**Classifier**
iPhone 6	Index	0.38	0.28	0.47	0.06	ANN
	Thumb	0.35	0.33	0.37	0.01	ANN
iPhone 6s	Index	0.44	0.35	0.54	0.06	SVM
	Thumb	0.37	0.33	0.41	0.02	ANN
iPhone 7	Index	0.54	0.45	0.59	0.04	SVM
	Thumb	0.38	0.30	0.42	0.04	ANN

Table 6.4: Classification results for the 5x4 tapping grid for both input modalities: thumb and index finger.

6.4 Body Posture Comparison

For the comparison between the two body postures (sitting and standing) the overall training material was filtered based on the device and body posture the user had while tapping. Furthermore, only index finger taps are considered in this experiment. As for the comparison of input modalities, the amount of training material was balanced.

		Accuracy				
Device	**Input Modality**	mean	min	max	std	**Classifier**
iPhone 6	Standing	0.30	0.19	0.45	0.08	SVM
	Sitting	0.35	0.29	0.41	0.03	ANN
iPhone 6s	Standing	0.37	0.34	0.42	0.03	SVM
	Sitting	0.47	0.34	0.61	0.08	SVM
iPhone 7	Sitting	0.58	0.44	0.82	0.11	SVM
	Standing	0.43	0.35	0.52	0.05	ANN

Table 6.5: Classification results for the 5x4 tapping grid for both body postures: sitting and standing.

Only the 5x4 grid is presented here, as similar results are found for the other grid sizes (see Appendix). The findings show that the mean accuracies between the two body modal-

ities differ (See figure 6.7). This is also indicated in a Wilcoxon signed-rank test showing that the classification results for both factors differed significantly Z = 31, p > 0.05.

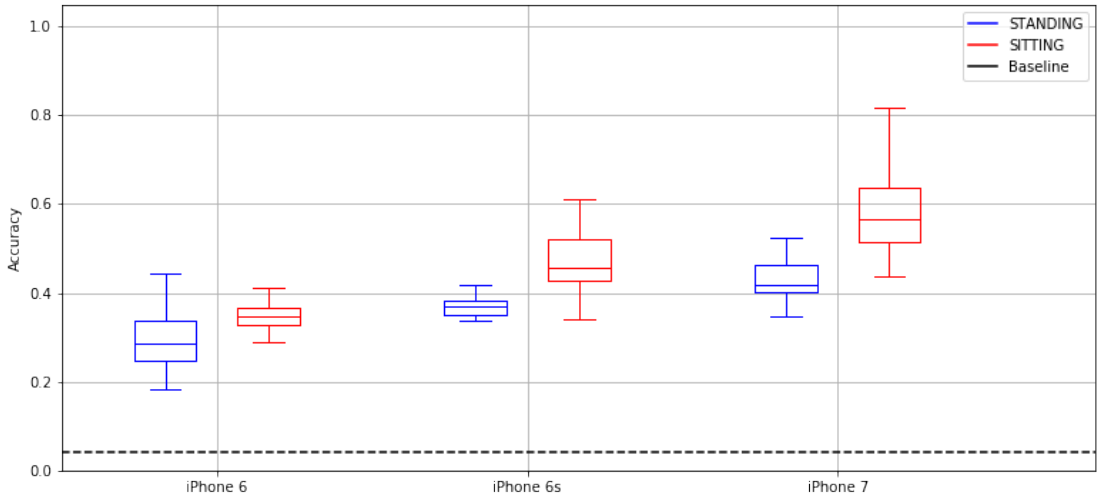

Figure 6.7: The figure shows the tap inference accuracies for the 5x4 grid of the 10-fold cross-validation.

6.5 Cross User Experiment

In order to determine if it is possible to train a classifier with data from a set of users to infer taps from people not involved in the training phase, a cross user experiment was performed. Here, a SVM classifier was trained for each subject with laboratory data from 4 randomly selected subjects sharing the same device. The subject's field data was then tested on the classifier and the accuracy was measured.

For each subject, the accuracy results for the 5x4 grid are displayed in in the bar chart 6.8. It can be seen that for user 86 the prediction accuracy was at minimum value of 0.09 whereas for user 87 the overall maximal value of 0.34 could be measured. The mean score measured was 0.25.

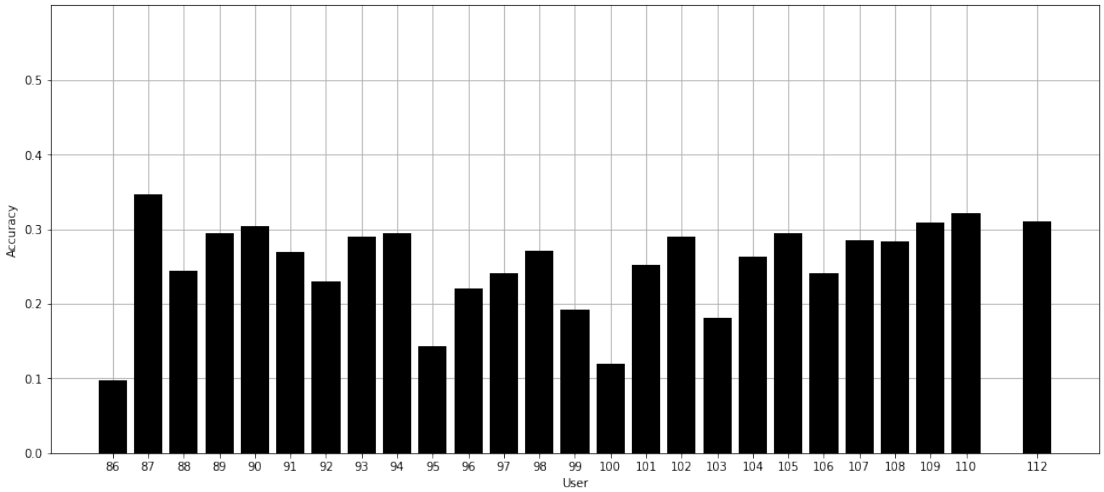

Figure 6.8: Results of the cross user experiment on the 5x4 grid.

7 Discussion

For the hypothesis tests, the assumptions made in section 5.1 will be either rejected or approved based on the results observed in the previous chapter.

H.1 The environment of recorded sensory data has an effect on the prediction accuracy.
H1.1: The prediction accuracy for a classifier trained with the data in the laboratory environment will score higher than one trained with data collected in the field.

Both assumptions can be approved as the results yield a significant difference between the performance measures of the estimators for both environments.

Across all available grid sizes the analysis shows that tap location inference is reasonably possible in the field as well as in the laboratory environment. Yet, for the field environment, an accuracy drop of approximately 20% was measured. Moreover, the results hint that PINs could be obtained in a real-world attack as the granularity of inferable locations is sufficient to project a PIN input mask on the 4x3 grid. Compared to previously proposed inference systems, the system presented in this work yields lower prediction accuracies than *TapPrints* [35]. However, as the scope of this work is to highlight the difference between both environments and not to display an upper bound to what is feasible, by interpreting the results, it is indicated that tap inference in the field is considerably more difficult.

Since device motion sensors are capable of capturing the slightest device vibrations, a vibrant environment or activity, one to which subjects where exposed during the field acquisition, is presumably prone to polluting the sensor signals with increased noise. This noise can distort the tap information encoded in the sensor signals aggravating clear predictions of the tap locations. Consequently, as subjects where free to perform tap generation trails where and how they wanted, this freedom is reflected in the recorded data sets with increased variability negatively impacting the classification accuracies.

When comparing the tested devices, the analysis has shown that the iPhone 7 taps could be predicted with higher accuracies compared to the iPhone 6 and iPhone 6s, respectively. During the feature extraction, it was observed that the iPhone 6 and the iPhone 6s have generated taps that were partially below the sampling rate of 100Hz, the rate initially defined in the TapSensing application. It is assumed that this is caused by high CPU loads on the devices. As a high CPU load causes the sampling rate to drop, due to lower resolution signals a decrease in estimator performance can be explained.

H2: The input modality has an effect on the prediction accuracy.
H2.2: The prediction accuracy for a classifier trained with index finger tap data will score higher than one trained with thumb tap data.

The analysis has shown that classification results of the computed models, when comparing the input modalities, differed significantly. As the index finger taps could be predicted at higher measures compared to the thumb taps, both hypothesis can be approved.

This outcome can be explained by comparing the motion of the individual input modalities. When a user taps the device with the index finger, the striking force of the finger hits the smartphone screen causing a shift towards the z-axis. As the other hand is used as a support, the applied force is partially resisted stopping the device from tilting. In contrast, when a user taps with the thumb, the striking force causes the device to rotate as the device is held in the same hand. This rotation causes a higher variance in the recorded data which results in an inferior predictability.

H3: The body posture has an effect on the prediction accuracy.
H3.1: The prediction accuracy for a classifier trained with taps where a user sat will score higher than one trained with taps where a user stood.

The results have shown that the difference in classification measures for both body postures, sitting and standing, differed significantly. The classification for sitting data yielded higher accuracies when compared to the standing data sets. Due to this finding, both assumptions can be approved.

The analysis indicates that the body posture poses an important influence factor on the variability in the motion data collected. This result can be explained based on two assumptions. Firstly, it is likely that subjects used their device while walking during the field study which poses a source for increased noise. Secondly, during the data acquisition in

the laboratory environment, it is known that subject did not walk while tapping the device. As this data was also contained in the training examples, it is assumed that standing on the spot also enables the user to make slight body movements which can effect the variability of the recorded samples.

Lastly, in the cross user experiment, it has been shown that it is possible to predict tap locations of users in the field that where not involved in the training process. This shows that in a real-world scenario an attacker could compute a model trained on several test users and, consequently, make predictions on unseen users in the field. However, as the accuracies measured ranged from 0.9 to a 0.34 for the individual users tested, the varying results indicate that the cross user inference can yield unreliable predictions. These findings align with previous results [35], as varying cross user accuracies were likewise to be found.

With the overall findings in this work, it has been shown that the performance of a tap inference system is strongly influenced by various sources of data variability. Consequently, if an inference system was to be deployed for a real-world attack, it would have to overcome the user switching input modalities, changing body postures and a potential increase in environmental noise from the user's current location in the field. As the impact of the environment was not modelled in related experiments [35, 11, 43], the proclaimed security threat of tap location inference has to be reassessed taking the variance of field data into account. However, as I believe that the performance gap between the field and laboratory environment could be bridged with appropriate filtering techniques or the design of more resilient features, the security threat of motion sensor emanation is yet prevalent.

8 Conclusion

In this thesis, *TapSensing* was presented, a data acquisition system that collects touchscreen tap event information with corresponding accelerometer and gyroscope readings. Having performing a data acquisition study with 27 subjects and 3 different iPhone models, a total of 45,000 labeled taps could be acquired from a laboratory and the field environment. After having performed a feature extraction on the acquired sensor recordings, several machine learning classifiers have been trained and compared in order to determine if the tap location inference is feasible for the field environment and secondly, to identify the sources of variability in the collected data. Furthermore, a real-world attack scenario has been evaluated where it has been tested if the user's field taps can be predicted based on a classifier trained on laboratory data from a different set of users.

The overall findings have shown that tap location inference is generally possible for data acquired in the field, however, with a performance reduction of approximately 20% when comparing both environments. Furthermore, it has been shown that the performance of the inference is dependant on the body posture and the input modality used to perform taps as these pose sources for an increased variability in the motion data. Lastly, it has been shown that it is possible to predict tap locations of users in the field that where not involved in the training process increasing the threat posed by motion sensor emanations. As the tap inference has now been shown on a more realistic data set and by aligning with the previous experiments [11, 35, 43], I hope that these findings furthermore raise the awareness of potential eavesdropping attacks due to non-restricted motion sensor access.

8.1 Further Outlook

It this work, it was identified that the field environment bears a potential source of variability in the motion data resulting in a general decrease of the predictability of tap locations. It is assumed that is due to an increase in noise originating from the user activity or vibrant surrounding. For future studies, it could be investigated if applying appropriate filtering on the sensor data could mitigate this "field effect" whereas a second option would be to design resilient features. However, as hand-crafting such features requires

high domain knowledge, convolution neural networks could be used to automatically extract features instead.

Convolution neural networks have shown to achieve high accuracies solving the Human Activity Recognition (HAR) problem [53] in which accelerometer signals are used to predict which activity the smartphone user currently has. As the gyroscope and accelerometer signals could be encoded as a single matrix, the convolution network is able to apply convolution filters on the input to automate the feature extraction process. This approach could not only be resilient against environmental noise but could also achieve higher accuracies than the currently proposed methods.

List of Acronyms

3GPP	3rd Generation Partnership Project
AJAX	Asynchronous JavaScript and XML
API	Application Programming Interface
AS	Application Server
CSCF	Call Session Control Function
CSS	Cascading Stylesheets
DHTML	Dynamic HTML
DOM	Document Object Model
FOKUS	Fraunhofer Institut fuer offene Kommunikationssysteme
GUI	Graphical User Interface
GPS	Global Positioning System
GSM	Global System for Mobile Communication
HTML	Hypertext Markup Language
HSS	Home Subscriber Server
HTTP	Hypertext Transfer Protocol
I-CSCF	Interrogating-Call Session Control Function
IETF	Internet Engineering Task Force
IM	Instant Messaging
IMS	IP Multimedia Subsystem
IP	Internet Protocol
J2ME	Java Micro Edition
JDK	Java Developer Kit
JRE	Java Runtime Environment
JSON	JavaScript Object Notation
JSR	Java Specification Request
JVM	Java Virtual Machine
NGN	Next Generation Network
OMA	Open Mobile Alliance
P-CSCF	Proxy-Call Session Control Function
PDA	Personal Digital Assistant
PEEM	Policy Evaluation, Enforcement and Management
QoS	Quality of Service

S-CSCF	Serving-Call Session Control Function
SDK	Software Developer Kit
SDP	Session Description Protocol
SIP	Session Initiation Protocol
SMS	Short Message Service
SMSC	Short Message Service Center
SOAP	Simple Object Access Protocol
SWF	Shockwave Flash
SWT	Standard Widget Toolkit
TCP	Transmission Control Protocol
Telco API	Telecommunication API
TLS	Transport Layer Security
UMTS	Universal Mobile Telecommunication System
URI	Uniform Resource Identifier
VoIP	Voice over Internet Protocol
W3C	World Wide Web Consortium
WSDL	Web Service Description Language
XCAP	XML Configuration Access Protocol
XDMS	XML Document Management Server
XML	Extensible Markup Language

Annex

Results

Data Aquisition

Input Modality

2x2 grid

For the comparison between input modalities, we find that for the 2x2 grid, all mean accuracies across devices were higher for index finger tap data compared to classifiers trained with the thumb tap data.

A Wilcoxon signed-rank test shows that the classification results for both input modalities differ significantly. The fold accuracies on thumb data were statistically lower than the fold accuracies on index finger data $Z = 27$, $p < 0.05$.

		Accuracy				
Device	**Input Modality**	mean	min	max	std	**Classifier**
iPhone 6	Index	0.82	0.72	0.95	0.07	SVM
	Thumb	0.79	0.71	0.87	0.05	SVM
iPhone 6s	Index	0.86	0.80	0.90	0.03	SVM
	Thumb	0.78	0.74	0.83	0.03	SVM
iPhone 7	Index	0.91	0.75	1.00	0.07	SVM
	Thumb	0.80	0.73	0.92	0.05	SVM

Table .1: Classification results for the 2x2 tapping grid for both input modalities: thumb and index finger.

4x3 grid

For the 4x3, all mean accuracies across devices were higher for index finger tap data compared to classifiers trained with the thumb tap data. The greatest difference in inference

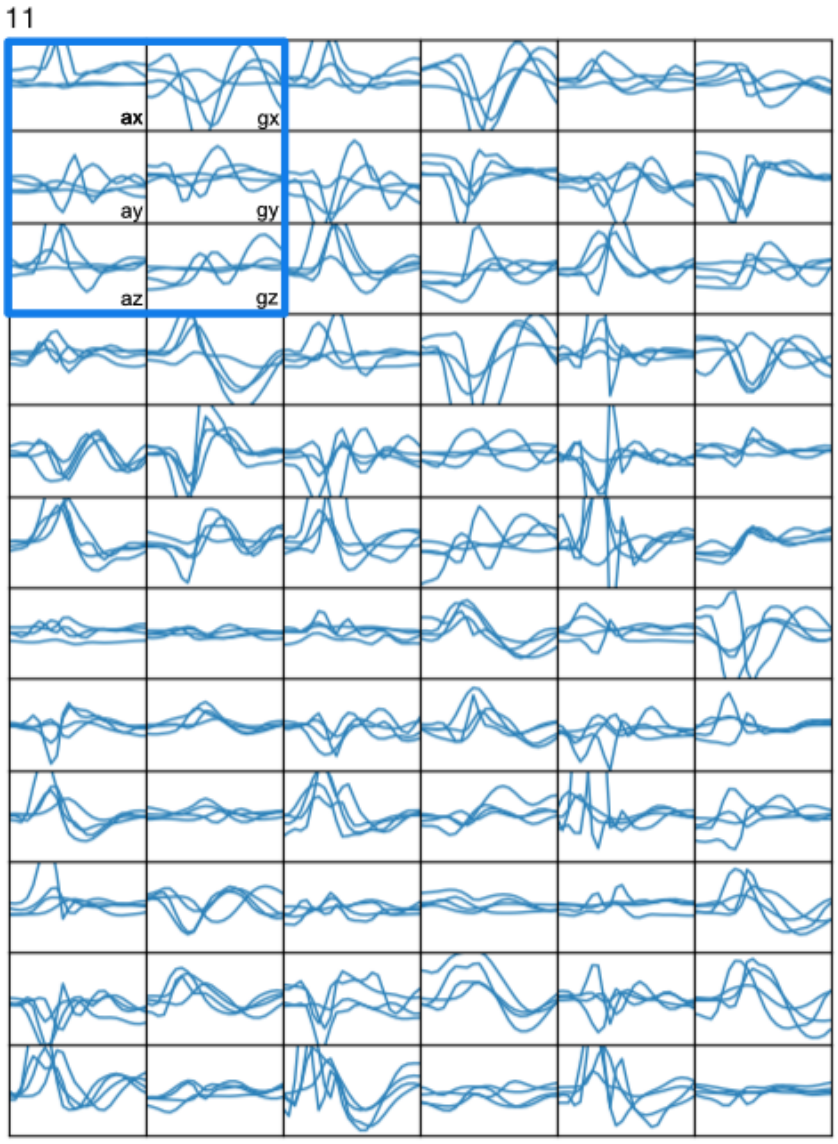

Figure .1: The visualization shows the collected gyroscope and accelerometer components for the 4x3 grid. In the top left corner the grid class 11 is shown which corresponds to the top left corner of the mobile device.

accuracies are to found on the iPhone 7 where 0.7 (+/- 0.08) for index finger records and 0.52(+/- 0.03) for thumb taps.

A Wilcoxon signed-rank test shows that the classification results for both input modalities differ significantly. The fold accuracies on thumb data were statistically lower than

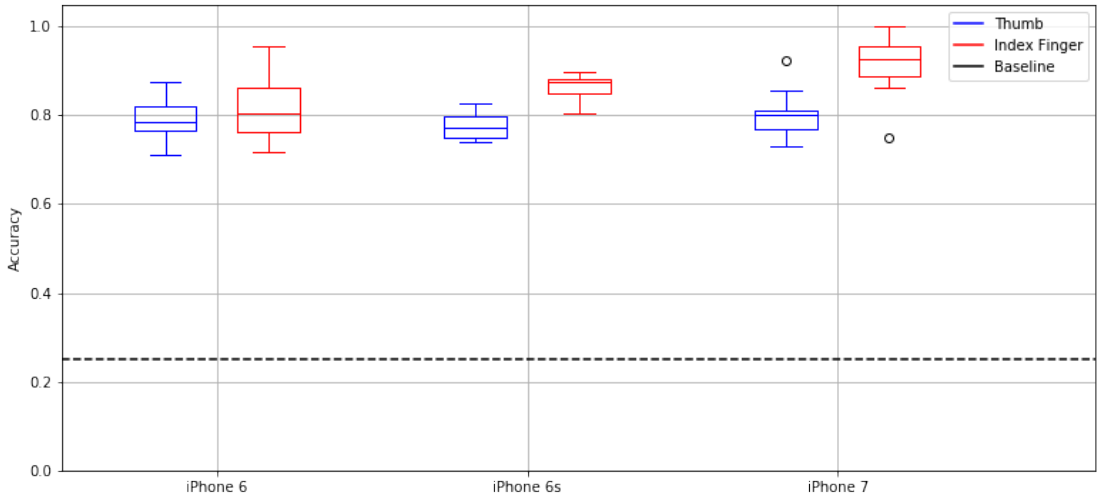

Figure .2: The figure shows the tap inference accuracies for the 2x2 grid of the 10-fold cross-validation.

the fold accuracies on index finger data $Z = 64.0$, $p < 0.05$.

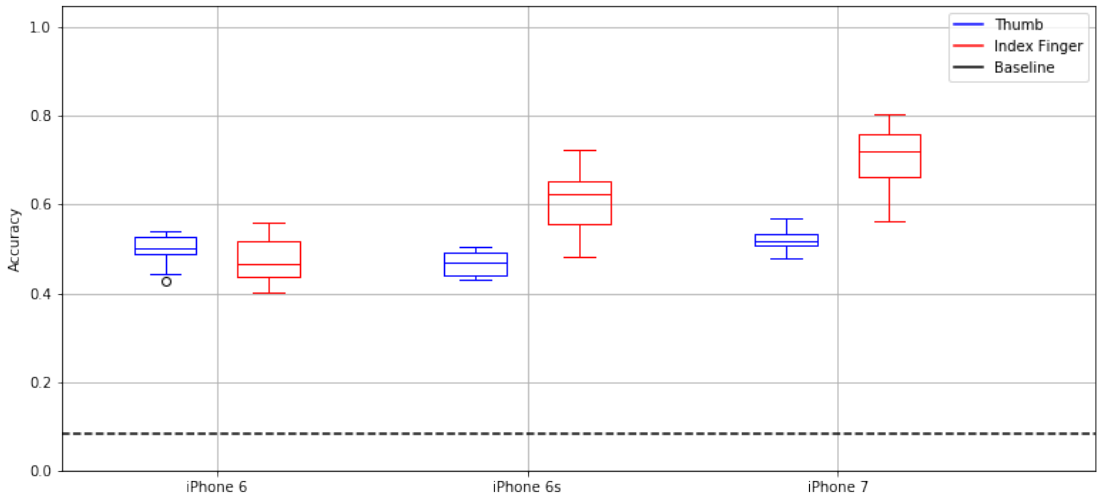

Figure .3: The figure shows the tap inference accuracies for the 4x3 grid of the 10-fold cross-validation.

		Accuracy				
Device	**Input Modality**	mean	min	max	std	**Classifier**
iPhone 6	Index	0.48	0.40	0.56	0.06	SVM
	Thumb	0.50	0.43	0.54	0.04	SVM
iPhone 6s	Index	0.61	0.48	0.72	0.07	SVM
	Thumb	0.47	0.43	0.51	0.03	SVM
iPhone 7	Index	0.70	0.56	0.80	0.08	SVM
	Thumb	0.52	0.48	0.57	0.03	SVM

Table .2: Classification results for the 4x3 tapping grid for both input modalities: thumb and index finger.

Body Posture

2x2

For the 2x2, all mean accuracies across devices were higher for sitting data compared to classifiers trained with the standing data. A Wilcoxon signed-rank test shows that the classification results for both body postures differ significantly. The fold accuracies on standing data were statistically lower than the fold accuracies on sitting data $Z = 55.0$, $p < 0.05$.

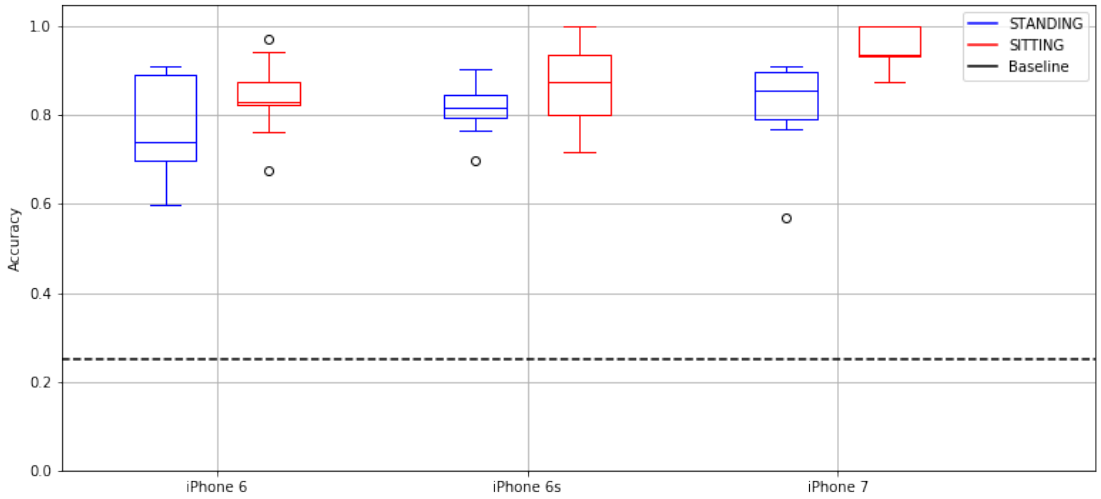

Figure .4: The figure shows the tap inference accuracies for the 4x3 grid of the 10-fold cross-validation.

		Accuracy				
Device	**Input Modality**	mean	min	max	std	**Classifier**
iPhone 6	Sitting	0.84	0.68	0.97	0.08	SVM
	Standing	0.80	0.63	0.94	0.10	SVM
iPhone 6s	Sitting	0.87	0.72	1.00	0.09	SVM
	Standing	0.85	0.73	0.93	0.05	SVM
iPhone 7	Sitting	0.95	0.88	1.00	0.04	SVM
	Standing	0.86	0.60	0.94	0.10	SVM

Table .3: Classification results for the 2x2 tapping grid for both input modalities: thumb and index finger.

4x3

For the 2x2, all mean accuracies across devices were higher for sitting data compared to classifiers trained with the standing data. A Wilcoxon signed-rank test shows that the classification results for both body postures differ significantly. The fold accuracies on standing data were statistically lower than the fold accuracies on sitting data $Z = 71.0$, $p < 0.05$.

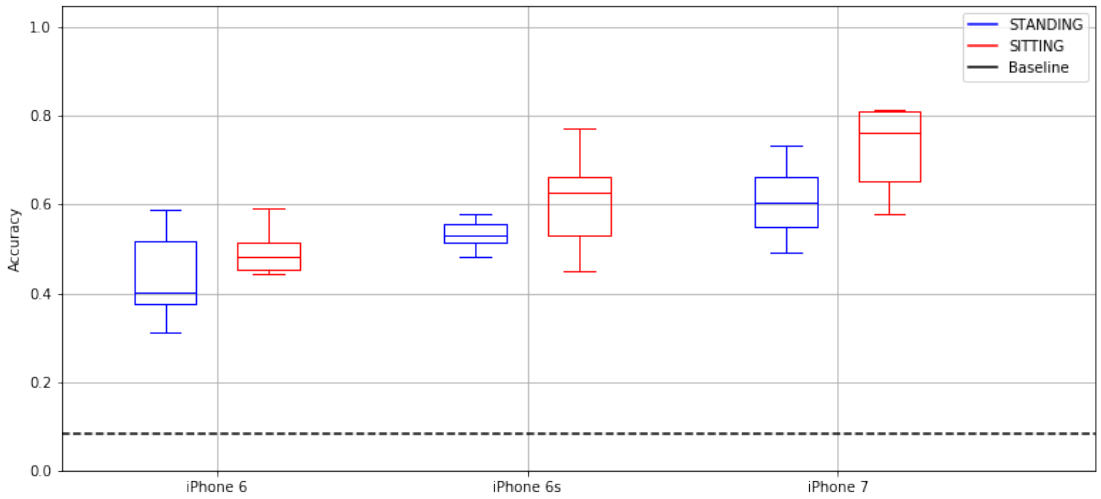

Figure .5: The figure shows the tap inference accuracies for the 4x3 grid of the 10-fold cross-validation.

		Accuracy				
Device	**Input Modality**	mean	min	max	std	**Classifier**
iPhone 6	Sitting	0.49	0.44	0.59	0.05	SVM
	Standing	0.47	0.34	0.62	0.08	SVM
iPhone 7	Sitting	0.73	0.58	0.81	0.08	SVM
	Standing	0.64	0.52	0.76	0.08	SVM
iPhone 6s	Sitting	0.61	0.45	0.77	0.11	SVM
	Standing	0.56	0.51	0.61	0.03	SVM

Table .4: Classification results for the 4x3 tapping grid for both input modalities: thumb and index finger.