Universität Zürich[UZH]

Geographisches Institut

Labelling Tappigraphy: A Ground Truth Study to Understand Mobile Map App Usage in Daily Life

GEO 511 Master's Thesis

Author: Oliva Schilling, 19-704-212

Supervised by: Dr. Tumasch Reichenbacher, Donatella Zingaro

Faculty representative: Prof. Dr. Sara Irina Fabrikant

23.10.2025

Acknowledgment

I wish to express my gratitude to my supervisors, Tumasch Reichenbacher and Donatella Zingaro, for their guidance, support and valuable feedback throughout the process of writing this thesis.

Further, I would like to thank Enea Ceolini for sharing his expertise in tappigraphy, machine learning and data analysis with patience and humour.

I would also like to thank all my participants for making this data collection possible.

I am grateful to Carmen Kull, Cassandra Zanivan and Nathan De Witte for proofreading the manuscript, providing helpful feedback and their warm support in general.

Finally, I want to thank my friends and family for their encouragement, and ongoing support throughout this journey.

Abstract

Mobile map applications have evolved into interactive tools offering functionalities beyond navigation. However, the context in which these apps are used is barely researched. This gap is caused by the limited availability of privacy-conscious data, which accurately reflects real-world mobile map app usage.

To address this, a new method called tappigraphy has recently been introduced to map app research. Tappigraphy only collects timestamps of taps and the app used, without capturing any other sensitive information. Therefore, by itself, tappigraphy does not reveal the specific tasks users perform within map apps. This thesis aims to address this limitation by collecting and labelling Google Maps ground truth data. This data was used to train a supervised machine learning classifier, which was then applied to a real-world tappigraphy dataset.

The ground truth tappigraphy data was collected in a controlled lab environment. Participants performed predefined tasks based on the previous research of Savino et al. (2021), allowing direct comparison to that study. The classification successfully replicated known usage patterns, especially when run solely on Google Maps data. Applying the classifier to other app combinations, such as similar map apps and public transport apps, revealed distinct patterns, which reflect app functionalities and assumed user behaviour.

Concluding, this thesis demonstrates the potential of using supervised machine learning to enrich tappigraphy data for task-specific within-app usage analysis. This allows an improved understanding of mobile map app usage context, which can inform the design of context-aware map applications.

Keywords: Mobile Map App Usage, Tappigraphy, Use Context, Supervised Machine Learning, Explorative Data Analysis, Smartphone, Google Maps, Within App Usage

List of Acronyms

CDF	Cumulative Distribution Function
CV	Cross Validation
GIScience	Geographic Information Science
GT	Ground Truth
ITI	Inter-Touch Intervals
MoT	MapOnTap
RBF	Radial Basis Function
SMLC	Supervised Machine Learning Classifier
std	Standard Deviation
SVC	Support Vector Classifier
UZH	University of Zurich

Table of Contents

1 Introduction 1

1.1 Research Motivation 1

1.2 Objectives and Research Questions 3

2 Related Work 4

2.1 Real-World Mobile Map App Usage 4

2.1.1 The Study of Savino et al. (2021) 4

2.2 Context of Map App Usage 6

2.2.1 Activity 8

2.2.2 Environment 8

2.2.3 Map Apps 9

2.3 Tappigraphy in Map App Research 9

3 Methods 12

3.1 Features of Tappigraphy 14

3.2 Real-World Tappigraphy Data 15

3.3 Ground Truth Data Collection 15

3.3.1 Data Collection Setup 16

3.3.2 Context Variables and Task Design 16

3.3.3 Practical Considerations and Observations 18

3.4 Ground Truth Data Preprocessing 18

3.5 Ground Truth Data Analysis 20

3.6 Supervised Machine Learning Classifier 21

3.7 Choice of Machine Learning Classifier 22

3.8 Weighting and Upsampling 24

3.9 Running the Classifier 24

4 Results 26

4.1 Real-World Data Analysis 26

4.2 Ground Truth Data Analysis 27

4.2.1 Descriptive Statistics 27

4.2.2 Group Comparison 30

4.3 Performance of Different Machine Learning Algorithms 31

4.3.1 Performance of the Support Vector Classifier 33

4.4 Applying the Support Vector Classifier on Real-World Data 35
4.4.1 Google Maps 36
4.4.2 Similar Map Apps 40
4.4.3 Public Transport Apps 42
4.4.4 Communications Apps 44
4.4.5 Remaining Apps 46
4.5 Applying the Support Vector Classifier on Ground Truth Data 48
4.5.1 Combined Google Maps Sessions 48
4.5.2 SBB App 49
5 Discussion 51
5.1 Real-World Dataset 52
5.2 Ground Truth Dataset 52
5.2.1 Ground Truth Data Collection 54
5.3 Performance of Different Machine Learning Classifiers 55
5.3.1 Performance of the Support Vector Classifier 56
5.3.2 Choice of Input Features 56
5.4 Applying the Support Vector Classifier 58
5.4.1 Google Maps 60
5.4.2 Similar Map Apps 62
5.4.3 Public Transport Apps 63
5.4.4 Communication and Remaining Apps 64
5.5 Limitations and Future Research 65
6 Conclusion 68
References 70
Appendix 76
Personal Declaration 81

Figures

Figure 1: Share of time spent in each interaction state within one app session, averaged for all users (Savino et al., 2021) 5

Figure 2: The four dimensions of map use context (Griffin et al. 2017) 6

Figure 3: Stages of the methodology implemented in this thesis for task-based real-world map app usage analysis 13

Figure 4: Tappigraphy only records taps within phone sessions (Reichenbacher et al. 2022) 14

Figure 5: The supervised machine learning classifier process from training to predicting 21

Figure 6: Most used map and travel apps by tap count 26

Figure 8: Distribution of log-transformed ITIs of Google Maps ground truth data 28

Figure 9: Distribution of log-transformed taps per session of Google Maps ground truth data 28

Figure 7: Distribution of log-transformed ITI differences of Google Maps ground truth data 28

Figure 10: Tapping distribution in milliseconds (ms) since session start for different interaction states, sample data of participant 6 in Google Maps 29

Figure 11: Correlation matrix heatmap of the five chosen metrics in the Google Maps GT data 29

Figure 12: Distribution of ITI log p25 values depending on the interaction state and app, GT data 30

Figure 13: Cumulative distribution function (CDF) of ITI log p25 per app, GT data 30

Figure 14: Distribution of ITI log p25 per participant for SBB and Google Maps, GT data 31

Figure 15: Permutation feature importances of all eight features without hyperparameter tuning (SVC RBF) 34

Figure 16: Permutation feature importance for the five main features, with hyperparameter tuning (SVC RBF) 34

Figure 17: Normalised confusion matrix of the SVC RBF, depicting the share of (mis)classified sessions per interaction state for GT Google Maps data 35

Figure 18: Prediction confidence of SVC RBF per interaction state for real-world Google Maps data 37

Figure 19: Number of app sessions classified per interaction state by SVC RBF for real-world Google Maps data 37

Figure 20: Share of total app sessions duration per interaction state for real-world Google Maps data (SVC RBF) 38

Figure 21: Average time share of phone session duration per predicted state for real-world Google Maps data (SVC RBF) 39

Figure 22: Prediction confidence of SVC RBF per interaction state for real-world data of similar map apps 40

Figure 23: Number of app sessions classified per interaction state by SVC RBF for real-world data of similar map apps 41

Figure 24: Share of total app sessions duration per interaction state for real-world data of similar map apps (SVC RBF) 41

Figure 25: Prediction confidence of SVC RBF per interaction state for real-world data of Public Transport Apps 42

Figure 26: Number of app sessions classified per interaction state by SVC RBF for Public Transport Apps 43

Figure 27: Share of total app sessions duration per interaction state for real-world data of Public Transport Apps (SVC RBF) 43

Figure 28: Prediction confidence of SVC RBF per interaction state for real-world data of Communication Apps 44

Figure 29: Number of app sessions classified per interaction state by SVC RBF for Communication Apps 45

Figure 30: Share of total app sessions duration per interaction state for real-world data of Communication Apps (SVC RBF) 45

Figure 31: Prediction confidence of SVC RBF per interaction state for real-world data of the Remaining Apps 46

Figure 32: Number of app sessions classified per interaction state by SVC RBF for the Remaining Apps 47

Figure 33: Share of total app sessions duration per interaction state for real-world data of the Remaining Apps (SVC RBF) 47

Figure 34: Share of prediction confidence per state depending on the combined state input, GT Google Maps data (SVC RBF) 48

Figure 35: Number of app sessions classified per interaction state by SVC RBF for the SBB GT data 49

Figure 36: Share of total app sessions duration per interaction state for the SBB GT data (SVC RBF) 50

Figure 37: Normalised confusion matrix of the SVC RBF, depicting the share of (mis)classified SBB GT sessions per interaction state 50

Tables

Table 1: Framework on map app usage, summarising common activities, environments and map apps 7

Table 2: Examples of tasks per interaction state in the GT data collection 17

Table 3: Explanation of the names of the eight available metrics 20

Table 4: Average session length and taps per session by interaction state, Google Maps GT data 27

Table 5: CV and balanced accuracy scores of the different SMLCs using different feature combinations 32

Table 6: Differences in CV and balanced accuracy scores depending on the weighting method (SVC RBF classifier on GT Google Maps data) 33

Table 7: Average prediction confidence of the SVC RBF for different app combinations 36

Table 8: Average session length and taps per session per interaction state, Google Maps real-world data 36

Table 9: Ten most common interaction state sequences within one phone session, real-world Google Maps data (SVC RBV) 39

Table 10: Task list for group A with corresponding interaction state (for group B the same blocks were used in a different order) 76

Table 11: Names of the apps included in the app combination public transport 80

1 Introduction

1.1 Research Motivation

Understanding the usage of mobile map apps in everyday life is a complex task due to the dynamic and context-dependent nature of portable devices. Mobile map apps, such as Google Maps and the SBB app (i.e. Swiss railway), are nowadays used for more than just navigation. They have evolved into applications that help explore surroundings, check place details, and facilitate spatial decisions (Savino et al., 2021). Due to its multifunctionality, Google Maps has been ranked as the fifth most-used app overall and the leading map application (Böhmer et al., 2011; Savino et al., 2021).

Despite the popularity of these apps, the understanding of how, why, when, and where people interact with mobile map apps in real-world contexts remains very limited (Zingaro et al., 2023). Most studies rely on controlled experiments for design decisions (Roth et al., 2017), which fail to capture real-world usage. In addition, existing studies often focus narrowly on geographic context retrieval, overlooking critical factors such as the user's contextual needs, activities and goals (Reichenbacher, 2004). Filling this gap requires access to in-app usage data that accurately reflects real-world behaviour. Meaning the data must have high ecological validity, which refers to how well research findings generalise to real-world settings. However, obtaining such data poses significant challenges due to strict privacy regulations and user confidentiality concerns that limit availability and accessibility (Savino et al., 2021)

To address these challenges, the mobile app MapOnTap (MoT) was developed by the Geographic Information Visualisation and Analysis group in collaboration with the University of Zurich (UZH) spin-off company QuantActions sàrl. MoT captures only the sequence of touchscreen interactions (i.e., taps) within apps used in the foreground when the phone is unlocked and stops when the phone is locked again (Zingaro & Reichenbacher, 2022). Furthermore, it gathers GPS data on the phone's location. The abstract tap data is called tappigraphy and is later described in detail (see Section 2.3).

While MoT was explicitly developed to address the challenges of studying mobile map application use in real-world contexts, providing insights into user behaviour solely relying on a limited set of information, such as taps, has significant limitations. This includes the lack of background information about users, the inability to track the specific content accessed within the app during taps, and the lack of details about the app's interface or design at the time of interaction. Hence, MoT by itself can't provide insight into specific usage conditions, such as the user's goal.

These limitations pose challenges in assessing and improving the usability of mobile map applications, which often include diverse and complex functionalities. It is important to emphasise that these limitations stem from the necessary trade-off between collecting information at a high level of granularity and protecting user privacy. Addressing this balance is critical for maintaining ethical standards for data collection and privacy (Reichenbacher et al., 2022).

This thesis aims to mitigate these limitations by collecting ground truth (GT) tappigraphy data. GT refers to data collected under controlled conditions where contexts are predefined and observable (Muller et al., 2021). The GT data collection is designed according to a literature-based framework. This involves performing specific common tasks in a controlled lab environment using Google Maps or the SBB app. These apps were chosen for their high frequency of use and relevance to navigation and travel tasks, particularly for activities like commuting (Xu et al., 2011). This GT dataset is then labelled based on predefined interaction states and used to train a machine learning classifier. This classifier is applied to the real-world MoT data collected in uncontrolled environments to classify tapping patterns into interaction states. This combined approach enables a deeper understanding of mobile map interactions. It improves the interpretability of tappigraphy data and allows the identification of distinct behaviours.

1.2 Objectives and Research Questions

This thesis aims to create a framework for analysing task-specific usage of mobile map applications, overcoming limitations in current data collection methods. The methodology includes collecting and labelling GT data in a controlled lab environment. Next, a supervised machine learning classifier (SMLC) is trained and optimised to distinguish different map app usage interaction states based solely on Google Maps tappigraphy data. The trained classifier is then applied to real-world data, and its results are compared to findings from prior research to validate this approach. Finally, the classifier is applied across different app combinations for exploratory analysis and comparison of usage patterns in relation to app functionalities and user goals.

Ultimately, this framework strengthens the ability to analyse and categorise mobile map usage captured by tappigraphy data, linking it to use cases. These insights are intended to improve the design for more intuitive, adaptive and user-centred mobile map applications, aligning functionality with users' cognitive and practical needs.

Research Questions

1. What are the common map app usage activities and environments found in literature?
2. Which features are suitable for labelling activity-specific tapping patterns?
3. To what extent do the tapping patterns observed in ground truth data align with those in real-world tappigraphy data?

2 Related Work

This section summarises related work in the field of real-world mobile map app usage, with an emphasis on usage context. In the first part, research on real-world mobile map app usage is discussed, focusing on the study of Savino et al. (2021). The second part reviews the context of map app usage, focusing on the three context dimensions: activity environment and map app. In the last part, previous research using tappigraphy to analyse map app usage is discussed.

2.1 Real-World Mobile Map App Usage

Despite the widespread use of mobile map applications, research into their real-world usage remains very limited. Existing studies, such as Carrascal & Church (2015), have explored general within-app usage. However, there is a lack of detailed, real-world data analysis specific to mobile map apps (Savino et al., 2021). Understanding how users interact with these apps is essential for improving usability and creating context-aware designs (Huang et al., 2018; Otebolaku & Andrade, 2016; Zingaro et al., 2023).

2.1.1 The Study of Savino et al. (2021)

Savino et al. (2021) conducted the most in-depth study of real-world Google Maps usage. This study guided the data collection design of this thesis and was later used to assess the classifier's results. Therefore, it is presented here in detail.

Savino et al. (2021) conducted a study on 28 locals in Bremen, Germany, over the course of four weeks. This data was collected before 2020 and was therefore unaffected by COVID-19 in Germany ("COVID-19 Pandemic in Germany," 2025). The researchers used a wrapped version of Google Maps (i.e. MapRecorder), which means users interacted with the Google Maps website as usual while usage data was collected unobtrusively. Therefore, they were able to record not only the touchscreen interactions within the app, but also the content of those interactions.

Their findings include the definition of four interaction states that characterise typical user activities in Google Maps:

- **Search (S):** Searching for destinations or points of interest in the search bar.
- **Place (P):** Viewing place details, such as reviews or pictures.
- **Direction (D):** Obtaining route information for a selected mode of transport.
- **Map (M):** Exploring, zooming, or panning on the map.

The results presented by Savino et al. (2021) are based on app sessions, defined as the time from when Google Maps is opened until it is closed again, without considering the phone session level. The study reported an average app session duration of 65 seconds, which aligns with the 71.56 seconds found by Böhmer et al. (2011). Furthermore, the time share of each interaction state per app session was analysed (see Figure 1).

Figure 1: Share of time spent in each interaction state within one app session, averaged for all users (Savino et al., 2021)

Savino et al. (2021) found Map to be the most prominent state with 67.5%, followed by Direction (21.1%), next Place (8.2%) and Search (3.2%). Furthermore, the study described that the most common sequences of interaction states within an app session were MSPD (Map-Search-Place-Direction, 12.1%) and MD (13%), showing that a minimum of 25.1% of the app sessions ended in getting directions.

While Savino et al. (2021) focus on a specific usage context, namely how Google Maps is used in daily life, they do not discuss the overall concept and importance of map app usage context. However, understanding context as a concept is crucial to support more context-aware design in the future.

2.2 Context of Map App Usage

Context-aware design is particularly important for mobile map apps since their usage often occurs in dynamic environments (Griffin et al., 2024). Additionally, environmental factors like time of day, weekday and location significantly influence app usage patterns (Do et al., 2011). However, there is a lack of research on the influence of task dependency (Griffin et al., 2024). To address real-world usage of mobile map apps, a clear definition of context is essential. This thesis applies the definition of context as the interaction of four dimensions of Griffin et al. (2017):

- **User**: Characteristics such as age or role (e.g., tourist vs. local) that shape interaction patterns
- **Map**: Variations in map design or functionality. In this thesis, this is called **Map App** due to the focus on mobile applications.
- **Activity**: Specific tasks users perform
- **Environment**: External factors that influence usage (e.g., time, location, distractions)

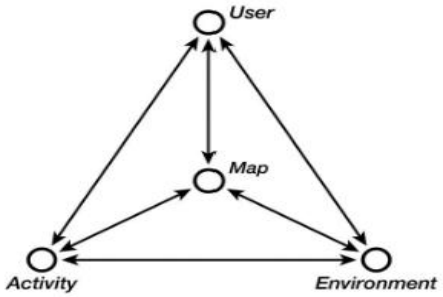

Figure 2: The four dimensions of map use context (Griffin et al. 2017)

Based on this definition, this thesis focuses on activity, while the environment and the map app are kept constant. The user group researched in the GT data is 22-39 years old and balanced in gender, its is therefore relatively homogenous.

The relevant literature on the context of map app usage was summarised in Table 1, which presents a refined framework focusing on the context dimensions activity, environment and map app. This framework aims to provide a foundation for future research by identifying critical aspects of mobile map app usage in existing literature.

Table 1: Framework on map app usage, summarising common activities, environments and map apps

Source	Activity	Environment	Map Apps
(Do et al., 2011)	Not addressed	- Home, work, and friend home (general phone usage) - Holiday, relaxing, restaurant, during transport - Natural environment	Not addressed
(Brown et al., 2013)	-Check the map while ordering a drink - Walk the blue dot to figure out the right direction - Walk to a sightseeing point	Non-workday in a city	Google Maps, Apple Maps (only iPhone users were researched)
(Yang et al., 2016)	Not addressed	On the move	App categories: music, web browsing, messaging, and map
(Tian et al., 2020)	-Regular info checking -Meal booking	-Morning 7-11:00 -Afternoon 15-21.00	App categories: Navigation, finance, weather, productivity, transport, food, communication App sequence: transport, communication, transport
(Fonseca et al., 2021)	-Finding the shortest route (42% of users). -Locating specific destinations like shops or restaurants (38%). -Collecting walking-related data (19%) (travel distance, travel time, optional routes, etc.).	- Map apps are predominantly used by younger adults and students (tech-savvy) -Tourists and commuters tend to use maps more than residents - Common usage frequency: weekly or occasionally - Usage in Porto is higher than in Bologna (place dependency)	95% Google Maps Other: Apple Maps, Here WeGo, and Maps.Me.
(Savino et al., 2021)	- Map (panning and zooming, ~65% of usage time). - Search for specific places (rather than entities) - Direction - Place details (menus, reviews)	- Everyday exploration and navigation (locals, radius: 8km) - Tourists: walking directions (radius: 400m)	Google Maps
(Li, Xia, et al., 2022)	- Listen to music - Check e-mails - Use navigation apps	Commute/Transportation (mainly during morning and evening rush hour), on weekends, more people use transport in the afternoon than in the evening	App sequence: Music, email, navigation
(Zingaro et al., 2023)	Not addressed	Two types of users: - Home behaviour – high map app usage close to home (<13km) - Travel behaviour – scattered map app usage within a big radius (0-200km)	App categories: 25 apps in the category “Maps and Navigation” 63 apps in “Travel and Local”

2.2.1 Activity

The reviewed studies confirm that map apps are used for more than just wayfinding. In several studies, the participants used the map to explore their surroundings, search for places, get place information and check possible travel options. In terms of navigation, the search for specific destinations and finding the (shortest) route was a common activity (Brown et al., 2013; Fonseca et al., 2021; Savino et al., 2021).

Overall, there are very few studies that analyse map app usage behaviour at the activity level. Most studies analyse phone usage on the app level (e.g., app sequences or share of usage time) or consider other factors like the environment. This is due to the lack of fine-grained usage data (Zingaro & Reichenbacher, 2022). This gap is addressed in this thesis using tappigraphy data and GT data, focusing on the interaction states as defined by Savino et al. (2021).

2.2.2 Environment

The environment includes external factors that influence map usage, such as time, location, and distractions. The reviewed papers found that usage-time patterns often align with daily commuting, with a common-use location being in transport (on the move). This implies tasks like checking for public transport or finding the best routes based on current traffic (Li, Li, et al., 2022; Tian et al., 2020).

In contrast, common-use locations also include typical leisure time activities like holidays, dining at restaurants, or relaxing. Most common places for map app usage are often first-time visits (Do et al., 2011). These environmental settings imply different use cases for mobile map apps, reflecting a division between work-related needs and leisure time activities, particularly during evenings and weekends (Do et al., 2011; Li, Li, et al., 2022). Distractions such as noise or crowds also impact map usage (Kim et al., 2019). However, these factors have received limited attention in existing research.

In this thesis's GT data collection, the environment will be held constant. The data collection takes place in a lab room without distractions.

2.2.3 Map Apps

Most existing studies analyse phone usage at a broader level, focusing on patterns across app categories rather than isolating map apps. These studies reveal frequent co-occurrences, such as map apps being used alongside transport, communication, or music apps, suggesting that map app usage often takes place in a multi-app interaction context (Li, Li, et al., 2022; Tian et al., 2020; Yang et al., 2016).

When focusing solely on maps, Zingaro et al. (2023) reported a high number of different navigation and travel apps being used. Nevertheless, among studies that specifically investigated mobile map applications, most focused exclusively on Google Maps, as it consistently emerged as the most widely used navigation app (Böhmer et al., 2011). The most used travel app in the real-world MoT dataset is the SBB app (*MapOnTap*, 2021), the Swiss national railway app, highlighting its relevance in the context of data collected in Switzerland. Currently, there is no open data research on the use of the SBB app.

The study of Savino et al. (2021) focused exclusively on Google Maps, whereas users typically rely on multiple apps for navigation tasks (Tian et al., 2020). They are often used sequentially, like combining travel and navigation apps for daily commuting (Xu et al., 2011). Therefore, in this thesis GT data on both Google Maps and the SBB app was collected. However, there were some issues in the SBB GT data, and consequently, the focus of this thesis remains on Google Maps.

2.3 Tappigraphy in Map App Research

The framework presented in this thesis underlines that most of the existing research fails to investigate map app usage context in depth. This research gap can be addressed by using data with high ecological validity, collected with tappigraphy.

Tappigraphy is a method for unobtrusively and continuously recording and analysing touchscreen interactions on smartphones in real-world settings. By capturing each tap at millisecond resolution, tappigraphy enables detailed analysis of tapping behaviour over time.

Initially developed in neuroscience to study cognitive processing speed, sleeping patterns, and disease states, this method has recently been applied to other fields, including Geographic Information Science (GIScience) (Borger et al., 2019; Duckrow et al., 2021; Reichenbacher et al., 2022).

This approach has proven valuable for studying mobile applications, including map apps. By collecting ecologically valid data continuously over long periods of time without requiring direct interaction with participants, tappigraphy avoids observational bias and supports large sample sizes. Its key strength lies in its ability to provide high-resolution data about everyday interactions in real-world contexts, providing insights into app usage patterns and user behaviour in different scenarios (Zingaro et al., 2023). Additionally, it does not require personal information, such as gender or nationality (Reichenbacher et al., 2022). This makes it a privacy-conscious tool for ecological momentary assessment in GIScience research. Ecological momentary assessment is defined as repeated real-time sampling of participants' actions within their natural environment (Shiffman et al., 2008).

Using tappigraphy, it is further possible to record taps across multiple apps, making it a flexible tool for evaluating various applications and providing a broader understanding of user behaviour across different contexts and tasks (Zingaro et al., 2023).

Other studies using tappigraphy showed that map apps are sparsely used compared to other app categories. Furthermore, map app usage follows daily and weekly cycles differing from other temporal app usage patterns, indicating a specific temporal footprint surrounding map app usage (Reichenbacher et al., 2022). Zingaro et al. (2023) further found two distinctive tapping patterns: "home behaviour," characterised by high phone usage near the user's home, and "travel behaviour," defined by less interaction distributed over greater distances. For some participants, the interaction increased with distance from home due to increased map app usage (Zingaro et al., 2023). Additionally, Zingaro et al. (2024b) uncovered clusters of active (more engaged, e.g. typing) and passive (viewing, scrolling) interactions within apps, offering insights into in-app usage patterns. Their study shows that these interaction modes vary by app category.

Looking into user differences, Zingaro et al. (2024a) found that women, who tend to have higher spatial anxiety, often switch more frequently between app categories during phone sessions, while men, who tend to have better spatial orientation skills and less spatial anxiety, spent more time on map apps.

In map app research so far, only simple metrics of tappigraphy, such as tapping speed, session length and taps per session, were used (Reichenbacher et al., 2022; Zingaro et al., 2023). However, the study of Ceolini et al. (2022) found that the 25th percentile of the inter-touch intervals (ITI), the time passing between two taps, can be used to estimate tapping speed.

They argue that despite the various ways in which people use their phones, tappigraphy is a proxy for underlying cognitive processes. For example, they found that tapping speed declines with age, indicating slower cognitive processing.

Despite these findings, tappigraphy alone struggles to capture context-specific aspects of map app usage. Tappigraphy does not provide information on specific activities performed within apps. This limitation results in generalised insights, which highlight the need for GT data collection under controlled conditions as an additional data source.

3 Methods

In this section, the different steps for using tappigraphy for task-based real-world map app usage analysis are described; they are also visualised in Figure 3. First, the existing real-world tappigraphy dataset and related literature are analysed, which informs the GT data collection design. Next, the GT data is collected, pre-processed and labelled. Thereafter, the GT data is analysed, and possible classification metrics are calculated. Then different SMLCs are trained, assessed and fine-tuned with different feature combinations. Next, the chosen Support Vector Classifier (SVC) is trained on the upsampled GT data with the optimal features, resulting in a predictive model. This model is applied to the real-world data, predicting the expected interaction states. The prediction on Google Maps is compared to the findings of Savino et al. (2021) to evaluate if the classifier finds realistic patterns. In a final step, the classification of real-world data for different application combinations (map vs public transport vs other apps) is visualised for exploratory comparison.

The data analysis and classification are done in Python 3 using different Python and Jupyter Notebook files. The documentation and the code itself are available on Git.

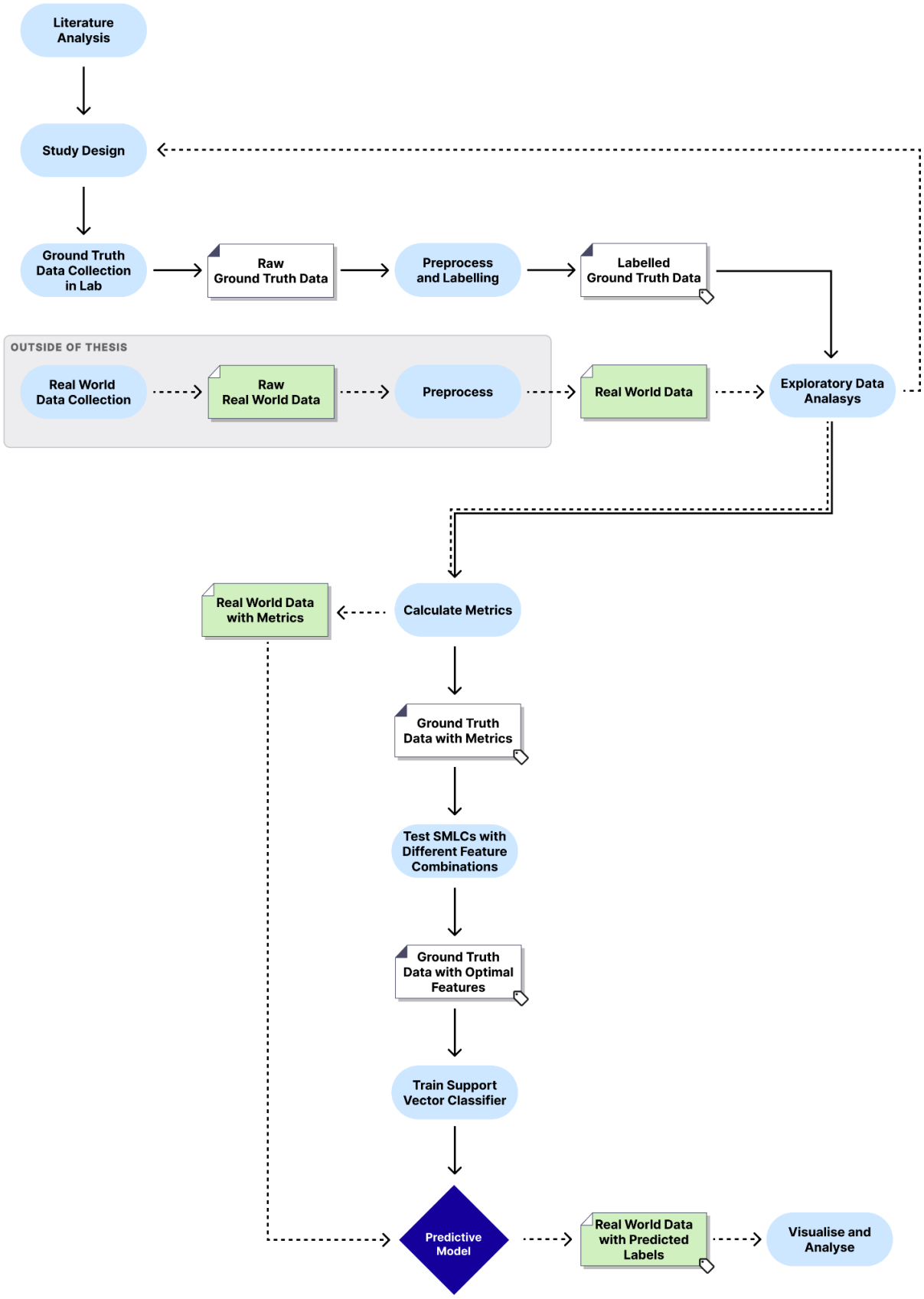

Figure 3: Stages of the methodology implemented in this thesis for task-based real-world map app usage analysis

3.1 Features of Tappigraphy

Tappigraphy unobtrusively records touchscreen interactions (taps) only when the phone is unlocked, capturing which taps were made in which apps and by which participant (only identified by their random ID). This results in data segmented into discrete phone sessions (see Figure 4) (Zingaro & Reichenbacher, 2022). One phone session can involve several app sessions, the time between opening and closing one app, the analysis in this thesis focuses on the app session level.

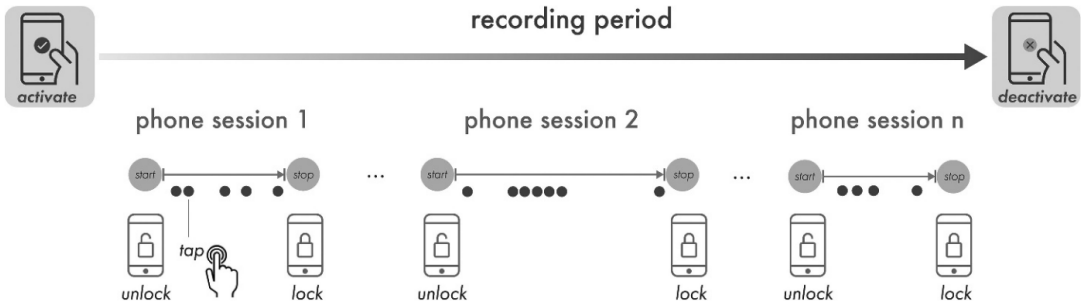

Figure 4: Tappigraphy only records taps within phone sessions (Reichenbacher et al. 2022)

Therefore, the metrics which can be derived from this type of data are all temporal and abstract. The three most intuitive metrics are app session length, number of taps per app session and taps per second (taps per app session divided by the app session length in seconds).

An additional metric is the inter-touch interval (ITI), which is simply the amount of time passing between two touches of the phone screen. Calculating all ITIs results in a list of ITI values for every app session. This list can then be analysed using statistical measures such as the 25th percentile, median and standard deviation. Summarising, ITIs capture the temporal dynamic of tapping sessions.

Another metric is the ITI difference, calculated by subtracting each ITI from the next, resulting in positive or negative values depending on whether tapping speed decreased or increased during a app session. This list of values for each app session can also be analysed using the aforementioned statistical measures.

Determining which of these metrics are suitable for labelling activity-specific tapping patterns was explored during SMLC training. Derived quantitative measures from the raw data are referred to as metrics. Only the metrics which will be used as input variables for the machine learning classifier are called features.

3.2 Real-World Tappigraphy Data

The real-world tappigraphy data analysed in this thesis was collected during the MoT study at the Geographic Information Visualization & Analysis group at UZH (*MapOnTap*, 2021). Anonymous participants downloaded the MoT app on their own smartphone and used their phones as usual for a minimum of two weeks. This resulted in a dataset of various app sessions, which are linked to phone sessions and participants by IDs. Each of these app sessions includes timestamps per tap, the app ID and location data; the latter was not used in this thesis.

To complement the literature framework and inform the data collection design, an analysis of the real-world dataset was conducted. This analysis is based on app sessions, and the real-world data was filtered to include only sessions from the app categories "Travel and Local" or "Maps and Navigation".

App usage was analysed by number of taps, number of sessions or total usage time, all of which consistently identified Google Maps and SBB as the two most used apps. This supported the decision to focus on these two apps for the GT data collection.

After this analysis, the real-world data was further pre-processed before the SMLC analysis. This involved calculating and log-transforming the required metrics (see Table 3) and dropping sessions with missing values.

3.3 Ground Truth Data Collection

To fill the gaps in task-specific map app usage research, GT data was collected in a controlled, lab-based setting. The design of the data collection was based on both pre-analysis of real-world tappigraphy data and insights from existing literature. Based on this analysis, the GT data collection was conducted in Google Maps and the SBB app.

To record tappigraphy data, the TapCounter app was used. This app functions the same as MoT, except it does not record sensor data such as GPS or accelerometer information. This was not relevant in a lab setting, and the TapCounter app had recently been updated. MoT, however, was facing technical issues at the time of the data collection, which further supported the decision to use TapCounter.

3.3.1 Data Collection Setup

The data collection for this thesis included 20 participants, who performed a series of predefined tasks using Google Maps and the SBB App on a provided smartphone (OnePlus 8T). It took place in a quiet lab room without interference or distractions. Participants were invited and instructed individually by the researcher.

Participants were recruited via mailing lists of GIScience master's students and PhD students at the Geography Institute of the UZH, as well as through the researcher's private network. Consequently, the sample was relatively homogenous. Most participants (95%) visited Zurich daily or weekly and were between 22 and 39 years old. Gender and smartphone operating systems (Android/iOS) were balanced across the sample. All participants were regular users of both the SBB app and Google Maps, which was an inclusion criterion to ensure familiarity with the app features.

Before starting the data collection, participants completed a brief questionnaire on demographics and app usage habits. Afterwards, they filled out the Santa Barbara sense-of-direction scale (Hegarty, 2002) and the spatial anxiety questionnaire (He & Hegarty, 2020). The participants app usage habits were compared to the given tasks in the data collection to analyse familiarity with the functionalities. Due to limited resources, neither the demographics nor the sense of direction or spatial anxiety scores were further analysed within this thesis. However, this data remains useful for future research.

3.3.2 Context Variables and Task Design

The variables for the data collection are derived from literature, data analysis, and the defined context framework. Thus, context is the interaction between the user, map app, environment, and activity as described by Griffin et al. (2017) (see Section 2.2).

- **Map App**: Controlled variable – only Google Maps and SBB, on the same Android device.
- **Environment**: Controlled variable – distraction-free lab room
- **User**: Partially controlled variable – participants are regular users of the two apps, aged 22-39 and balanced in gender
- **Activity**: Independent variable – specific tasks like searching for a destination

The dependent variable is the temporal distribution of taps (e.g. ITIs). To ensure generalizability and capture within-task variance, each task was repeated multiple times with varying destinations. This structured approach allows for the analysis of task-specific tapping patterns while maintaining consistency across other contextual factors.

The tasks were intended to reflect common Google Maps and SBB app usage scenarios (see Table 2). For Google Maps, the tasks were based on the four interaction states: Search, Place, Direction and Map as defined by (Savino et al., 2021). For the SBB app, Search and Map were kept, while Direction and Place were replaced by Buy and Check. The additional interaction states describe the buying of a ticket (Buy) and the checking of saved routes or previously bought tickets (Check). These additional interaction states for the SBB App were designed based on the researcher's experience and peer consultation.

Table 2: Examples of tasks per interaction state in the GT data collection

App	Interaction State	Example Task
Google Maps	Map	Explore Basel as if you are planning to have dinner there - what options are there
	Search	Search for the Zara in Basel
	Place	Look at the details - opening hours etc.
	Direction	Plan a route from Solothurn to Basel using public transport and choose a specific route
SBB	Map	Look at the map that shows you where to find the bus and orientate yourself (click Walk)
	Search	Look up a train connection from here to Solothurn, save the route
	Check	Check for valid tickets or abonnements, look at the ticket of today
	Buy	press ticket, choose return and 2nd class (do not actually buy it)

The data collection was divided into four blocks, with two blocks focusing on Google Maps and two on the SBB app. Per app, each block includes a series of the same predefined tasks, with task order varying to balance learning effects and decrease in focus. Participants were split into two groups (A and B), completing the blocks in different sequences. A break separated the two halves of the experiment, during which all saved routes in the SBB app were deleted while search suggestions in both apps were kept.

This captures variations in tapping behaviour related to searching with and without suggestions, as both are common usage in the real world. The task list for Group A is presented in the Appendix. The destinations to be looked up and explored included Zurich, Solothurn, and Basel. As expected, all participants were regularly in Zurich, but rarely or never in Solothurn and Basel. This introduces the within-task variance of exploring known versus unknown places.

3.3.3 Practical Considerations and Observations

While the setup was designed for consistency and control, and two test runs were done, the actual data collection process revealed several practical considerations. Half of the participants were not familiar with certain app functions, like saving a route in the SBB app. Therefore, the experiment setup was slightly changed from the second participant onward. Participants were now asked at the beginning if they knew all the functions and were able to try out unknown functions with hints to ensure clarity.

To minimise distraction, participants locked the phone after each task and unlocked it only once instructions for the next task were clear, although reminders were occasionally necessary. The researcher monitored tasks using screen recordings, the Tap Counter app, and manually noted timestamps, though these timestamps sometimes differed slightly from actual start times. This difference was later handled during data preprocessing. Rare incidents of app switching led to the exclusion of some phone sessions.

3.4 Ground Truth Data Preprocessing

After the GT data collection, the tap data, app data and metadata were retrieved from the QuantActions database connected to the Tapcounter app (*Taps.Ai*, n.d.). The parquet files were parsed to produce a pandas DataFrame using code provided and maintained by QuantActions via GitHub.

Each row of the resulting DataFrame represents one app session, defined as the period from unlocking to locking the device within a single application. The raw data includes the app session ID, phone session ID, participant ID, app name, app category, start and end time of the sessions and the timestamps for each tap, all at millisecond resolution. The data frame was filtered to include only sessions of either the SBB app or Google Maps, excluding any sessions with noise (e.g., sessions where participants left the app).

Furthermore, for each participant a file containing manually noted timestamps was collected. These files included the interaction state label for every task, the participant ID, experiment half and block number. All files were standardised and merged into a single dataset including all participants. This dataset was then combined with the tappigraphy data by matching session start times to the manually noted timestamps within a maximum deviation of five seconds. This threshold was established based on screen recording checks. This approach made it possible to assign the interaction state label to each app session of the tappigraphy data.

Any missing or inconsistent entries were verified against screen recordings and corrected or excluded where necessary. Later, several metrics were derived from this combined dataset, including app session length, number of taps per app session, taps per second, ITIs, and differences between consecutive ITIs.

During metrics calculation, it became clear that sessions with fewer than four taps lead to missing values for the summary statistics of the ITI differences. Since three taps result in only two ITIs and one ITI difference, from which, e.g. no standard deviation (std) can be derived. Most sessions of the classes Check and Buy had fewer than four taps, and temporal dynamics were not sufficiently captured in such short sessions. Thus, sessions with fewer than four taps were excluded from further analysis.

This led to the SBB app data only including two states with a sufficient amount of app sessions, making it unsuitable for training an SMLC specifically for the SBB app. The data was still analysed and used to explore the performance of the Google Maps classifier on public transport data.

The final metrics were log10-transformed, and summary statistics such as the 25th percentile of the ITIs were computed in accordance with prior tappigraphy research. The eight final metrics are listed in Table 3, along with an explanation of their names. The terms “log” and “log10” refer to the same transformation, where the logarithm was calculated using base 10.

Table 3: Explanation of the names of the eight available metrics

Metrics Name	Meaning
ITI_log_p25	25th percentile of log-transformed ITIs
taps_per_sec_log10	Log-transformed taps per second
SessionLength_sec_log10	Log-transformed session length (in seconds)
tapsSession_log10	Log-transformed taps per session
ITI_log_diff_std	Standard deviation of log-transformed ITI differences
ITI_log_diff_median	Median of log-transformed ITI differences
log_ITIs_std	Standard deviation of log-transformed ITIs
log_ITIs_median	Median of log-transformed ITIs

3.5 Ground Truth Data Analysis

After the GT data preprocessing, the goal was to determine which of the available metrics were suitable for labelling activity-specific tapping patterns. This analysis included descriptive statistics, metric-distribution analysis, metric-correlation checks, groupwise comparisons, and interpreting different data visualisations.

Feature distributions were visually inspected through histograms, cumulative distribution functions (CDFs) and plots of the temporal distribution of taps within a session. This allowed a comparison between apps and interaction states.

Each app session was treated as an individual observation in the dataset, enabling app session-level granularity for later classification modelling. Since the ITIs and the ITI differences are lists of values per session, they cannot be directly used as input features for machine learning. Instead, summary statistics like median, std and the 25th percentile were calculated. Thus, eight metrics were available to test as input features for the SMLC (see Table 3).

3.6 Supervised Machine Learning Classifier

Before explaining how the SMLC was built and which features were used, it is essential to clarify what an SMLC is and how it will be used in this thesis. Machine learning, a subfield of Artificial Intelligence, allows computers to generalise and behave more intelligently than, for example, a simple database would. A key task in machine learning is supervised classification, where a classifier is trained on labelled data to predict unknown labels in unlabelled datasets (Muhammad & Yan, 2015), as visualised in Figure 5.

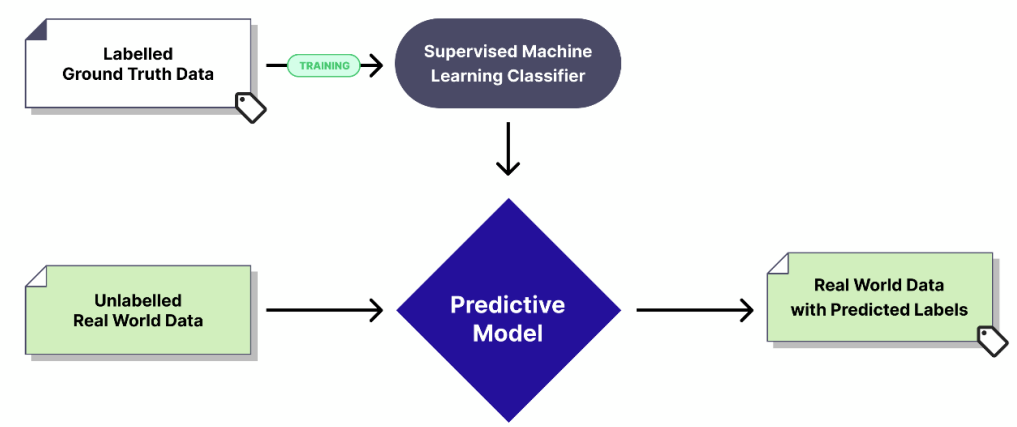

Figure 5: The supervised machine learning classifier process from training to predicting

In this thesis, GT tappigraphy data with labelled interaction states was collected to train an SMLC, which is then applied to predict interaction states in an unlabelled real-world tappigraphy dataset. The final goal is to analyse real-world data on activity-level, to understand what tasks people perform in mobile map apps.

Applying supervised classification to map app tappigraphy data is an exploratory approach. Therefore, the choice of the features, classifier and hyperparameters is based on trial and error rather than existing literature.

3.7 Choice of Machine Learning Classifier

One of the first steps in building an SMLC is the analysis of the labelled dataset (e.g. the GT data) to derive available metrics from which appropriate features are selected. The best-case scenario is to choose features based on expert suggestions. The other option is to analyse all available metrics, hoping to isolate all relevant features. This is less ideal and involves tedious preprocessing (Muhammad & Yan, 2015).

When analysing the GT dataset, all eight metrics available within the abstract tappigraphy data were analysed (see Table 3). The metric calculation was informed by Zingaro & Reichenbacher (2022) for session length, taps per session, and taps per second, and by Ceolini et al. (2022) for the ITIs, specifically the 25th percentile of the ITIs as an estimation of tapping speed. The ITI differences have not yet been used in another study. However, not all metrics were used for the final classifier. Instead, the input features were chosen by analysing data characteristics and using trial and error to evaluate which feature combinations yield the highest SMLC performance.

SMLC performance was assessed using cross-validation (CV) to evaluate the model's generalisation ability on unseen data. Specifically, the dataset was divided into 20 folds corresponding to the 20 participants, where each fold was used once as the validation set while the remaining 19 folds were used as training data. The average performance across all validation folds is the CV score, providing a robust estimate of expected model accuracy on new data (Pedregosa et al., 2011).

The balanced accuracy score was also inspected since the size of the classes (interaction states) was imbalanced. This score penalises a model if it performs well on bigger classes but poorly on smaller classes (Pedregosa et al., 2011).

Features were standardised where required by the classifier. Different feature combinations were tested, including using just one feature (ITI_log_p25), all eight features and reducing features step by step based on (permutation) feature importance, until the best CV score and balanced accuracy with a minimum number of features were achieved. The goal of minimising features is to avoid using features that barely contribute to model performance but carry the risk of introducing noise, therefore worsening the performance (Arinze, 2024).

Feature importance refers to the magnitude of the coefficient assigned to a feature and can be directly accessed in some SMLC, such as the Random Forest Classifier. In other models, like the Radial Basis Function (RBF) SVC, this is not possible. Instead, the permutation feature importance can be examined. It assesses the impact on model performance if one feature is randomly shuffled (Pedregosa et al., 2011).

The feature reduction process resulted in selecting five key features: ITI_log_p25, tapsSession_log10, log_ITIs_std, log_ITIs_median and the ITI_log_diff_std. The implications of this feature choice will be discussed in Section 5.3.2. Choice of Input Features

Selecting the best SMLC depends on the task. Therefore, it is common practice to try different SMLCs and assess their performance (Muhammad & Yan, 2015) in addition to analysing the data characteristics. In this thesis, Random Forest, Support Vector Machine, Gradient Boosting Classification and Logistic regression were tested on the GT dataset. They were chosen for being common SMLCs which can be used for the classification of discrete classes (GeeksforGeeks, 2025).

The SVC classifier was chosen due to having the highest performance scores, reliability and processing speed. SVCs have several advantages, namely being memory efficient, effective in high-dimensional space and versatile due to different available kernel functions (Muhammad & Yan, 2015). The classes (interaction states) in my GT dataset are not separable by a hyperplane, like most real-world problems. This issue can be solved by mapping the dataset into a higher-dimensional space and dividing it there. This is usually done using a RBF kernel (Eskandar, 2023).

Choosing the best kernel and hyperparameters is important for optimising the model fit and needs to be adjusted to the distribution of the classes in the dataset. It can be determined by using automated hyperparameter tuning (Muhammad & Yan, 2015). More specifically, grid search with cross-validation was applied to all evaluated machine learning classifiers, including the chosen SVC. This involved assessing combinations of kernel types (linear, RBF, polynomial) and associated hyperparameters, and selecting the hyperparameters that yielded the highest average CV score (Pedregosa et al., 2011). The best hyperparameters for the GT Google Maps dataset were C = 10, gamma = 0.01, and kernel = RBF.

3.8 Weighting and Upsampling

Most machine learning algorithms assume balanced classes, which can lead to a bias towards the majority class in the case of an imbalanced dataset. This bias can be minimised by using techniques like class weighting and upsampling. The choice of the technique depends on the use case, which is why it is recommended to try different approaches and assess which improves the performance and reliability of the classifier the most (Bakırarar & Elhan, 2023).

The GT dataset is imbalanced, due to the cutting of too short sessions and shortcomings of the data collection design. Therefore, upsampling and three weighting techniques, namely standard balanced class weighting, inverse number of samples and inverse square number of samples, were assessed in this thesis. The best results were achieved with upsampling; thus, it was used to prepare the dataset before training the final Google Maps classifier. Upsampling balances the class distribution by duplicating random minority class instances until classes are balanced (Murel, 2024).

3.9 Running the Classifier

After training and fine-tuning the classifier on the selected features of the upsampled Google Maps dataset, it was run on the real-world dataset. It was not trained on the SBB App data because most Check and Buy sessions were too short, resulting in mostly Search and Map sessions in that dataset.

Within the real-world dataset, the classifier was run on different apps and combinations thereof, namely Google Maps by itself, similar map apps, public transport apps, communication apps and remaining apps (not map, public transport or communication apps). The goal of the classifier is not to evaluate which app was used, as this information is provided in the tappigraphy dataset. Instead, the classifier was specifically trained to analyse interaction states in Google Maps.

The other app combinations were inspected to research if the resulting patterns in terms of prediction confidence per state and share of each interaction state are realistic and not, for example, the same for any (non-map) app combination. Furthermore, differences between various types of apps, such as map versus public transport apps, were investigated. The classifications of communication and remaining apps were conducted for exploratory interpretation, such as understanding what the interaction state Map could represent in a non-map app.

The average prediction confidence per interaction state for the different app groups was inspected to analyse how well the real-world data fits the defined classes. However, this is not an actual performance measure and there is no real-world test dataset to verify the classifier's performance on the unlabelled real-world dataset. Running the Google Maps classifier on the SBB GT dataset can be used as an indicator for how well the model performs on public transport apps. However, it is not an actual validation since the SBB data is GT data collected in a lab setting and not real-world data. The average prediction confidence per interaction state is a measure directly outputted by the SVC, based on Platt scaling (Pedregosa et al., 2011). While explaining this in detail goes beyond the scope of this thesis, it is important to clarify what this score implies. Since the classification involved four classes, a score of 0.25 is the chance level, indicating high uncertainty (a bad match), while a score of 1 is a perfect match, representing high certainty.

There are some app sessions which were classified with low prediction confidence. These sessions were removed from the classified data frames for all app combinations to prevent distortion effects in analysing the interaction state shares. A threshold of 0.55 was set at approximately the level of the 25^{th} percentile of the prediction confidence distribution of the least confident interaction state.

To inspect app (combination) differences, the share of the interaction states overall and per phone session was visualised. This comparison was based on the count of app sessions per state as well as the time spent in each state. The real-world Google Maps dataset was further analysed based on the sequence of interaction states within one phone session. This was determined by grouping app sessions by phone session ID and ordering them by session start time. The classification results on Google Maps were compared to existing research.

4 Results

This section presents the results of the different analysis steps described in the methods. First, the characteristics of the real-world dataset are summarised. Next, the collected GT dataset is analysed through descriptive statistics and group comparisons. Then, the performance of various machine learning classifiers using different input features and weighting methods is compared. Finally, the performance of the chosen SVC on different app combinations of the real-world dataset is assessed.

4.1 Real-World Data Analysis

The real-world dataset includes 51 participants and a total of 254’379 app sessions, of which 6’460 are Google Maps sessions. The average duration of a Google Maps app session is 65 seconds. Overall, only 4% (10’512) of app sessions in the real-world tappigraphy dataset are Google Maps or SBB usage. The study period ranged from a minimum of two weeks to over a year, from February 2021 to April 2022.

For the analysis, the app sessions were filtered by the categories “Travel and Local” and “Maps and Navigation”, resulting in 101 unique apps. Figure 6 presents the ten most frequently used map and travel apps by tap count. Analyses of session count and total usage time similarly identified Google Maps and SBB as the most used apps.

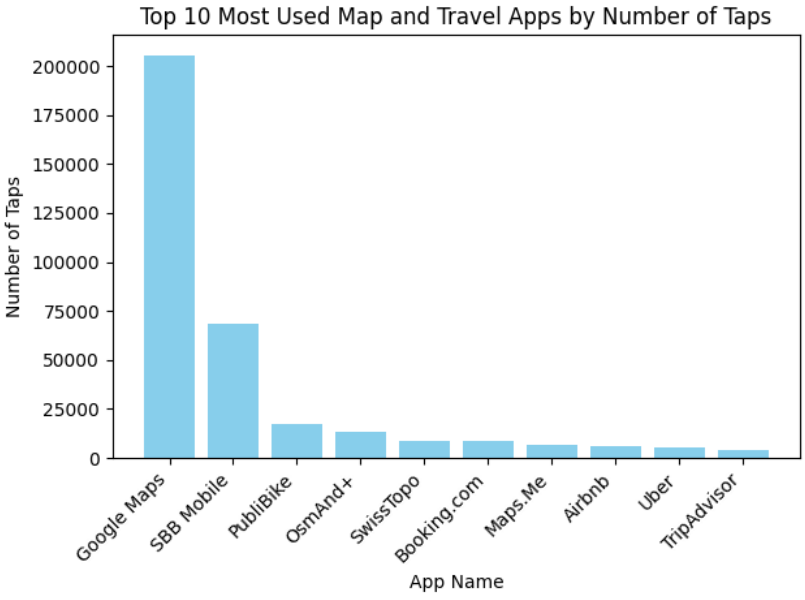

Figure 6: Most used map and travel apps by tap count

4.2 Ground Truth Data Analysis

An in-depth analysis of the GT data was conducted to guide feature selection. First, descriptive statistics and data distribution of the relevant metrics are presented, and next, group differences are analysed based on visualisations. The compared groups are SBB and Google Maps, along with the different interaction states and the participants.

4.2.1 Descriptive Statistics

The GT dataset consists of 20 participants, each with 76 tasks, resulting in a total of 1'520 app sessions. It was collected in April and May 2025. After filtering out sessions due to incorrect app usage or other issues, 1'473 app sessions remained. Of these sessions, 677 were SBB and 796 were Google Maps sessions. As explained earlier, the interaction states of the SBB app were only partially suitable for analysis, as most Check and Buy sessions included four or fewer taps. The statistics on the SBB interaction states will therefore not be presented here, but they are available on GitLab, and the SBB dataset was used for the group comparison.

The Google Maps dataset consists of 252 Search sessions, 120 Map sessions, 231 Place sessions, and 134 Direction sessions. This imbalance was partially caused by filtering out noisy sessions and by the experiment design. The average session duration and taps per session, by interaction state, are presented in Table 4. Map is clearly the longest interaction state, with the most taps per session.

Table 4: Average session length and taps per session by interaction state, Google Maps GT data

State	Average Session Duration (s)	Average Taps per Session
Search	12.763	12.806
Direction	18.411	11.861
Place	25.672	21.643
Map	61.622	53.683

After the feature reduction process, the final classifier was trained using five features: ITI_log_p25, tapsSession_log10, log_ITIs_std, log_ITIs_median, and ITI_log_diff_std. Therefore, only the statistics on the ITIs, the ITI differences and the taps per session are presented here. Further analysis on additional metrics is documented on GitLab. The distributions of the three log-transformed metrics were visualised in histograms. The distribution of the taps per Session is right-skewed with clear outliers of low tap counts (Figure 8).

Conversely, the distributions of ITIs and ITI differences are close to a normal distribution, indicating that most values fall within a similar range with only a few outliers (Figure 9 and Figure 7). However, when testing for normal distribution, none of the three metrics follows a normal distribution.

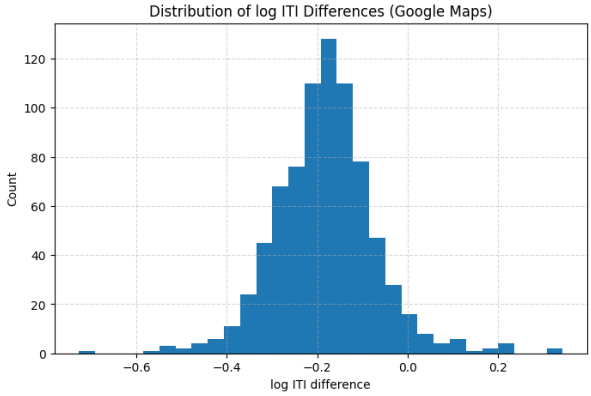

Figure 9: Distribution of log-transformed ITI differences of Google Maps ground truth data

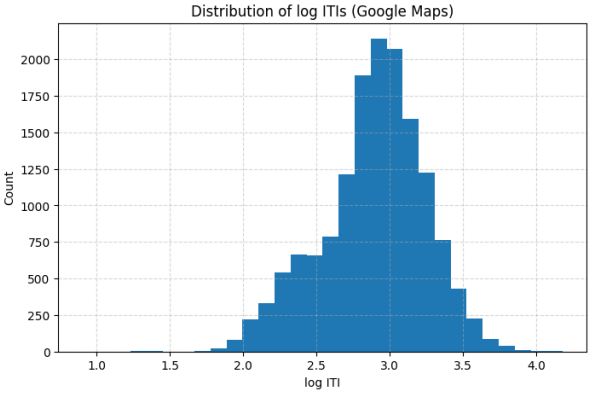

Figure 7: Distribution of log-transformed ITIs of Google Maps ground truth data

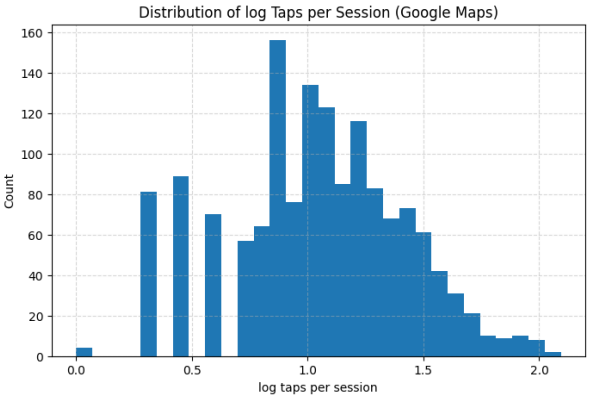

Figure 8: Distribution of log-transformed taps per session of Google Maps ground truth data

Another way to explore the temporal distribution of taps within app sessions is to visualise each tap from the session start, for sample sessions of a random participant (see Figure 10). This illustrates the type of data captured by tappigraphy. Overall, the taps of the interaction states, Search, and Direction seem to be more clustered. Furthermore, Figure 10 highlights that Map sessions are often longer than others.

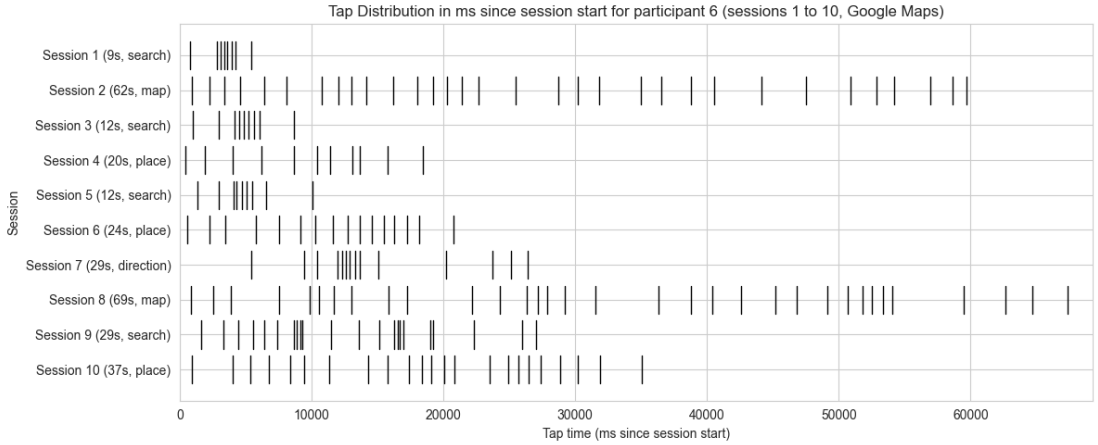

Figure 10: Tapping distribution in milliseconds (ms) since session start for different interaction states, sample data of participant 6 in Google Maps

Next, the correlation matrix overall and the individual correlation coefficients were inspected (see Figure 11). The correlation coefficient r represents the strength and direction of the correlation between two variables. Values around 0 indicate a weak correlation, while values close to 1 suggest a strong positive correlation, and values close to -1 indicate a strong negative correlation.

The correlation matrix displayed that there were strong positive correlations between ITI_log_p25 and log_ITIs_median (r = 0.84), as well as between ITI_log_diff_std and log_ITIs_std (r = 0.85). Conversely, ITI_log_p25 was negatively correlated with ITI_log_diff_std (r = -0.54) and log_ITIs_std (r = -0.69). Additionally, taps per session showed only weak correlations with all the other features, with a maximum correlation of r = -0.22.

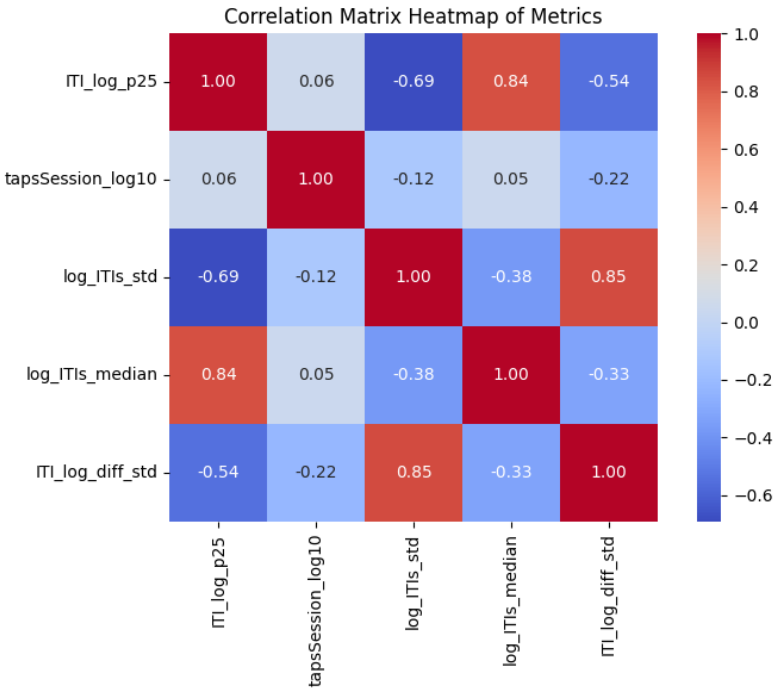

Figure 11: Correlation matrix heatmap of the five chosen metrics in the Google Maps GT data

4.2.2 Group Comparison

Differences between interaction states, apps, and participants, considering various metrics, were explored in depth. However, only a few examples will be presented here; the rest is documented on GitLab. The following group comparisons were based on ITI_log_p25, as prior research has identified it as an estimate of tapping speed (Ceolini et al., 2022).

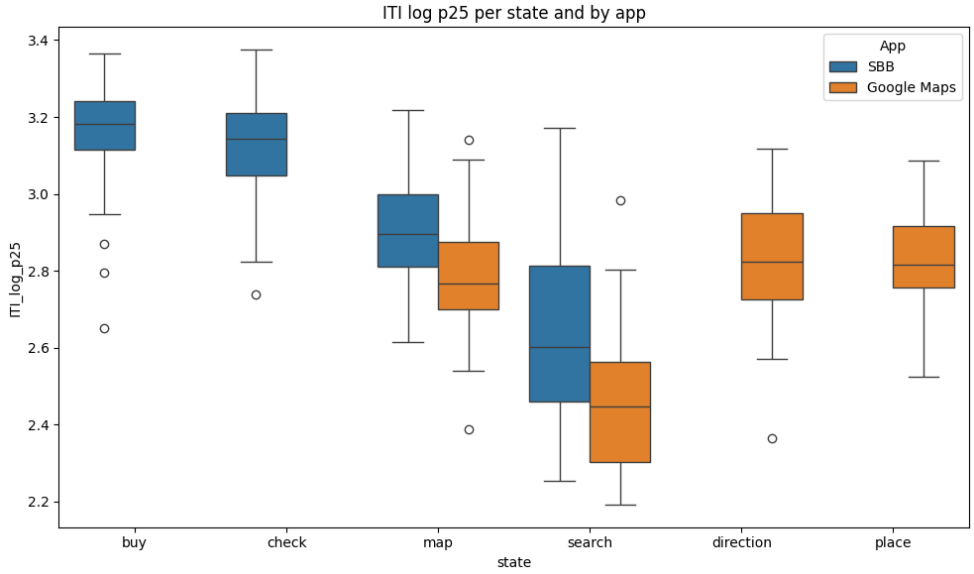

Figure 12: Distribution of ITI log p25 values depending on the interaction state and app, GT data

Figure 12 shows that Buy and Check are interaction states with relatively slow tapping, and therefore high ITI_log_p25 values (long intervals between taps). Furthermore, for the states Search and Map, the ITI_log_p25 values are, on average, shorter in Google Maps than in the SBB app.

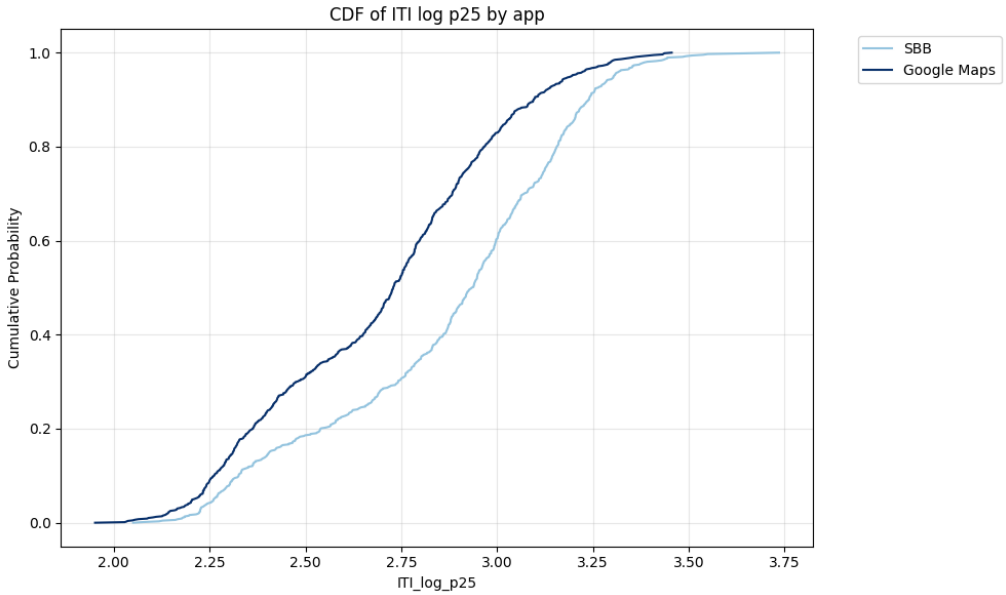

Figure 13: Cumulative distribution function (CDF) of ITI log p25 per app, GT data

Visualising the sum of all interactions in each app using a cumulative distribution function (CDF) reveals that the ITI_log_p25 in Google Maps is, on average, lower, indicating shorter breaks between taps (see Figure 13).

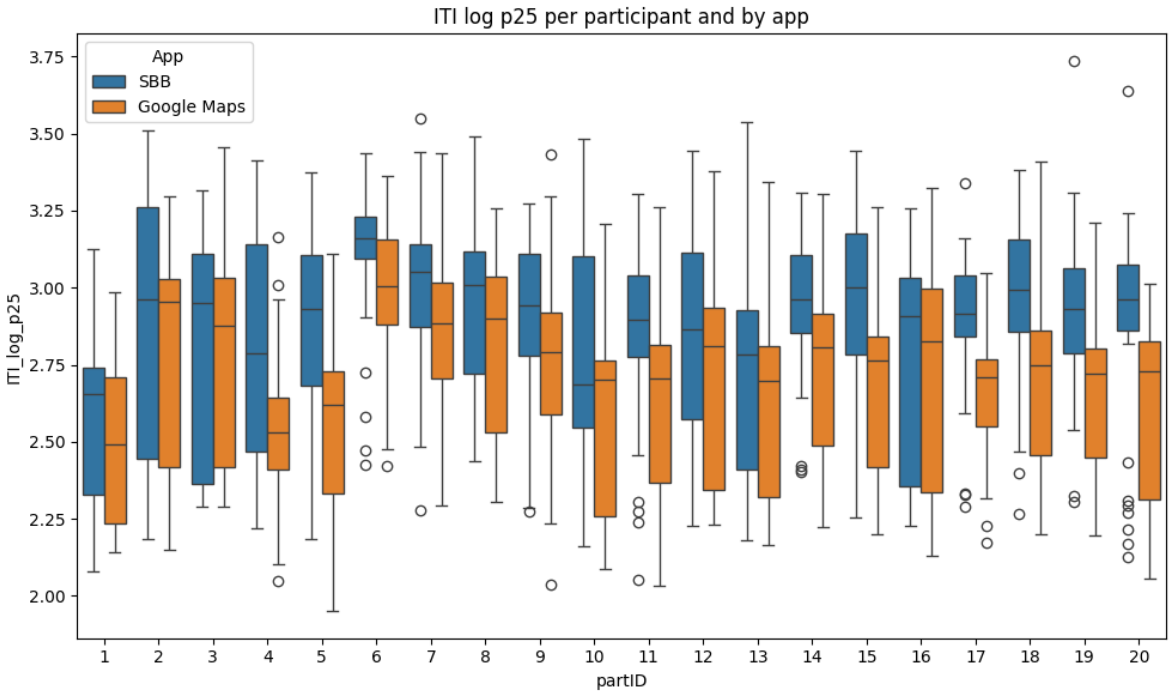

Figure 14: Distribution of ITI log p25 per participant for SBB and Google Maps, GT data

Figure 14 shows that the ITI_log_p25 values vary strongly between participants. The overall trend of shorter ITI_log_p25 values for Google Maps becomes apparent again. Additionally, the variability within participants and within apps is substantial.

4.3 Performance of Different Machine Learning Algorithms

In this section, the focus lays on evaluating the performance of different machine learning algorithms in classifying the four interaction states of the Google Maps GT data. Although performance was also assessed on the SBB dataset, no classifier specific to SBB or combining the SBB data with the Google Maps data was trained or run on real-world data. Thus, the performance of the different SMLCs on the SBB GT data will not be presented, but can be checked on GitLab.

The following results are therefore all based on training classifiers on the Google Maps GT dataset. The four classifiers compared in this study are Random Forest, SVC, Gradient Boosting Classification, and Logistic Regression.

Eight features were available and scaled when necessary (see Table 3). These features were then used in different combinations to achieve the highest CV and balanced accuracy score with the minimum number of features for each classifier. Furthermore, automated hyperparameter tuning with grid search CV was used to improve model fit.

In Table 5, the scores of each model using one feature (ITI_log_p25), all features, and the optimal feature combination with hyperparameter tuning are presented. For Logistic Regression and the SVC, the five optimal features were ITI_log_p25, tapsSession_log10, log_ITIs_std, log_ITIs_median, and ITI_log_diff_std. The optimal four features for the Random Forest classifier and the Gradient Boosting classifier were ITI_log_p25, SessionLength_sec_log10, log_ITIs_std, and log_ITIs_median. The only difference between these sets is the inclusion of tapsSession_log10 and ITI_log_diff_std among the five features, instead of SessionLength_sec_log10 among the four features. The feature names were explained in Table 3.

The performance of the models using only one feature is approximately 0.5. The differences in scores between using all features and the optimal features are minor.

Table 5: CV and balanced accuracy scores of the different SMLCs using different feature combinations

Model	Features	Hyperparameters	Average CV Score	Average Balanced Accuracy Score
Random Forest classifier	all	default	0.7806	0.7647
	one (ITI_log_p25)	default	0.4494	0.3965
	optimal (4)	tuned	**0.7586**	**0.7448**
Gradient Boosting classifier	all	default	0.7864	0.7676
	one (ITI_log_p25)	default	0.5161	0.4311
	optimal (4)	tuned	**0.759**	**0.7419**
Logistic Regression	all	default	0.7947	0.7793
	one (ITI_log_p25)	default	0.57969	0.4398
	optimal (5)	tuned	**0.7868**	**0.7673**
SVC (RBF)	all	default	0.7843	0.7675
	one (ITI_log_p25)	default	0.597	0.4706
	optimal (5)	default	0.7942	0.776
	optimal (5)	tuned	**0.7999**	**0.7855**

All classifiers achieved high accuracy scores, ranging from 0.74 to 0.8, with the differences between the scores being minor. For the SVC, the performance measures increased when using fewer features. For all the other classifiers, that was not the case.

The SVC classifier has the highest average CV and balanced accuracy score. Furthermore, it also has a high processing speed independent of the number of features (Singh et al., 2016). Therefore, it was chosen as the final classifier.

4.3.1 Performance of the Support Vector Classifier

For the SVC, the score was further enhanced by applying different weighting techniques and upsampling. The scores are presented in Table 6. The highest average CV score and balanced accuracy were achieved by using upsampling, resulting in 0.807 for both performance measures.

Table 6: Differences in CV and balanced accuracy scores depending on the weighting method (SVC RBF classifier on GT Google Maps data)

Method	Average CV Score	Average Balanced Accuracy Score
None (no weighting)	0.799	0.786
Standard Balanced Class Weighting	0.797	0.801
Inverse Number of Samples	0.760	0.771
Inverse Square Number of Samples	0.797	0.793
Upsampling	**0.807**	**0.807**

The permutation feature importance of all eight features for the SVC RBF is depicted in Figure 15. This was used as a starting point for the trial and error of different feature combinations. It led to the exclusion of three features. One of them, SessionLength_sec_log10, although it has relatively high permutation feature importance, turned out to be a less relevant feature.

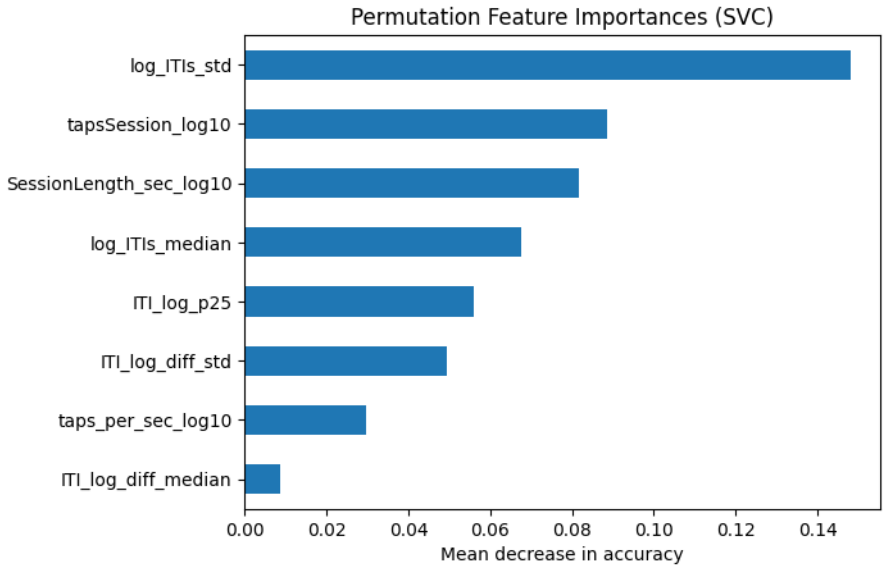

Figure 15: Permutation feature importances of all eight features without hyperparameter tuning (SVC RBF)

After the feature reduction and the hyperparameter tuning, the permutation feature importance changed (see Figure 16). TapsSession_log10 was now the most relevant feature, and ITI_log_p25 was the least relevant one.

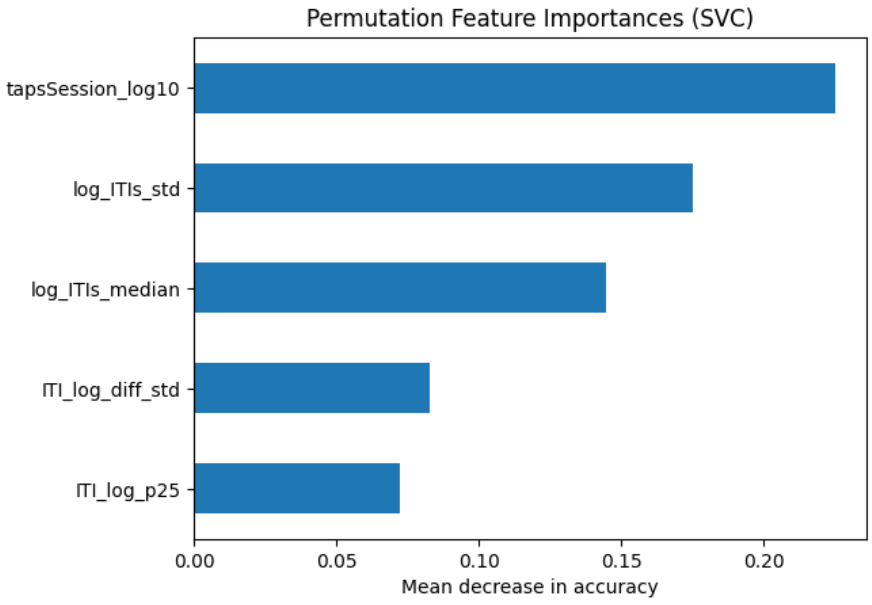

Figure 16: Permutation feature importance for the five main features, with hyperparameter tuning (SVC RBF)

The classified phone sessions were further analysed in a confusion matrix (see Figure 17). It depicts the share of (mis)classified sessions per interaction state. The dark blue diagonal highlights that most sessions are classified correctly. The medium-dark blue fields indicate that some interaction states are more likely to be misclassified as each other, specifically Search vs. Direction and Map vs. Place.

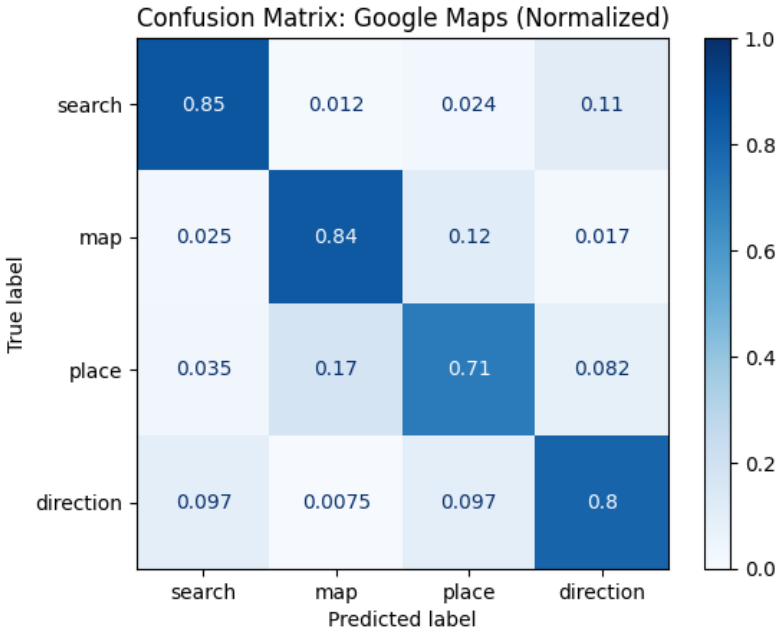

Figure 17: Normalised confusion matrix of the SVC RBF, depicting the share of (mis)classified sessions per interaction state for GT Google Maps data

4.4 Applying the Support Vector Classifier on Real-World Data

The final SVC with RBF kernel was trained on the complete upsampled Google Maps GT dataset. The upsampled version of this dataset includes a total of 1008 app sessions divided into four interaction states, with 252 app sessions per state, and the five input features without missing values.

To evaluate the classifier's assumed prediction accuracy on real-world data, it was applied to different groups of mobile applications. For each app or app combination, the overall share of the predicted interaction states and their associated prediction confidence values were visualised. Specifically, the classifier was tested on Google Maps itself, similar map apps, public transport apps, communication apps and all remaining apps (no map, public transport or communication apps). No app was included in more than one app combination group to prevent two combinations from looking similar due to overlap effects.

The focus lies on the analysis of the real-world Google Maps data. On the one hand, because the quality of this classification can be assessed by comparison to the in-depth study of Savino et al. (2021), which is not possible for the other apps. On the other hand, tappigraphy data includes the information on the app used. The classifiers' goal is not to detect which app was used but to classify interaction states for its specific use case, which is map apps.

However, it is also explored whether the classifier produces reasonable results when applied to similar map apps and public transport apps. The comparison to communication and the remaining apps aims to verify that the classifier does not produce uniform results, regardless of the dataset. The average prediction confidence for each application combination is presented in Table 7.

Table 7: Average prediction confidence of the SVC RBF for different app combinations

App combinations	Average prediction confidence
Google Maps	0.762
Similar map apps	0.7928
Public transport apps	0.7957
Remaining apps	0.812
Communication apps	0.7872

After applying the classifier to the real-world dataset and evaluating the average prediction confidence, the low-confidence sessions (below 0.55) were removed from the classified dataset for all the app combinations to analyse the share of each interaction state without distortion effects. For example, for Google Maps, 1'215 of 6'460 app sessions were excluded. The following subsections present the results for each of the app combinations listed in Table 7.

4.4.1 Google Maps

Table 8 shows the average session duration and taps per session for the classified real-world Google Maps data. Map sessions have the longest average length, followed by Direction, while Place and Search are relatively short. It further stands out that the taps per session for Direction are low compared to its high session length. These descriptives were analysed as a comparison to the corresponding GT data descriptives.

Table 8: Average session length and taps per session per interaction state, Google Maps real-world data

Predicted State	Average Session Duration (s)	Average Taps per Session
Search	13.100	12.333
Place	16.847	17.018
Direction	72.091	12.665
Map	129.072	76.134

The average prediction confidence for applying the classifier to Google Maps by itself is the lowest (0.7620). However, it is clearly higher than the chance level of 0.25 for four classes. Figure 18 shows the prediction confidence per interaction state. The highest average prediction confidence was achieved in the interaction state Map with 0.896, while the lowest was Place with 0.683. Overall, the prediction confidence per app session ranges from 0.286 to 1.

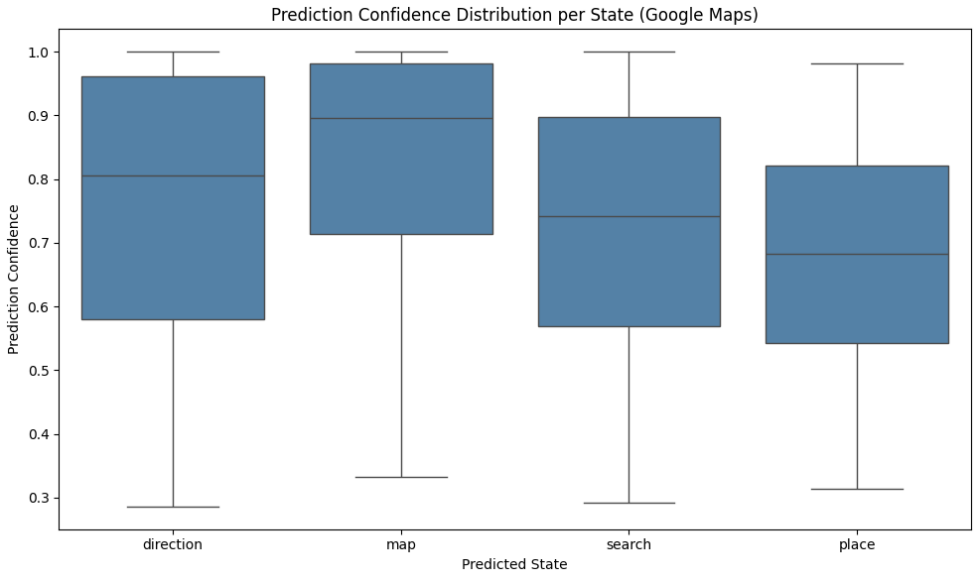

Figure 18: Prediction confidence of SVC RBF per interaction state for real-world Google Maps data

After the exclusion of the low confidence predictions, the number of app sessions classified per state is depicted in Figure 19. It shows that Direction and Map are equally common, while Search and Place are equally less common. However, this does not account for the time spent in each state.

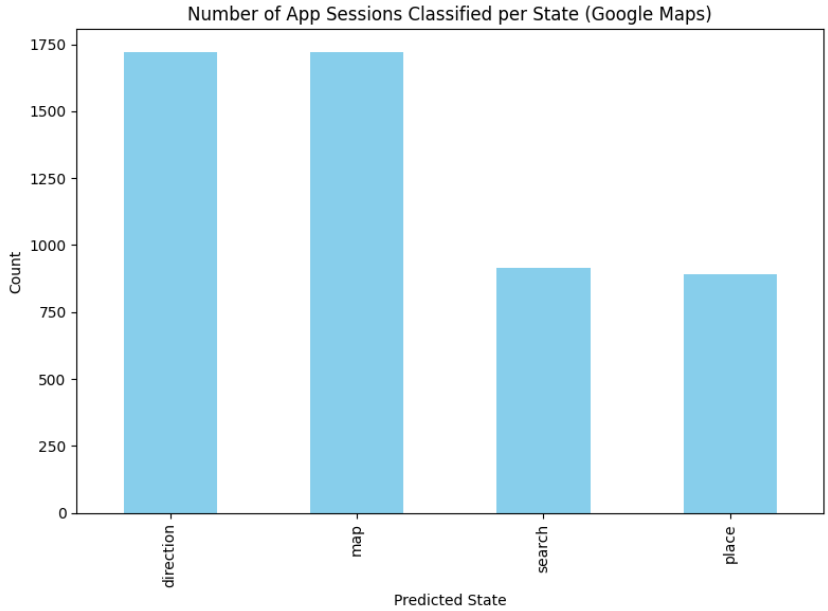

Figure 19: Number of app sessions classified per interaction state by SVC RBF for real-world Google Maps data

Conversely, Figure 20 visualises the share of the total app session duration per interaction state. Map is clearly the interaction state with the largest share of total app session duration, at 59.5%, while Direction is used about half as much, at 33.2%. Search and Place are still on a similar level, with 3.2% and 4% of the usage time, respectively.

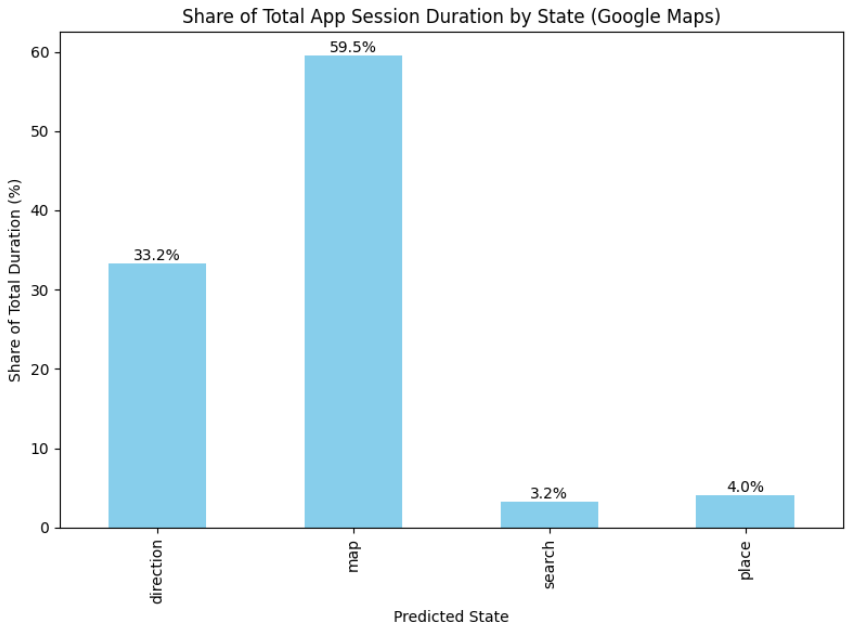

Figure 20: Share of total app sessions duration per interaction state for real-world Google Maps data (SVC RBF)

This pattern of the overall duration share distribution per interaction state, as seen in the real-world Google Maps data, is similar to the time share of interaction states within an app session found by Savino et al. (2021) (see Figure 1).

Instead of analysing the overall duration share distribution, it can also be analysed within phone sessions (see Figure 21). The resulting pattern is similar to that of Figure 19, describing the number of app sessions classified per interaction state. This was the case for all app combinations. Therefore, the duration share of interaction states within phone sessions is not displayed for the other app combinations.

The share of time spent in each interaction state differs when calculated over all app sessions (Figure 20) compared to within the respective phone sessions (Figure 21). The biggest difference is that the state map has a smaller share per phone session than overall.

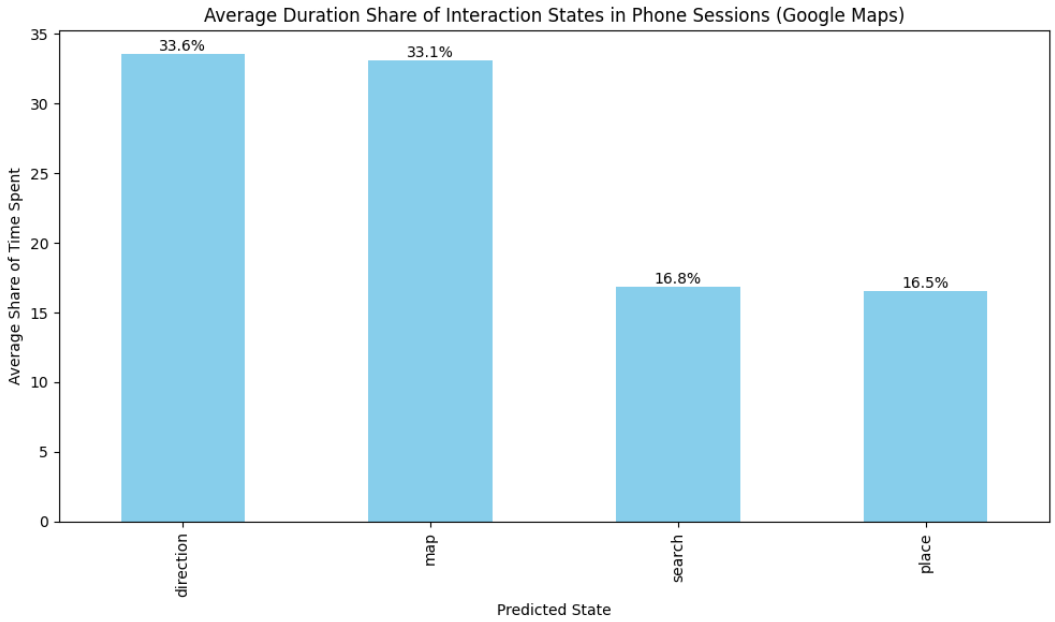

Figure 21: Average time share of phone session duration per predicted state for real-world Google Maps data (SVC RBF)

The sequence of interaction states within one phone session is presented in Table 9. In most phone sessions by far, only one interaction state was used. However, sometimes two app sessions with the same interaction state were used within a single phone session. Furthermore, the most common combinations are Map with Direction (0.76%), Direction with Map (0.57%), Map with Place (0.57%) and Search with Direction (0.48%).

Table 9: Ten most common interaction state sequences within one phone session, real-world Google Maps data (SVC RBV)

Predicted Interaction State Sequence in Phone Session	Number of Occurrences
Direction	1365
Map	1219
Search	710
Place	679
Map, Map	62
Direction, Direction	49
Map, Direction	40
Direction, Map	33
Map, Place	33
Search, Direction	25

4.4.2 Similar Map Apps

The similar map apps are a selection of the five map apps in the real-world data that are the most alike Google Maps: OsmAnd, Swisstopo-App, OsmAnd+ (pro version), HERE WeGo, and MAPS.ME. It was assumed that these map apps are similar enough to Google Maps, to allow the four interaction states to be applied to them.

Figure 22 shows the prediction confidence per interaction state for the similar map apps. The highest average prediction confidence was achieved in the interaction state Map with 0.908, the lowest being Place with 0.635. Overall, the prediction confidence ranges from 0.310 to 1. These results are almost the same as for the Google Maps data (see Figure 18).

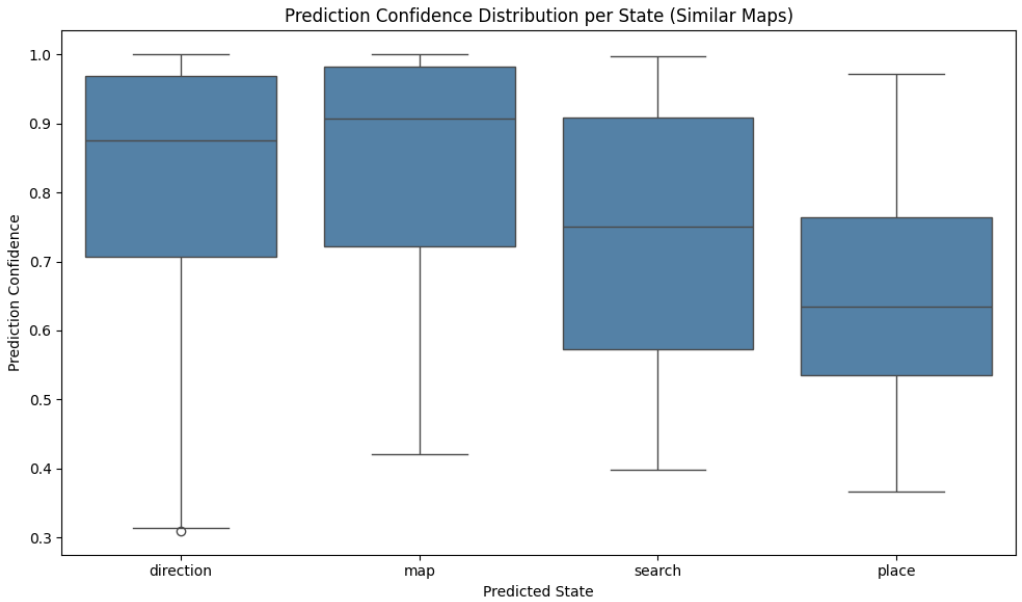

Figure 22: Prediction confidence of SVC RBF per interaction state for real-world data of similar map apps

After the exclusion of the low confidence predictions, the number of app sessions classified per state is depicted in Figure 23. The visualisation shows that the largest share of sessions is Direction, while Search has the smallest share. This is different from Google Maps, see Figure 19.

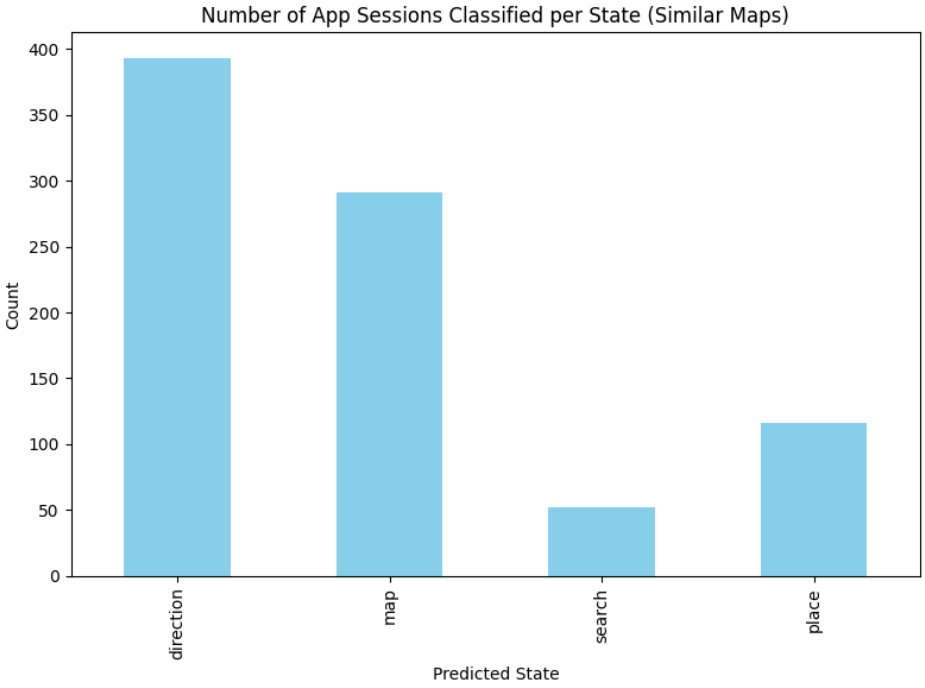

Figure 23: Number of app sessions classified per interaction state by SVC RBF for real-world data of similar map apps

The overall pattern of the share of total app session duration by state (Figure 24) has similarities to the number of app sessions classified per state (Figure 23). However, the share of Search and Place is even smaller than before, while the share of Map has increased. Compared to Google Maps, the shares of Search and Place are also small for similar maps, while Direction is clearly bigger and Map is smaller.

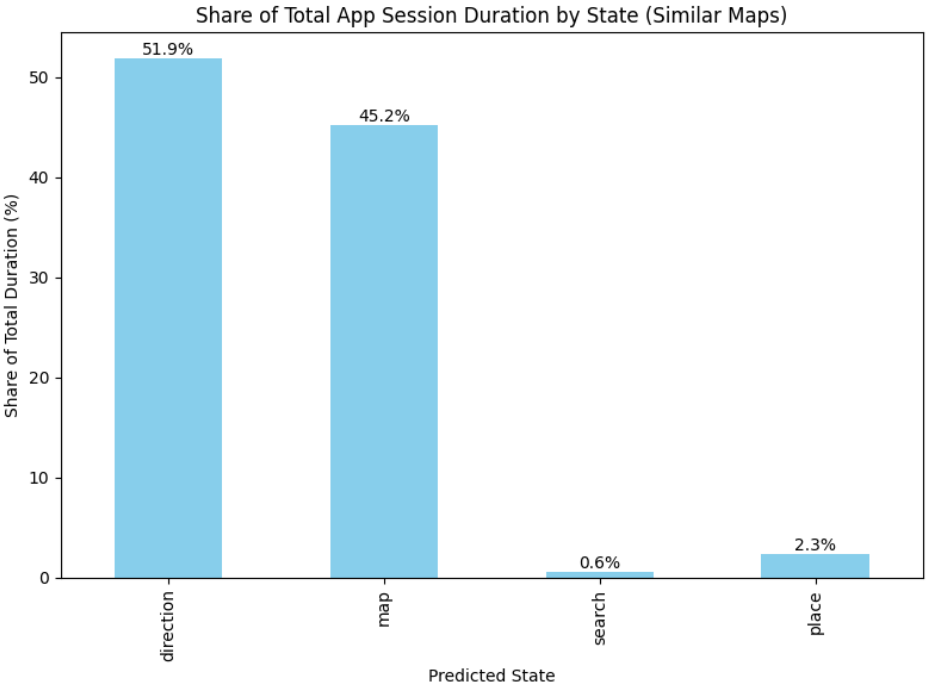

Figure 24: Share of total app sessions duration per interaction state for real-world data of similar map apps (SVC RBF)

4.4.3 Public Transport Apps

The 17 public transport apps are a sub-selection of the app store categories “Travel and Local” and “Maps and Navigation” from the real-world dataset. Only public transport apps were selected to avoid analysing a too broad category of travel apps. The specific app names can be found in the appendix (see Table 11). This data subset consists of 4’457 app sessions, of which 4’052 (91%) are SBB app sessions, which highlights that the SBB app is clearly the most used public transport app in this data collection.

The highest average prediction confidence was achieved in the interaction state Direction with 0.902, the lowest being Place with 0.573. Conversely to the distribution for the similar map apps and Google Maps, where Map had the highest prediction confidence. Overall, the prediction confidence spreads between 0.296 and 1, see Figure 25.

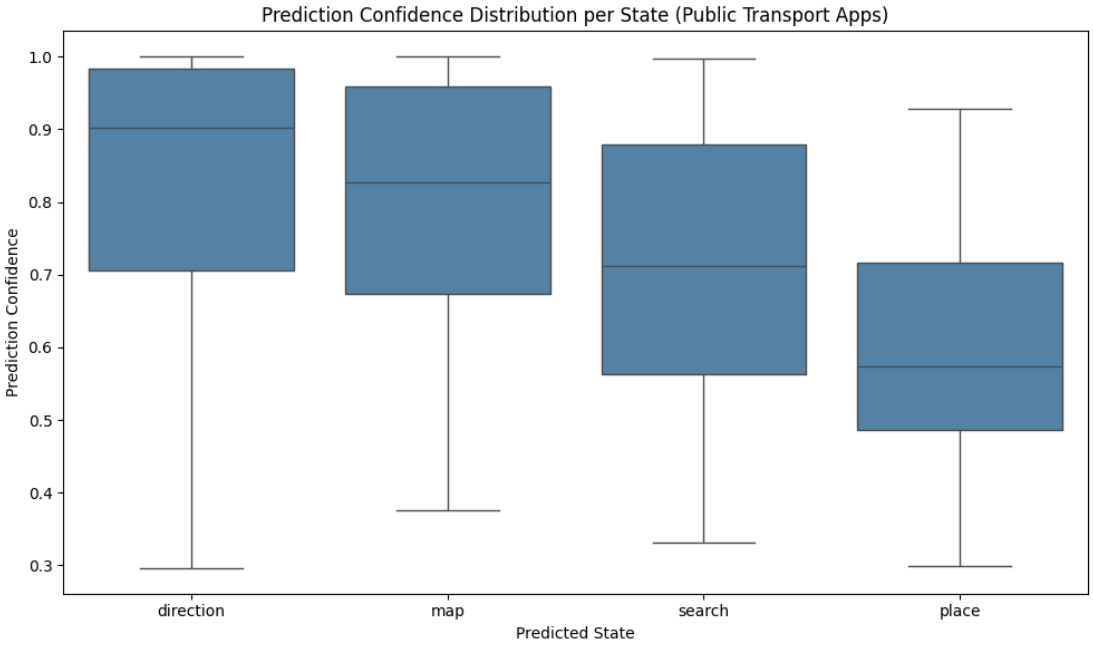

Figure 25: Prediction confidence of SVC RBF per interaction state for real-world data of Public Transport Apps

Direction is not only the interaction state with the highest average prediction confidence but also the state that occurs most frequently in terms of share of app sessions and share of app session duration (59.8%) (see Figure 26 and Figure 27). The same applies to Place, having the lowest average prediction confidence and occurring the least often (2.1% of session duration).

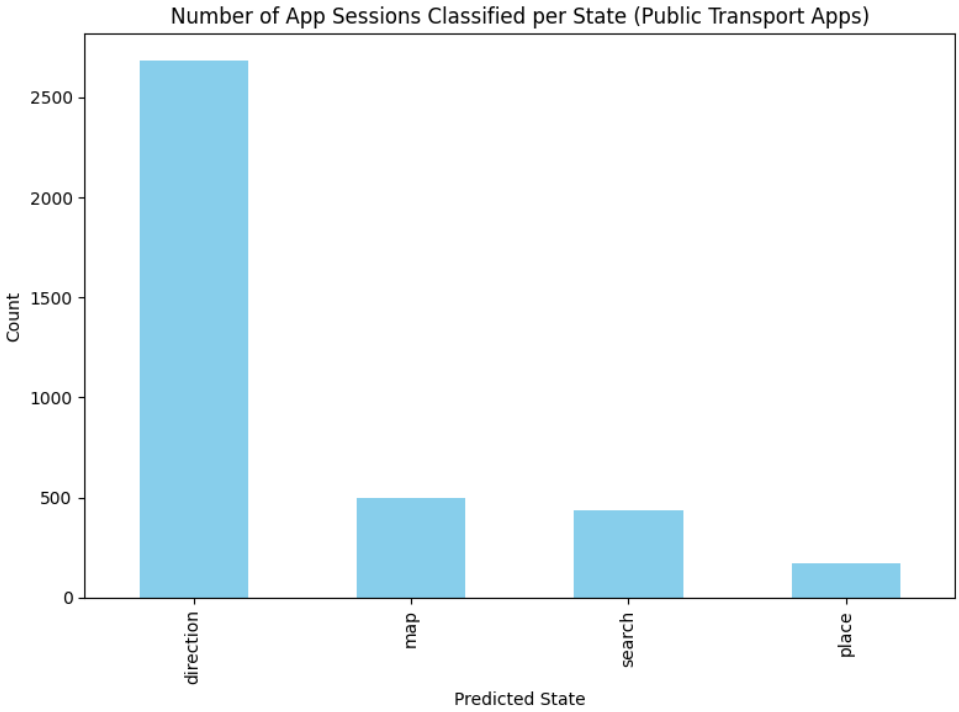

Figure 26: Number of app sessions classified per interaction state by SVC RBF for Public Transport Apps

In Figure 26, it becomes apparent that most sessions of public transport apps are Direction sessions, while for the map apps, it was Map and Direction. Looking at the share of time spent (Figure 27), Map has a bigger fraction again with 34.6%, while Search and Place have even smaller shares with 3.6% and 2.1% respectively.

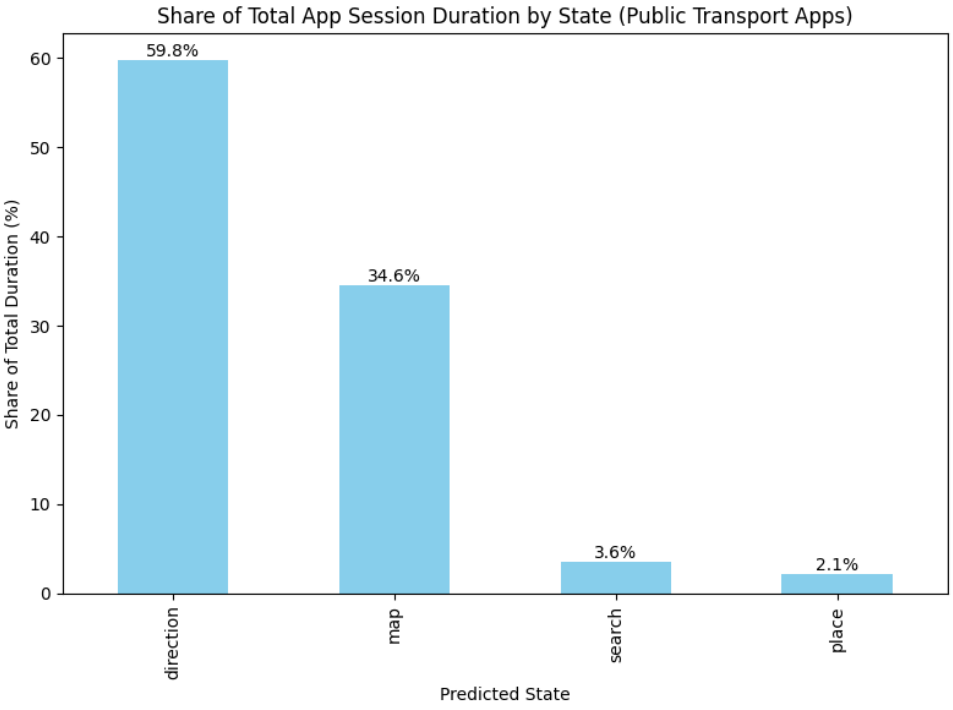

Figure 27: Share of total app sessions duration per interaction state for real-world data of Public Transport Apps (SVC RBF)

4.4.4 Communications Apps

The communication apps consist of 99’482 app sessions within 32 apps, which are simply all apps of the app store category ‘Communication’ that were included in the real-world dataset. The classifier was not trained for this type of data. Therefore, the following results are used as an exploratory comparison.

The average prediction confidence is high at 0.7872. The interaction state map has the highest average prediction confidence at 0.904, while the lowest is Place at 0.603. The overall prediction confidence spreads between 0.264 and 1, see Figure 28.

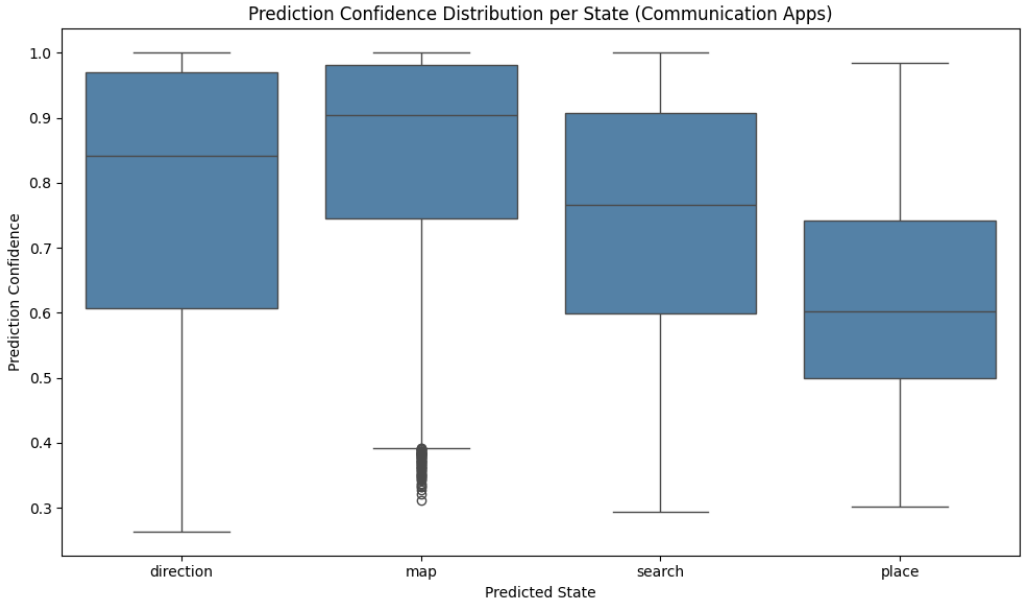

Figure 28: Prediction confidence of SVC RBF per interaction state for real-world data of Communication Apps

Figure 29 shows that most app sessions are classified as Map, followed by Search, then Direction, and finally Place. This pattern is different from that of the map apps and public transport apps. In both of these categories, Direction was always one of the two most common interaction states.

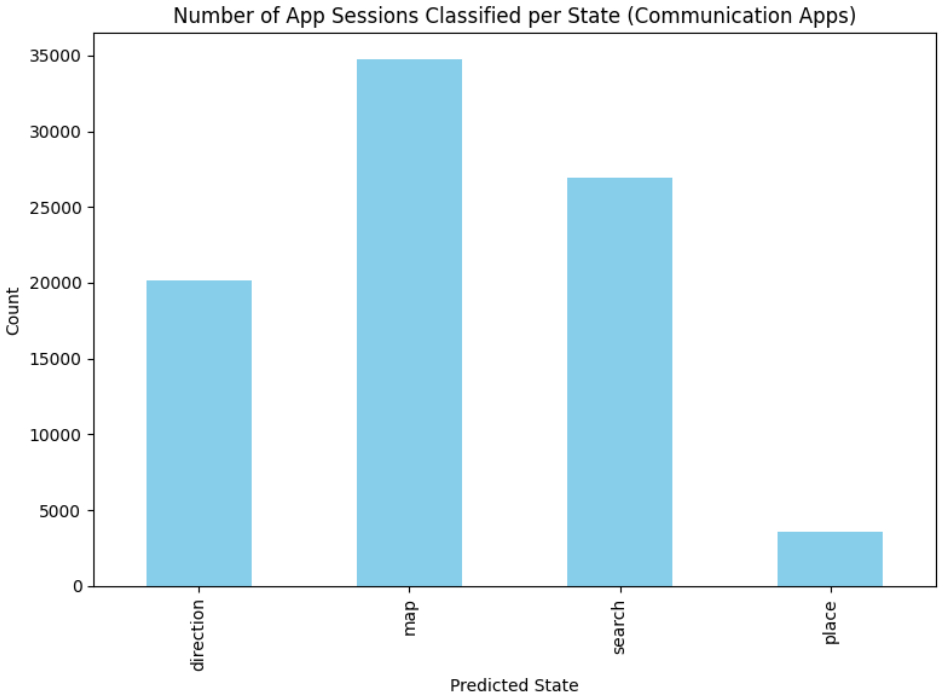

Figure 29: Number of app sessions classified per interaction state by SVC RBF for Communication Apps

When analysed based on the share of session duration, Map is by far the most used interaction state with 74.5%. The second biggest share is now Direction with 18.2%, followed by Search 6.3% and Place 0.9% (see Figure 30). This pattern is different from the one of public transport apps but has certain similarities with the one of Google Maps.

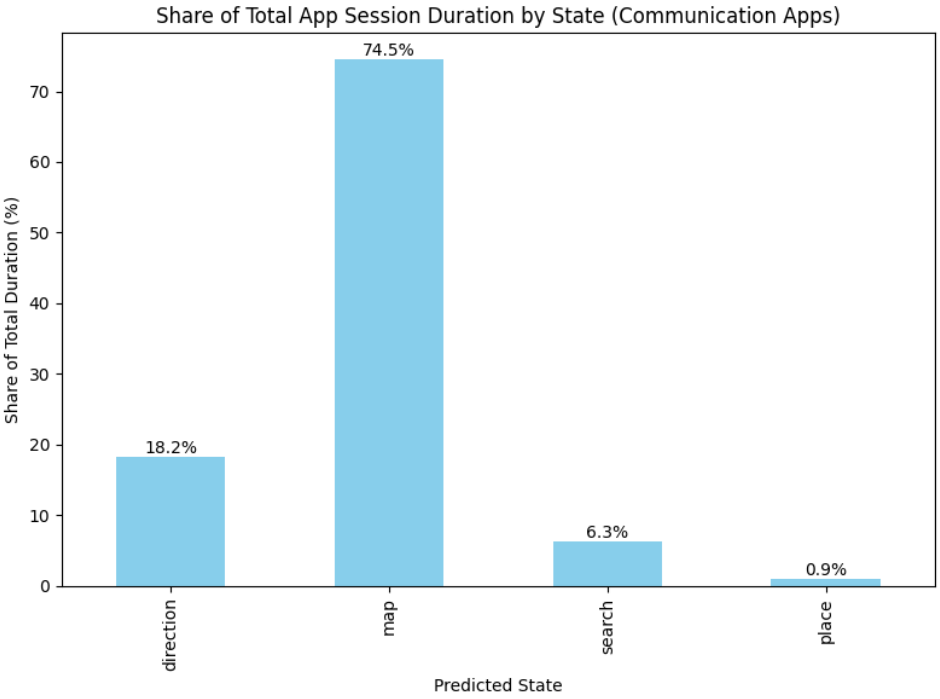

Figure 30: Share of total app sessions duration per interaction state for real-world data of Communication Apps (SVC RBF)

4.4.5 Remaining Apps

As the name implies, the remaining apps are all the apps in the real-world dataset which do not belong to the app store categories 'Maps and Navigation', 'Travel and Local' and 'Communication'. This app combination consists of 736 apps with a total of 137'346 app sessions and is therefore the largest data subsection. Again, the following classification is only an exploratory comparison, as the classifier was only trained for map apps.

The average prediction confidence lies at 0.812, which is the highest of all the app combinations. The interaction state with the highest average prediction confidence is Map, with 0.975, while the lowest is Place, with 0.627. Overall, the prediction confidence spreads between 0.258 and 1 (see Figure 31). This pattern has some similarities with those of the map apps and communications apps, but the confidence spread is larger for the remaining apps.

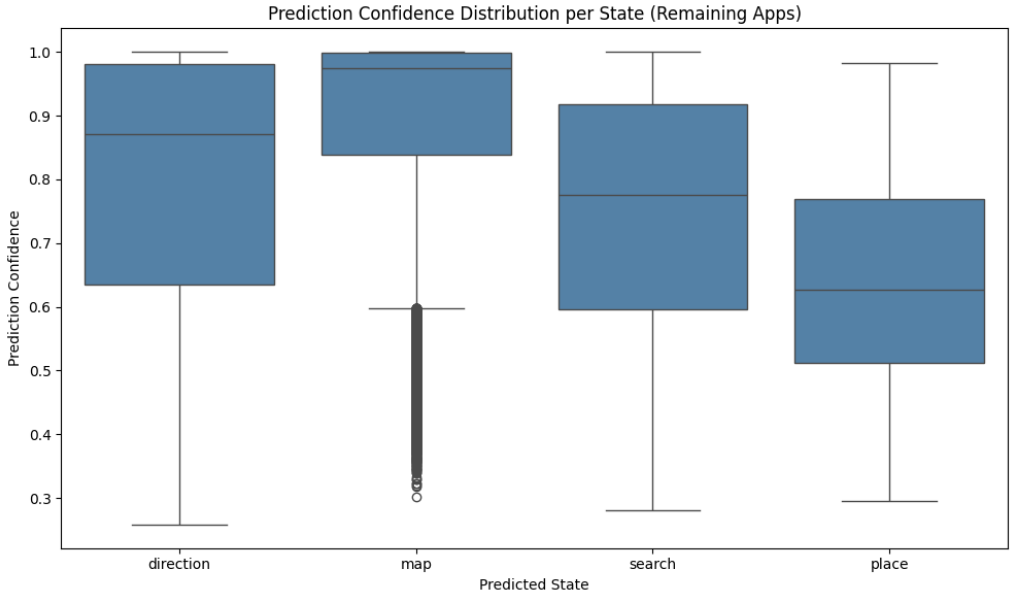

Figure 31: Prediction confidence of SVC RBF per interaction state for real-world data of the Remaining Apps

The pattern of the number of sessions per interaction state is different from all other app categories (see Figure 32). Most sessions were classified as Map (almost 50'000), followed by Direction, and then with notably fewer sessions Search and Place.

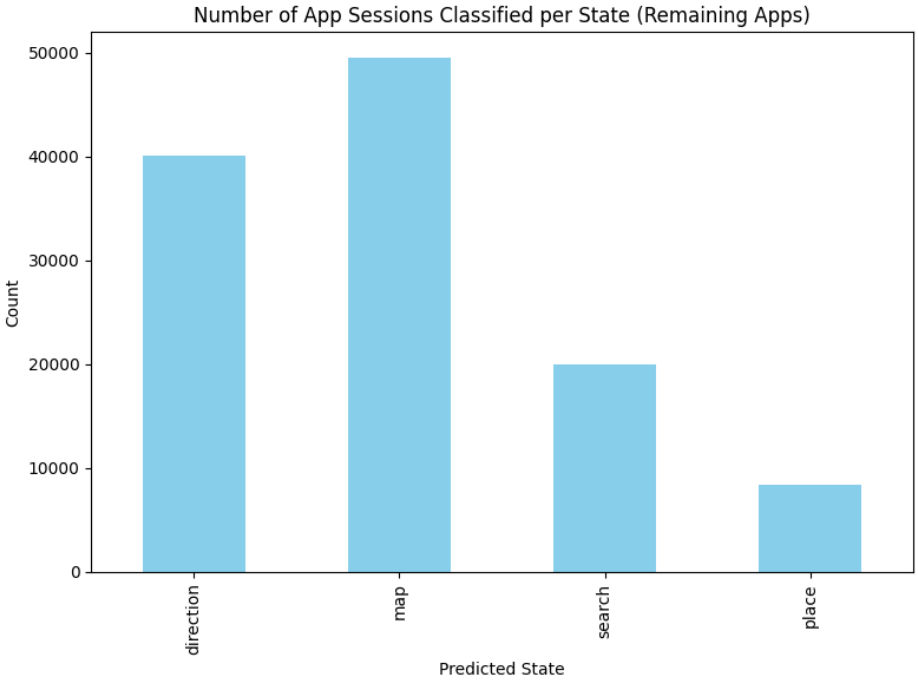

Figure 32: Number of app sessions classified per interaction state by SVC RBF for the Remaining Apps

The share of time spent in the different interaction states is similar to that of Google Maps and the communication apps. Map has the largest duration share with 71.2%. The share of the other states is smaller than when looking at the number of app sessions per state, specifically, 26.8% for Direction, 1.3% for Search, and 0.7% for Place (see Figure 33).

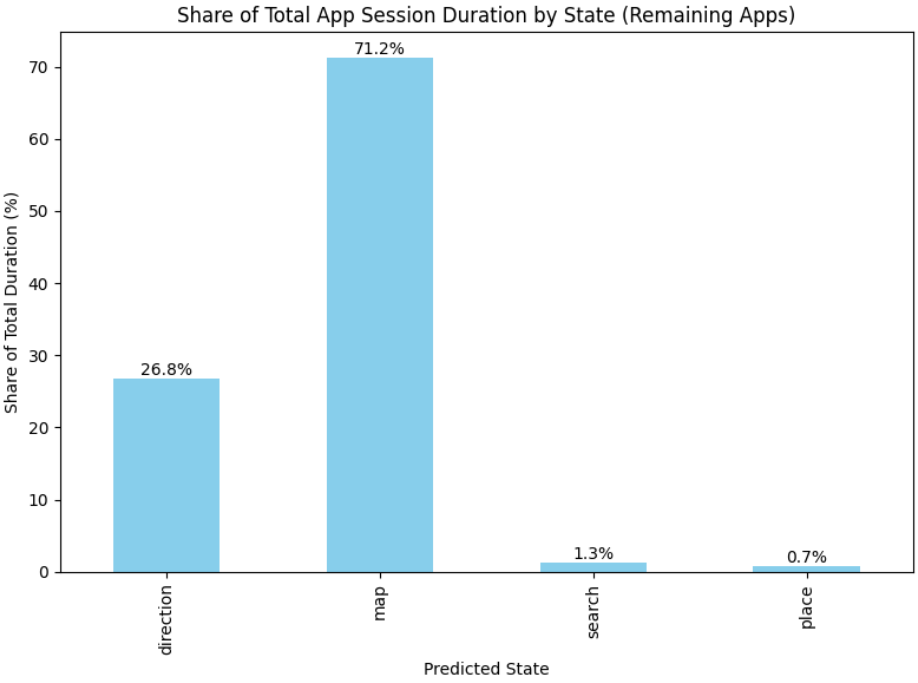

Figure 33: Share of total app sessions duration per interaction state for real-world data of the Remaining Apps (SVC RBF)

4.5 Applying the Support Vector Classifier on Ground Truth Data

Besides applying the SVC to real-world data, it was further explored how the classifier performs on synthetic combined interaction state sessions. Additionally, the classifier was run on the SBB App GT data, which was not part of the training set.

4.5.1 Combined Google Maps Sessions

Since Savino et al. (2021) found that app sessions often consist of more than one interaction state, the classifier's performance on combined app sessions was tested. This was done by producing synthetic app sessions by combining GT app sessions of different states within participants.

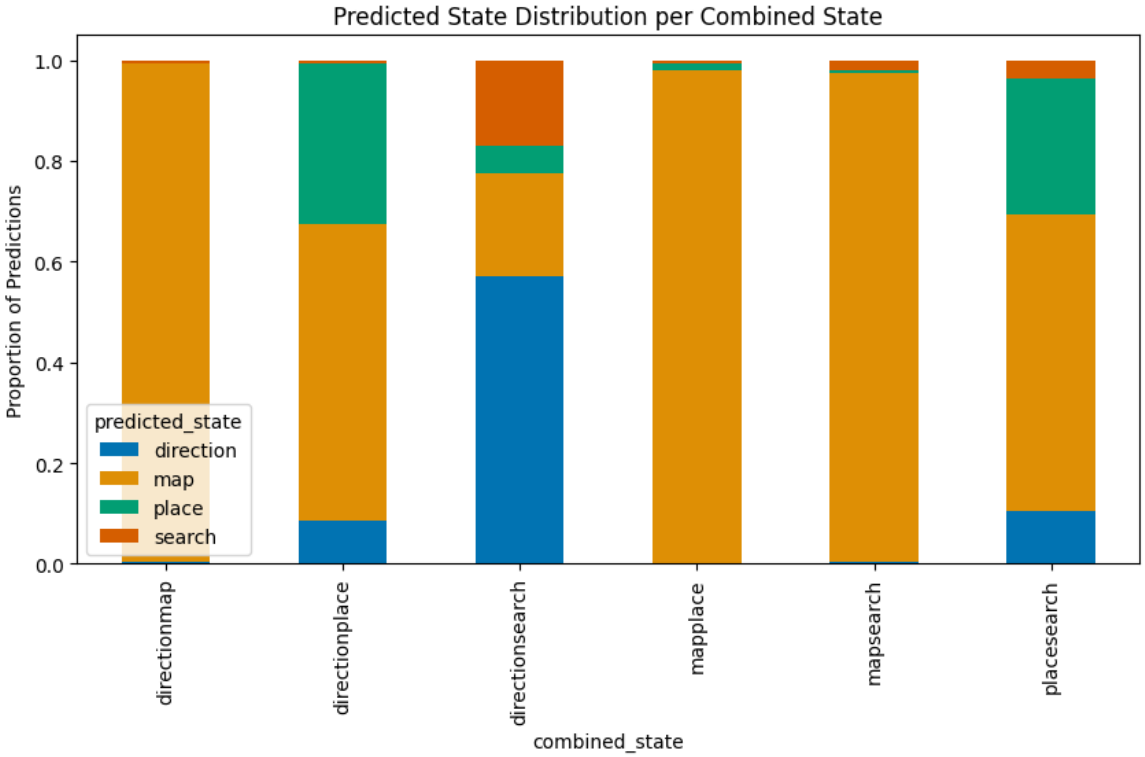

Figure 34: Share of prediction confidence per state depending on the combined state input, GT Google Maps data (SVC RBF)

Figure 34 shows that for all combined sessions, except for direction-search, the prediction confidence is highest for Map. This means all these sessions would be classified as Map, even those that do not contain Map at all. However, for the ones not containing Map, the prediction confidence was always below 0.55. This highlights the importance of setting a minimum prediction confidence threshold.

4.5.2 SBB App

The SBB GT data was not used to train a classifier specifically for public transport apps, nor was it used to train the Google Maps classifier. However, the classifier was applied to the SBB GT dataset to enable comparison with the classification results of real-world public transport apps. It should be noted that the SBB data is not a real-world test dataset since it is GT data.

The average prediction confidence was not very high, with 0.778. The overall pattern of the interaction state shares for the number of app sessions (see Figure 35) and the session duration (see Figure 36) is the same. Direction is the most common state with 52.8%, which aligns with the public transport apps, where it was 59.8%. However, Place is the second most common state, and Map is the least common state, which is different from all the app combinations of the real-world data.

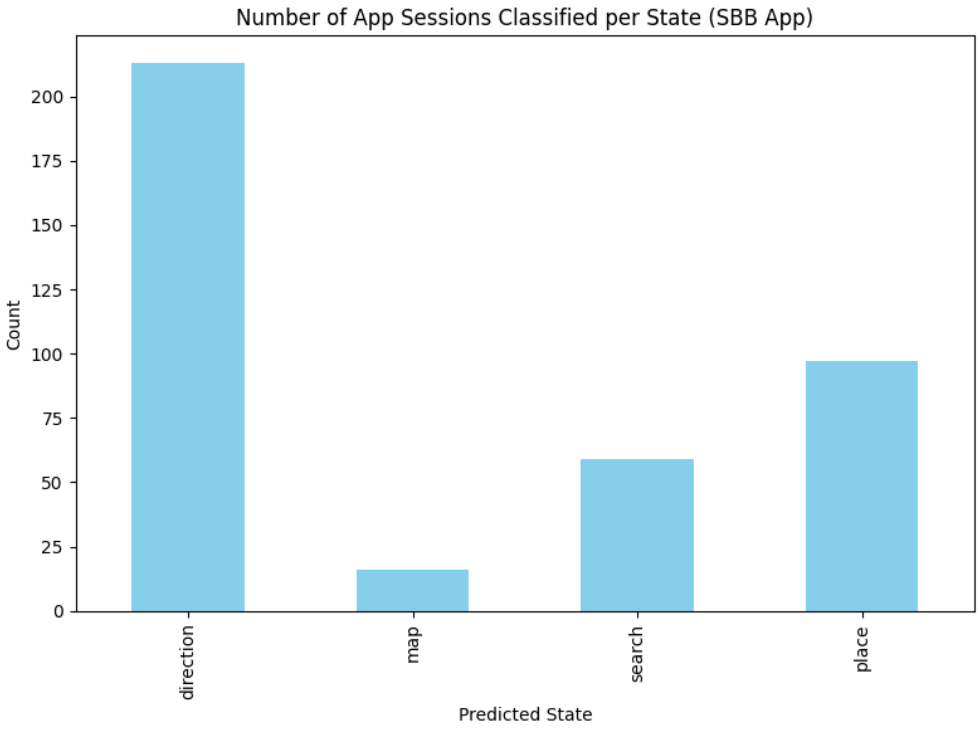

Figure 35: Number of app sessions classified per interaction state by SVC RBF for the SBB GT data

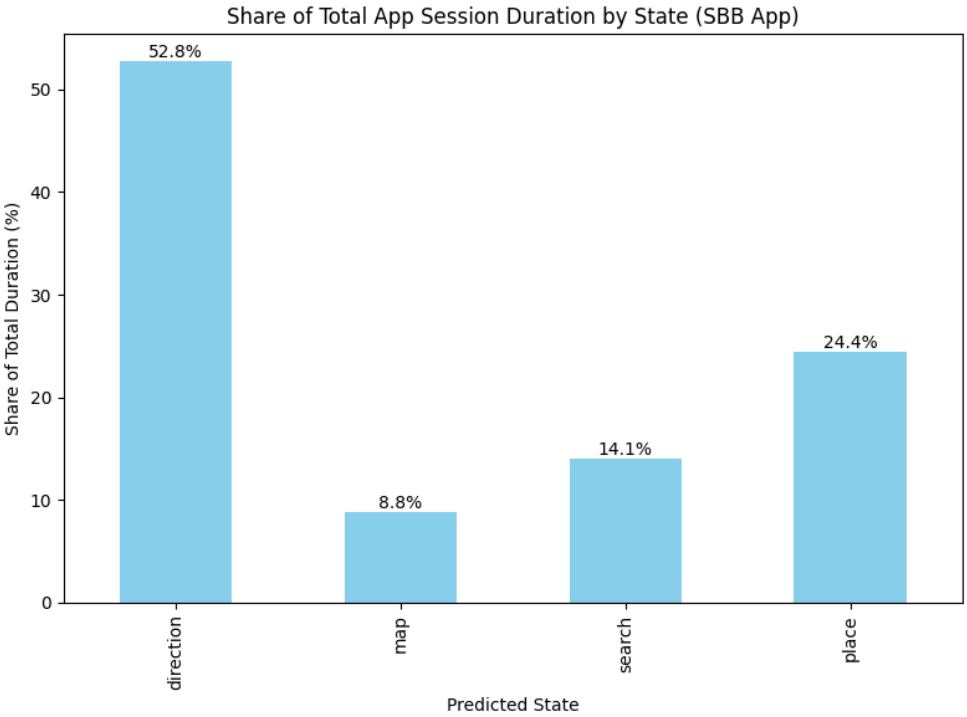

Figure 36: Share of total app sessions duration per interaction state for the SBB GT data (SVC RBF)

Since the SBB GT data is labelled, the confusion matrix can be assessed to evaluate which sessions were misclassified (see Figure 37). Map was often misclassified as Place, although the interaction state Place does not exist in the SBB app or other public transport apps. Search was further frequently misclassified as Direction, and Map was also often misclassified as Direction.

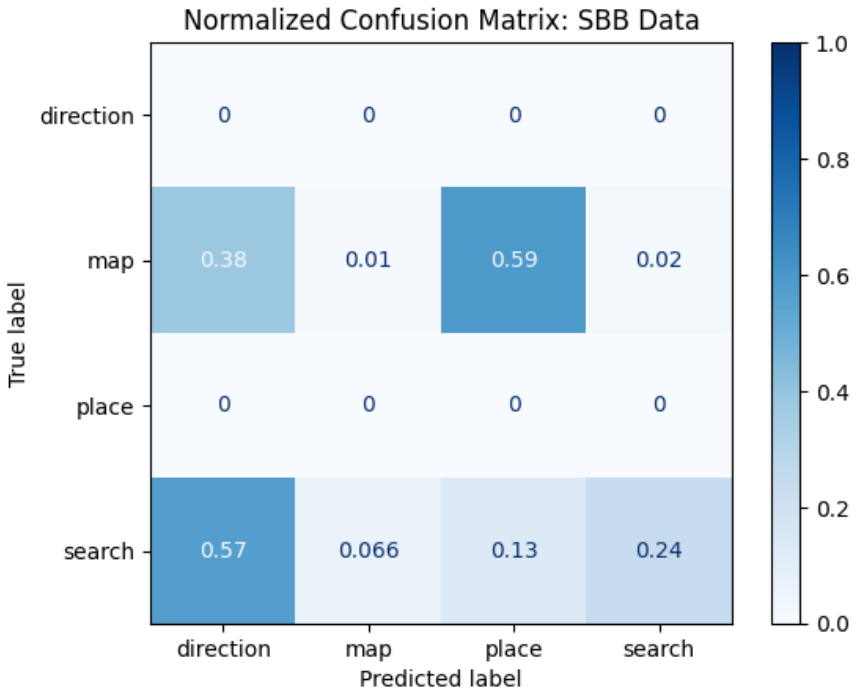

Figure 37: Normalised confusion matrix of the SVC RBF, depicting the share of (mis)classified SBB GT sessions per interaction state

5 Discussion

The goal of this discussion is to critically reflect on the data collection design, the data analysis, the choice of features and the SMLC, as well as the application of the chosen SVC on different app combinations. This includes examining the characteristics and comparability of the two datasets (real-world data and GT data), and analysing the GT data using descriptive statistics and group differences. Furthermore, research question 2 on suitable features for the SMLC will be answered, and the meaning of these features will be studied. Finally, research question 3, on the alignment of tapping patterns observed in the GT data with real-world tappigraphy data, is discussed.

The focus of this discussion lies on analysing the results of the SVC applied to the real-world Google Maps data. By comparing the classification of Google Maps data to the study of Savino et al. (2021), it can be assessed if the results align with prior research. Furthermore, comparing the classification of different app groups, such as map apps versus public transport apps, provides insight into category-specific usage patterns. Moreover, the classification of communication apps and the remaining apps shows how the classifier reacts to usage data beyond navigation. On the one hand, it can be assessed whether the classification is the same, regardless of the dataset. On the other hand, the classification of these non-map apps can be interpreted, for example, by examining what the interaction state Map could represent in an app without a map. Especially the comparison of these different app groups is exploratory, and the classifier's results cannot be verified.

To address research question 1, which concerns common map usage activities and environments reported in the literature, relevant studies were summarised in Table 1. Typical activities go beyond navigation and include searching for places, inspecting place information, exploring the map and checking route information. Environments are typically either commuting or leisure settings, reflecting the broad range of use cases for mobile map apps. This in-depth literature analysis guided the design of the GT data collection and informs the interpretation of the classification of the real-world tappigraphy dataset.

5.1 Real-World Dataset

The real-world dataset was collected between February 2021 and April 2022. Therefore, the usage data was affected by the second lockdown in Switzerland and subsequent regulations, which only allowed vaccinated people access to public spaces ("COVID-19 Pandemic in Switzerland," 2025). This can influence the usage of map apps and other apps and differs from the data collection context of Savino et al. (2021), who collected their data before the pandemic.

The participants are anonymous, so there is no information on age, gender or residency. However, it can be assumed that the majority are Swiss residents who are somehow connected to UZH since they were mainly recruited through the university network.

Furthermore, the amount of data per participant varies, since some participants only activated MaponTap for two weeks while others recorded their smartphone usage for over a year. This leads to the overall data pattern being influenced more by participants with more usage data.

The dataset analysis showed similarities to prior research; for example, Google Maps is clearly the most used map app (see Table 1). However, the data collection focus on Switzerland is depicted in the most used map and public transport apps, which include Swisstopo and the SBB app. This highlights regional differences in app preferences, which emphasises the importance of considering context when studying map app usage. The average Google Maps app session was 65 seconds, which aligns with the 65 seconds reported by Savino et al. (2021) and the 71.56 seconds reported by Böhmer et al. (2011). Finally, of the 254'379 app sessions, only around 4% were Google Maps and SBB sessions, resulting in a relatively small dataset for classification.

5.2 Ground Truth Dataset

The Google Maps GT dataset consists of 20 participants and 796 app sessions. It was not feasible within the scope of this thesis to collect more data. The dataset is imbalanced, partially due to filtering out noisy sessions and limitations in the data collection design. The data collection design is discussed in the next section (5.2.1). Although uneven class sizes are not ideal, the differences between classes were relatively small. Addressing the issue by trying different weighting techniques and upsampling resulted in only slight improvements in performance scores (see Table 6), indicating that the class imbalance did not substantially affect classification performance.

When looking at the average session length per state, the interaction state Map stands out with the longest session duration and average taps per session. Comparing the GT data to the later-calculated average session length per predicted interaction state in the real-world data, some similarities are observed. Search has a similar length in both datasets, with 12.762s versus 13s, while Direction is longer than Place for the real-world data. Furthermore, Direction and Map are clearly longer in the real-world data than in the GT data. This aligns with the fact that the GT Direction state only involves entering the start and destination and selecting an available route. In contrast, in the real world, it might also involve following directions while on the move. This theory is further reinforced by the low number of taps compared to the session length, which indicates that part of the interaction state Direction is visually inspecting the route without touching the screen. For the interaction state Map, one reason for the longer sessions could also be distractions when checking a map while on the move or in a busy environment. Another reason could be that people explore the map more extensively when, for example, really picking out a restaurant, than when told to do so in a lab setting. Additionally, the real-world dataset is from 2021, while the GT dataset was collected in 2025. Therefore, it is possible that small changes to the Google Maps interface influence differences in average session length across states.

All available metrics in the GT dataset were log-transformed to better capture variability in data characterised by outliers, which is typical for human behavioural data. Human behavioural data is often close to log-normally distributed (Gualandi & Toscani, 2019), which was also the case for the log ITIs and the log ITI differences, although the Shapiro-Wilk test indicated that they did not precisely follow a normal distribution. A distribution that is close to normal makes statistical analysis more robust (Gualandi & Toscani, 2019). Conversely, the log taps per session did not have a close-to-normal distribution; instead, there are many outliers, especially for low tap counts. Even though this data distribution was not ideal, it was handled by standardising the training data before it was used as input to the SMLC.

The GT data was also analysed visually by plotting the tap distribution from the session start for singular sessions (see Figure 10). The taps in Search and Direction sessions appear to be more clustered than in the other interaction states. This pattern might be caused by these two states, mainly including typing, while Map and Place include more scrolling.

The correlation matrix heatmap exposed that all metrics are correlated rather strongly, except for taps per session, which is only weakly correlated with the other metrics (see Figure 11). This correlation was expected, since three of the five metrics are summary statistics of the ITIs. The strong correlation is not ideal, but classifiers can generally handle correlated features. However, it can lead to overfitting. This issue can be addressed through hyperparameter tuning (Ashraf, 2023), which was later applied during SMLC training.

Finally, group differences within the Google Maps GT dataset and differences to the SBB GT dataset were studied. Inspecting differences across interaction states and apps, the Map and Search interaction states have shorter ITIs in Google Maps than in the SBB app (see Figure 12). This could be caused by user experience, e.g., better search suggestions in the SBB App might result in less typing there than in Google Maps. This can lead to shorter ITIs in Google Maps because typing generally produces shorter ITIs than other tapping. The CDF visualisation also shows that, overall, the ITIs in Google Maps are shorter than those in the SBB app (see Figure 13). Part of this effect is generated by the different interaction states used for the two apps, since Check and Buy are interaction states with relatively slow tapping. Moreover, most participants were familiar with all Google Maps functionalities, but not with all SBB app functionalities. This could also partly explain the slower tapping observed in the SBB app.

The distribution of the ITI_log_p25 per participant and app shows that the ITI log p25 varies strongly both within and between participants. This variability highlights the heterogeneity of app usage data, which depends on many factors. Capturing this variability in the GT data is beneficial since it replicates the expected variability in the real-world data. The overall trend of shorter ITIs for Google Maps than for the SBB App is observable across all participants.

5.2.1 Ground Truth Data Collection

The GT data collection in this thesis faced several challenges. Since no prior study had collected GT tappigraphy data, the implementation aligned with general data-collection rules but remained somewhat exploratory. For Google Maps, the definition of the interaction states was based on Savino et al. (2021), whose in-depth study was very valuable to ensure defining tasks which reflect common real-world usage. This was also reflected by most participants being familiar with all Google Maps functionalities. It further allowed training an SMLC whose predictions on real-world tappigraphy data are comparable to those of the reference study by Savino et al. (2021).

Conversely, for the SBB app, there was no prior study on common within-app usage. This made defining the tasks more challenging. Thus, tasks were based on the researcher's experience and peer consultation. This lack of prior insight led to difficulties, including half of the participants being unfamiliar with certain functionalities, such as saving a route. From the second participant onward, the unknown functionalities were explained before the experiment began. This approach is not ideal, as the first participant's experience differed, and the explanation could have influenced how participants used the functionalities. However, the tapping patterns of the first participant were similar to those of the other participants, suggesting comparability.

The definition of the SBB app states Buy and Check was not ideal, as they are both short and include few taps, which limits the possibility to analyse tappigraphy patterns within these states. This led to the exclusion of most sessions of these states from the classifier training, which then further made it unreasonable to train a separate classifier specifically for the SBB data, with mainly the remaining two states, Search and Map. It can be assumed that only two interaction states do not adequately capture the usage variability in the SBB app. This is unfortunate, since a classifier specific to public transport apps could have provided valuable insights.

Furthermore, the share of sessions per interaction state for Google Maps was uneven. Besides the cutting of noisy sessions, this was partly caused by planning issues and partly due to logistics. For example, participants first had to search for a city before they could explore the map there, resulting in more Search sessions. The data collection took place in a controlled lab setting to ensure the collection of enough data to train an SMLC; therefore, it does not allow insights into different usage environments. Further, including different environments would have required investing much more time per participant, which would also likely have made participant recruitment more difficult. This was not feasible within the scope of this thesis, especially since the focus was on building and running the SMLC.

5.3 Performance of Different Machine Learning Classifiers

There is no one-fits-all solution when it comes to choosing an SMLC. Hence, the performance of four SMLC, namely Random Forest, SVC, Gradient Boosting Classification, and Logistic Regression, using different feature combinations was assessed. All four classifiers reached high performance scores, with small differences between them. Interestingly, the performance of the models using only one feature was already around 0.5, well above the 0.25 chance prediction for four classes. The choice of input features is further discussed in Section 5.3.2.

The average balanced accuracy was consistently lower than the average CV score for all classifiers, which is caused by the uneven class distribution. This issue was addressed through hyperparameter tuning, which reduced the difference between the average CV score and the average balanced accuracy.

This indicates that this measure successfully mitigated class imbalance. Finally, the SVC classifier was chosen for its highest average CV and balanced accuracy scores, as well as its high processing speed.

5.3.1 Performance of the Support Vector Classifier

Due to the uneven class distribution, different weighting techniques and upsampling were applied to further improve the performance of the SVC. The highest average CV score and balanced accuracy were achieved by using upsampling, resulting in 0.807 for both performance measures.

The performance of this tuned SVC on upsampled GT data was further analysed using a confusion matrix (see Figure 17). It showed that some interaction states are more likely to be misclassified as each other than others. This is the case for Search and Direction, likely caused by both of these interaction states, mainly consisting of typing, whereas Map and Place do not involve typing. Map and Place were misclassified even more often. This is the case because when exploring the map (e.g. to find a restaurant), most participants also checked place details. Therefore, there is a certain overlap between these two categories.

5.3.2 Choice of Input Features

In the following paragraphs, research question 2 is answered concerning which features are suitable for labelling activity-specific tapping patterns. First, the specific features used for the SVC are listed, and then it is discussed what they describe and how they are correlated with each other.

The starting point for selecting the input features was the permutation feature importance score. However, this score is only an indication of the importance of a feature. Trial and error showed that simply selecting the features with the highest score does not necessarily lead to the best classification performance. Furthermore, feature importance changes depending on the hyperparameter tuning and the number of input features.

For the chosen SVC, the five optimal features were ITI_log_p25, log_ITIs_std, log_ITIs_median, tapsSession_log10, and ITI_log_diff_std. The first three metrics are summary statistics of the list of ITIs per app session, the correlation between these metrics is moderate to high (r -0.38 to 0.84). It can be assumed that they highlight different characteristics of the temporal dynamic within an app session.

The ITI_log_p25 was found to describe the tapping speed of phone usage since it covers the faster part of the tapping, like typing, compared to scrolling, etc. (Ceolini et al., 2022). This makes it a key indicator for tapping speed within an app session.

The log_ITIs_median indicates whether overall, there was more active interaction, like typing or slower passive interaction, like scrolling. The log_ITIs_std indicates whether there is a large average spread between slow and fast tapping or whether the app session was rather homogeneous. This can be an indication if an app session consisted of active and passive interaction or only one of them. The existence of these clusters of active and passive interactions was described by Zingaro et al. (2024b). It appears logical that taking these different types of interaction into account is important for the classification of interaction states.

The ITI differences show whether there was an acceleration or deceleration in tapping speed within an app session. For example, in the interaction state Direction, it can be expected that first the ITIs are small (fast tapping) while typing in the destination, and then the ITIs become larger when scrolling to choose a specific connection. This would be depicted by increasingly negative ITI difference values. The ITI_log_diff_std describes the average spread of these positive (acceleration) or negative (deceleration) ITI differences within an app session. This feature indicates changes in interaction type and the strength and abruptness with which they are performed. ITI_log_diff_std has the highest correlation with log_ITIs_std (r = 0.85). That makes sense, since both metrics describe how homogeneous tapping speed is within a session in their own way.

The feature tapsSession_log10 represents the number of taps per app session. It is influenced by both the session duration and the share of active interaction. The latter can be assumed to be higher for interaction states that involve more typing, like Search and Direction. Unlike the other features, tapsSession_log10 does not describe the speed or variability of tapping behaviour. This is also reflected by the low correlations with the other features (r = -0.22 to 0.06). Because it captures a different aspect of usage behaviour, it provides additional information and therefore increases the discriminative power of the feature set.

To answer research question 2, the five features ITI_log_p25, tapsSession_log10, log_ITIs_std, log_ITIs_median, and ITI_log_diff_std were found to be suitable to label activity-specific tapping patterns. These features capture different characteristics of the temporal dynamics within an app session, such as speed, variability, and overall activity. Together with information about the app used, they allow the analysis of interaction states solely using tappigraphy data, without requiring additional data on context, demographics, or app content.

5.4 Applying the Support Vector Classifier

The main objective of this thesis is to evaluate whether an SMLC trained on GT data with labelled interaction states collected in a controlled lab setting produces valid results when applied to real-world tappigraphy data. Specifically relating to research question 3, it is explored whether the tapping patterns observed in the GT data align with the ones in the real-world tappigraphy data.

To assess the performance of the chosen SVC, it was applied to different app combinations. The first one included only Google Maps data. This data subset corresponds exactly to the type of data the classifier was trained on, since the training data consisted only of Google Maps sessions. Furthermore, the classifier's performance on Google Maps data can be evaluated by comparing its results to the in-depth study of Savino et al. (2021). There are no such in-depth studies for other similar map apps or public transport apps.

Therefore, the classifier's performance on the other app combinations can only be used for comparison with the Google Maps classification. For example, for similar map apps, similar interaction state shares are expected when assuming that these apps are used equally. For public transport apps, a different pattern is expected, since certain interaction states, such as Place, are absent in these apps. These comparisons can be an indication of whether the classifier can generalise its learned patterns to different types of data while still producing realistic predictions.

Finally, the classifier was also applied to communication apps and to all remaining apps (those not categorised as map, travel, or communication). A classification with low prediction confidence and evenly distributed interaction state shares would have indicated that the classifier found patterns specific to map apps and was unable to find these patterns in unrelated apps. This was not the case. Since tappigraphy data includes information about the app used, this result does not imply that the exploratory approach failed. Instead, the classification of the other apps can be ground for interesting interpretation, such as what the interaction state Map could represent in a communication app.

Comparing the average prediction confidence of the SVC on the different app combinations, it is surprising that the lowest score was achieved for Google Maps (0.76) and the highest for the remaining apps (0.81). This highlights that the classifier finding matching patterns for the interaction states does not mean the resulting classification is correct, especially when applied to non-map apps.

Place has the lowest prediction confidence of all interaction states in every app combination. Map has the highest confidence across all app combinations, except for travel apps, where Direction is highest. The state with the highest confidence is also mostly the one in which the majority of app sessions were classified.

For all app combinations, a prediction confidence score threshold of 0.55 was applied; sessions with a classification below this confidence were excluded. The value aligns approximately with the 25th percentile of the prediction confidence distribution for the least confident interaction state. The goal was to cut as few predictions as possible while minimising the risk of including misclassifications. Including misclassified sessions bears the risk of distorting the interaction state share distribution. Excluding these sessions is therefore crucial to ensure a more meaningful analysis and comparison of the distributions.

The average time share of the interaction states per phone session depicted the same overall pattern across all app combinations as the number of app sessions per interaction state. This finding is notable because the share of the interaction state Map was larger when considering session duration overall than when considering the number of sessions or session duration per phone session. This pattern is likely caused by the longer average duration of Map in the real-world data (129.1s) than all other interaction states. Furthermore, many phone sessions consist of a single state, especially Direction and Map, making the time spent on Map seem less prominent when looking at it per phone session. The time share of interaction states per phone session is also more closely linked to participants, because if one participant has, e.g., many short phone sessions, these influence the overall time share per phone session more than a participant with fewer sessions but longer ones.

In addition to these overall patterns, app-combination-specific patterns were observed and are discussed in the next sections.

5.4.1 Google Maps

Looking at basic descriptive statistics, the duration of the interaction states in the GT data is comparable to that of the real-world Google Maps data. Search is clearly the shortest state (GT: 12.8s, real-world: 13.1s), and Map is the longest for both datasets (GT: 61.6s, real-world: 129.1s). Direction is clearly shorter in the GT with 25.7s compared to 72.1s. This difference is reasonable, as the Direction in the GT dataset only consisted of searching for and starting a route. In contrast, in the real-world dataset, it could involve following directions while moving. This is a shortcoming of using GT data from a lab setting rather than from common map-use environments.

Analysing the number of app sessions classified per state, Map and Direction are evenly common, and there are almost twice as many Map and Direction sessions as Place and Search sessions. This pattern doesn't fully align with the one found by Savino et al. (2021). However, they analysed the share of time spent in each interaction state within one app session (see Figure 1), whereas the SVC classifier in this thesis was trained by GT data with only one interaction state per app session. This is not ideal for comparison with Savino et al. (2021), but this decision was made because the real-world data's finest granularity is app sessions; therefore, labelling within app sessions would have been even more complex. Additionally, labelling GT data collected at the granularity of app sessions, with a share for each interaction state, would be challenging.

A first step to make the classifiers' results more comparable to the research of Savino et al. (2021) was to analyse the time share per interaction state (instead of app session count) overall and per phone session. However, the time share per phone session had a similar pattern to the number of app sessions. Studying the time share per interaction state across all app sessions revealed a pattern very similar to that found by Savino et al. (2021). The interaction state Map has the highest share with 59.5%, followed by Direction (33.2%), Place (4%) and Search (3.2%). Savino et al. (2021) found the Map state to be even more prominent with 67.5%, Direction was used less with 21.1%, Place was used a little more with 8.2% while Search was the same with 3.2%. This is a promising result, suggesting that the classifier identifies realistic patterns.

The slight deviations from the patterns found by Savino et al. (2021) could be caused by the relatively small sample size of both their study, with 28 participants, and the real-world data used here, including 51 participants. Furthermore, there are differences in usage context: their data collection was conducted in Bremen, Germany, included only locals, and lasted 4 weeks, while the MapOnTap (2021) data was mainly collected in Switzerland, lasted 2 weeks to over a year, and was affected by Covid-19.

However, it is surprising that there were more Direction sessions in the real-world data, which was affected by Covid-19, than in the unaffected dataset of Savino et al. (2021). It could be assumed that the restriction on moving within public spaces would have the opposite effect. Participants may have explored new outdoor locations due to restrictions on other activities. This would align with the findings of Do et al. (2011), who suggest that Map usage often occurs in first-time visit places. Additionally, Savino et al. (2021) used a wrapped version of the Google Maps website for their data collection, while in the real-world tappigraphy data, the Google Maps app was used. Differences in interface design and usability, especially of the Direction functionality, could therefore also explain the increased usage of Direction.

Furthermore, Savino et al. (2021) found that the most common sequences of interaction states within an app session were MSPD (12.1%) and MD (13%). Again, this is not directly comparable to the results of this thesis's classifier, since it classifies only one interaction state per app session. However, the closest possible comparison is inspecting sequences of interaction states within a phone session, possibly including more than one (Map) app session. Most phone sessions of the real-world Google Maps data consist of only one interaction state. Very rarely, more than two interaction states were combined in a single phone session. The most common combinations are Map-Direction (0.76%), Direction-Map (0.57%), Map-Place (0.57%) and Search-Direction (0.48%). Therefore, there are some similarities with MD being the most common combination. This more fragmented use could be connected to the real-world data having been collected during COVID-19, when the restrictions of the Swiss government led to people staying at home more. Therefore, maybe fewer interaction sequences from Map to Search to Place and Direction, or from Map to Direction, were completed. Again, the classifier assigns only one interaction state per app session, meaning that if more than one interaction state is used within an app session, it is not depicted, which makes the classifier's results less comparable to the interaction state sequences found by Savino et al. (2021).

To test how the classifier performs on app sessions, including two interaction states, it was run on synthetic data generated by combining GT app sessions. Most combined states were classified as Map. This could be connected to the average session length per state, which is more than double in Map in the GT data (61.662 s) compared to all the other states (12–26 s). This leads to Map being the dominating interaction state of these sessions; therefore, the classification is reasonable.

However, combined states that do not include Map but do include Place are still often classified as Map. This is likely caused by the partial overlap between the interaction states Place and Map, because exploring a map (as defined in the GT data) often involves checking place details. Nonetheless, this misclassification should be avoided. The misclassified sessions in this combined app sessions dataset never reached a prediction confidence above 0.55. Therefore, the previously defined threshold of 0.55 seems valid and avoids including this type of misclassification in the interaction state share analysis.

In conclusion, some sessions classified as Map likely involved a share of other interaction states. However, mixed app sessions should not be entirely misclassified. Furthermore, if this were the case for many Map sessions, the overall time share of Map sessions compared to the time share per app session reported by Savino et al. (2021) would be expected to be larger, which is not the case. Instead, it is smaller in the real-world dataset.

Overall, the answer to research question 3 is yes, the tapping patterns observed in the GT Google Maps data seem to align with the real-world tappigraphy data, especially for the Google Maps data. The classifier's performance confidence lies at 76%, which is high, and the classified interaction state patterns align with prior research.

5.4.2 Similar Map Apps

The prediction confidence per interaction state and the average prediction confidence (0.79) for the similar map apps are similar to those observed for Google Maps. This is an indication that the datasets are somewhat comparable. However, both the number of app sessions and the share of total app session duration display that the interaction state Direction was used more than Map, which is different from the pattern for Google Maps. This difference may be explained by the diverse functionality Google Maps offers for map exploration. For example, Google Maps allows searching for categories such as restaurants or sightseeing spots and provides easy access to extensive information on places, including reviews and photos contributed by its large user community.

This may explain why Map (45.2% versus 59.5%) and Place (2.3% versus 4%) accounted for a smaller share of total duration in similar map apps than in Google Maps, while Direction was used substantially more (51.9% versus 33.2%). This suggests that similar map apps are mostly used for getting directions, followed by exploring the map. This differs from the results of Savino et al. (2021), but their research only included Google Maps.

One of the similar apps was the Swisstopo-App, which allows users to explore different types of maps (satellite, historic, etc.) and to plan routes for outdoor activities like hiking (*Swisstopo - Apps on Google Play*, n.d.). In this context, the predominance of Direction and Map seems plausible.

Overall, these patterns seem realistic, indicating that the classifier, which was only trained on GT Google Maps data, can be applied to similar map apps. However, it is important to emphasise that, given the available datasets, the classifier's results cannot be validated.

5.4.3 Public Transport Apps

Contrary to Google Maps and similar map apps, the highest average prediction confidence for public transport apps was achieved for the interaction state Direction (0.902). Both the number of app sessions and the share of total app session duration indicate that Direction is also the most used state, accounting for 59.8% of total app session duration, followed by Map (34.6%), Search (3.6%), and Place (2.1%). This represents the highest share of Direction and the lowest share of Map compared to all other app combinations.

This aligns well with what most people presumably use public transport apps for, namely getting directions. Although the share of Map usage (34.6%) is relatively high, since exploring a map is arguably not a main task when using public transport apps. However, the SBB app, which is the main part (91%) of this data subset, includes a map. This map feature allows users to inspect entire routes, check surrounding areas during transfer and even view different floors of train stations. These various functionalities potentially explain the elevated time share of the interaction state Map. Furthermore, this complex and potentially confusing map interface might also lead to a decrease in usability, which further increases usage time. Consequently, Map sessions are on average longer, which is why the share of Map is more prominent in the temporal share than in the number of app sessions per interaction state.

The classifier, which was trained only on Google Maps data, cannot capture the full range of interaction states used in public transport apps. However, it does seem capable of depicting that the main interaction state in these apps is, as expected, Direction. These insights are exploratory and cannot be validated.

However, for comparison, the GT SBB data was also classified. The results differ somewhat from those of real-world public transport apps. Direction is still the most used interaction state in terms of both the number of sessions and time spent, followed by Place and Search, and lastly by Map.

The normalised confusion matrix (Figure 37) shows that Map was often misclassified as Place (0.59). Since the interaction state Place does not exist in the SBB app, this could indicate that sessions classified as Place in the public transport apps should also have been classified as Map. Additionally, Map was sometimes misclassified as Search (0.38). That is unexpected, as Map usually does not involve typing, while Search does. As expected, Search is often classified as Direction (0.57). This is likely due to different state definitions: in the SBB app, Search was defined as entering a start and destination, then choosing a route. This definition is closer to the definition of Direction in Google Maps than to its Search interaction state. This misclassification is therefore caused by different definitions of interaction states, rather than classification errors.

5.4.4 Communication and Remaining Apps

All interaction states, except Search, do not apply to communication or remaining apps, since these typically lack Map, Direction or Place functionalities. Nevertheless, it is interesting to compare the classifiers' results on these apps with those on map and public transport apps.

Both app combinations have a high average prediction confidence (communication apps: 0.787, remaining apps: 0.812). This highlights that a high prediction confidence does not necessarily imply that the classification is meaningful or correct. The prediction confidence pattern across interaction states is similar for both categories, with Map showing the highest confidence. Interestingly, this pattern is comparable to that of similar map apps and Google Maps.

The number of app sessions classified per state and the share of total app session duration have similar patterns for communication and the remaining apps. Map is by far the most used interaction state, followed by Direction and small shares of Search and Place. This differs from public transport apps but shares some similarities with the pattern observed for Google Maps.

One objective of this comparison was to evaluate whether the classifier finds the same distribution patterns regardless of the dataset. The observed divergences of more than 10% between different app combinations indicate that this is not the case.

The high share of Map in apps that likely do not include a map feature is interesting. One possible interpretation is that Map reflects passive interaction patterns, such as scrolling and exploring content, whereas Direction and Search mainly consist of active interactions, such as typing. Under that assumption, the classifier’s results indicate that users spend most of their time exploring content (71-75%) and less time typing (24-28%). However, this is just an assumption and cannot be verified with the available datasets.

In relation to research question 3, these results suggest that the identified interaction patterns are not unique to map apps because the distributions in non-map apps are not uniform. This finding indicates that there are comparable user behaviours, such as active and passive interaction, across different types of apps.

5.5 Limitations and Future Research

This thesis explored the approach of collecting and labelling GT data to train an SMLC and then applying that classifier to real-world tappigraphy data. This complex and exploratory process has several limitations, which may affect the accuracy and generalisability of the final classification.

There is no prior study using GT data to analyse patterns in tappigraphy at the activity level, leading to a lack of related research, especially to guide design decisions of the GT data collection. Due to limited resources, GT data was collected in a controlled environment (lab) rather than in real-world environments. While this controlled setting ensures that differences in tapping behaviour were almost exclusively influenced by the performed tasks rather than environmental factors, it limits ecological validity.

For Google Maps, the four defined interaction states were based on the framework by Savino et al. (2021). However, the exact tasks that the participants had to carry out were specified by the researcher. The task for Map was to explore restaurant options, which often included checking their place details. This led to a certain overlap between Map and Place, which later resulted in misclassifications between the two states. Future work could address this issue by applying a finer-grained classification beyond a single interaction state per app session. The state Direction involved route searching and selection, but participants did not move in space. This restricts the ecological validity of this state. Moreover, the Google Maps GT data contained imbalances among interaction states, which can be an issue for machine learning models and should be improved in future data collection designs.

Future work should take the environmental aspect of context into account. Especially, the interaction state Direction cannot be fully replicated in a stationary lab setting. If different usage environments lead to differences in tapping patterns, this would allow to research typical usage environments of map apps using tappigraphy.

For the SBB app, there was no prior study on within-app usage. Therefore, new interaction states had to be defined by the researcher. The chosen states Buy and Check often produced very short sessions, limiting the possibility of analysing tappigraphy patterns. Moreover, around half of the participants were not familiar with some of the SBB functions, which might have led to slower tapping in the SBB app. All these constraints made the training of an SMLC specific to public transport apps unfeasible. Nevertheless, such a classifier could provide valuable insights into the usage of public transport apps. However, a prior study on typical interaction states would be needed. Future research should address this gap, as there is a general lack of within-app research, especially for public transport apps.

The sample size of 20 participants was relatively small. However, the main goal of this data collection was to explore the feasibility and potential of the GT matching approach rather than to achieve statistical representativeness. The participant group was relatively homogenous in terms of demographics and app usage experience. Half of the participants are iPhone users, which could influence tapping speed on the provided Android phone. The small sample size and demographic homogeneity of the participants limit the generalisability and robustness of the classifier, as factors such as age influence tapping speed.

Another limitation concerns feature engineering and input selection for the SMLC. The derived metrics were informed by prior research in tappigraphy (Ceolini et al., 2022; Zingaro & Reichenbacher, 2022). However, this related work did not focus on training a classifier; therefore, the choice of input features in this thesis was driven by trial and error, guided by model performance scores, which does not guarantee optimal feature selection. Some of the chosen features were highly correlated with each other, which may lead to overfitting despite mitigation through hyperparameter tuning.

The selection of the best SMLC was also guided by trial and error, due to a lack of comparable studies. The performance differences of the different SMLCs were minor. The chosen SVC classifier was trained on the entire upsampled Google Maps GT dataset. The size of this dataset is small, which limits the model's robustness and generalisability. Future GT data collections would benefit from including more participants.

A critical limitation in interpreting the classification of the real-world data is the absence of a labelled real-world validation dataset. Instead, the prediction confidence and the distribution of the interaction state shares compared to existing research were considered. In particular, the prediction confidence should be treated with care, as a high score does not guarantee accurate predictions.

Broadly, future tappigraphy research could extend beyond map-based applications to explore tapping patterns across different app types. Exploring if certain patterns correspond to specific usage behaviour, such as scrolling and exploration versus typing for searching, texting, etc., could shed light on within-app usage behaviour.

Furthermore, the GT data collection included questionnaires on participants' spatial anxiety and orientation skills. Analysing this data would have gone beyond the scope of this thesis. However, it would be interesting for future work to analyse if, for example, relying on a specific state may be related to anxiety levels. There could be connections, like when solely relying on Direction, the user doesn’t orient themselves within the map, so they feel more lost.

This could be connected to prior tappigraphy research on anxiety leading to fragmented map usage (Zingaro et al., 2024a).

By acknowledging current limitations and addressing them through refined GT data collection and model training, future studies could strengthen the interpretability of tappigraphy-based use context analysis and contribute to the development of more adaptive, context-aware, and user-centred mobile applications.

6 Conclusion

This thesis explored the within-app usage patterns of mobile map apps depending on the usage context, focusing on the activity dimension. The goal was to evaluate the feasibility of labelling task-specific tapping patterns to enhance the interpretability of real-world tappigraphy data on the example of Google Maps and the SBB App.

To ensure ecological validity of the GT data collection, common map app usage activities and environments were reviewed and compared (see Table 1), providing the base for this data collection design. For Google Maps, the four interaction states defined by Savino et al. (2021) were used as the labelled categories for classifier training, allowing direct comparison to prior research. For the SBB app, there was a lack of studies on within-app usage, leading to design limitations, which made training a separate classifier with meaningful categories unfeasible.

The classifier's input feature selection process was led by exploratory analysis and informed by prior studies using tappigraphy (Ceolini et al., 2022; Zingaro & Reichenbacher, 2022). Highly correlated features were reduced to mitigate introducing noise and overfitting. Finally, the SVC was trained on labelled and upsampled Google Maps GT data and applied to previously collected real-world tappigraphy data from 2021. The resulting classification was analysed based on the prediction confidence per state and the distribution of the interaction state share across app combinations in relation to prior research.

The classification of the real-world Google Maps data revealed usage patterns aligning with the findings of Savino et al. (2021) concerning the time share of the interaction states per app session. The distribution of time spent in each classified interaction state across all app sessions diverged only marginally from the reference dataset. The largest difference was a higher share of Direction in the real-world dataset, which might have been caused by participants exploring new outdoor spaces, due to COVID-19 restrictions on public areas.

Notably, Savino et al. (2021) analysed the share of time spent in each interaction state within one app session, whereas the SVC classifier in this thesis was designed to classify only one interaction state per app session. This coarser level of analysis was chosen due to data granularity and limited resources, leading to reduced comparability and resolution. Additional analysis with synthetic sessions showed a tendency towards classifying mixed sessions as Map, likely due to the longer average session length of Map. However, mixed app sessions should not be entirely misclassified, due to the set prediction confidence threshold (0.55).

Broadening the analysis to similar map apps, public transport apps, communication and remaining apps showed the generalizability and limitations of the classifier. Direction was the dominating interaction state for similar map apps and public transport apps, which reflects app functionalities and assumed user behaviour. For the communication and remaining apps, the state distribution was not uniform, possibly indicating that the tapping patterns found describe universal interactions beyond map apps, like scrolling and exploring content versus typing.

In answer to research question 3, this thesis found a clear alignment between ground truth and real-world tapping patterns. Specifically, for Google Maps the patterns align with prior findings, indicating a realistic classification. Furthermore, the labelled tapping patterns were not unique to map applications, but instead there seem to be comparable user behaviours across different type of apps.

Despite these findings, there were also several limitations. The need for prior within-app usage studies to define meaningful classification categories for the SMLC limits its applicability to apps that lack such research, complicating broader analysis across diverse applications. Furthermore, due to resource limitations, the GT dataset was collected in a controlled lab setting and the sample size was small. Additionally, the classification on app session level limits insights into more granular user activity.

The framework established in this thesis provides a scalable method for studying task-specific app usage while maintaining user privacy and ecological validity. Future research should focus on refining GT data collection to include more dimensions of context, such as usage environments and more diverse user groups as well as finer grained classification on within app session level. Gaining detailed insights into usage context of map applications can inform the design of adaptive mobile map applications, ultimately improving their usability. Future work could further explore the definition of interaction states in other apps to explore their respective activity specific tapping patterns.

References

Arinze, C. P. (2024). Effective Strategies for Handling Noisy Data in Machine Learning. *Medium*. Retrieved September 20, 2025, from https://medium.com/@InsightCoder/effective-strategies-for-handling-noisy-data-in-machine-learning-79f02f216b63

Ashraf, A. (2023). Correlation in machine learning—All you need to know. *Medium*. Retrieved September 15, 2025, from https://medium.com/@abdallahashraf90x/all-you-need-to-know-about-correlation-for-machine-learning-e249fec292e9

Bakırarar, B., & Elhan, A. (2023). Class Weighting Technique to Deal with Imbalanced Class Problem in Machine Learning: Methodological Research. *Turkiye Klinikleri Journal of Biostatistics*, *15*, 19–29. https://doi.org/10.5336/biostatic.2022-93961

Böhmer, M., Hecht, B., Schöning, J., Krüger, A., & Bauer, G. (2011). Falling asleep with Angry Birds, Facebook and Kindle: A large scale study on mobile application usage. *Proceedings of the 13th International Conference on Human Computer Interaction with Mobile Devices and Services*, 47–56. https://doi.org/10.1145/2037373.2037383

Borger, J. N., Huber, R., & Ghosh, A. (2019). Capturing sleep–wake cycles by using day-to-day smartphone touchscreen interactions. *Npj Digital Medicine*, *2*(1), 1–8. https://doi.org/10.1038/s41746-019-0147-4

Brown, B., McGregor, M., & Laurier, E. (2013). iPhone in vivo: Video analysis of mobile device use. *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 1031–1040. https://doi.org/10.1145/2470654.2466132

Carrascal, J. P., & Church, K. (2015). An In-Situ Study of Mobile App & Mobile Search Interactions. *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, 2739–2748. https://doi.org/10.1145/2702123.2702486

Ceolini, E., Kock, R., Band, G. P. H., Stoet, G., & Ghosh, A. (2022). Temporal clusters of age-related behavioral alterations captured in smartphone touchscreen interactions. *iScience*, *25*(8). https://doi.org/10.1016/j.isci.2022.104791

COVID-19 pandemic in Germany. (2025). In *Wikipedia*. Retrieved September 5, 2025, from https://en.wikipedia.org/w/index.php?title=COVID-19_pandemic_in_Germany&oldid=1299721533

COVID-19 pandemic in Switzerland. (2025). In *Wikipedia*. Retrieved September 5, 2025, from https://en.wikipedia.org/w/index.php?title=COVID-19_pandemic_in_Switzerland&oldid=1301879531

Do, T. M. T., Blom, J., & Gatica-Perez, D. (2011). Smartphone usage in the wild: A large-scale analysis of applications and context. *Proceedings of the 13th International Conference on Multimodal Interfaces*, 353–360. https://doi.org/10.1145/2070481.2070550

Duckrow, R. B., Ceolini, E., Zaveri, H. P., Brooks, C., & Ghosh, A. (2021). Artificial neural network trained on smartphone behavior can trace epileptiform activity in epilepsy. *iScience*, *24*(6), 102538. https://doi.org/10.1016/j.isci.2021.102538

Eskandar, S. (2023). Introduction to RBF SVM: A Powerful Machine Learning Algorithm for Non-Linear Data. *Medium*. Retrieved September 1, 2025, from https://medium.com/@eskandar.sahel/introduction-to-rbf-svm-a-powerful-machine-learning-algorithm-for-non-linear-data-1d1cfb55a1a

Fonseca, F., Conticelli, E., Papageorgiou, G., Ribeiro, P., Jabbari, M., Tondelli, S., & Ramos, R. (2021). Use and Perceptions of Pedestrian Navigation Apps: Findings from Bologna and Porto. *ISPRS International Journal of Geo-Information*, *10*(7), Article 7. https://doi.org/10.3390/ijgi10070446

GeeksforGeeks. (2025). *Supervised Machine Learning*. GeeksforGeeks. Retrieved August 26, 2025, from https://www.geeksforgeeks.org/machine-learning/supervised-machine-learning/

Griffin, A. L., Reichenbacher, T., Liao, H., Wang, W., & Cao, Y. (2024). Cognitive issues of mobile map design and use. *Journal of Location Based Services*, *0*(0), 1–31. https://doi.org/10.1080/17489725.2024.2371288

Griffin, A. L., White, T., Fish, C., Tomio, B., Huang, H., Sluter, C. R., Bravo, J. V. M., Fabrikant, S. I., Bleisch, S., Yamada, M., & Picanço, P. (2017). Designing across map use contexts: A research agenda. *International Journal of Cartography*, *3*(sup1), 90–114. https://doi.org/10.1080/23729333.2017.1315988

Gualandi, S., & Toscani, G. (2019). Human behavior and lognormal distribution. A kinetic description. *Mathematical Models and Methods in Applied Sciences*, *29*(04), 717–753. https://doi.org/10.1142/S0218202519400049

He, C., & Hegarty, M. (2020). How anxiety and growth mindset are linked to navigation ability: Impacts of exploration and GPS use. *Journal of Environmental Psychology*, *71*, 101475. https://doi.org/10.1016/j.jenvp.2020.101475

Hegarty, M. (2002). Development of a self-report measure of environmental spatial ability. *Intelligence*, *30*(5), 425–447. https://doi.org/10.1016/S0160-2896(02)00116-2

Huang, H., Gartner, G., Krisp, J. M., Raubal, M., & Van de Weghe, N. (2018). Location based services: Ongoing evolution and research agenda. *Journal of Location Based Services*, *12*(2), 63–93. https://doi.org/10.1080/17489725.2018.1508763

Kim, J., Chang, Y., Chong, A. Y. L., & Park, M.-C. (2019). Do perceived use contexts influence usage behavior? An instrument development of perceived use context. *Information & Management*, *56*(7), 103155. https://doi.org/10.1016/j.im.2019.02.010

Li, T., Li, Y., Hoque, M. A., Xia, T., Tarkoma, S., & Hui, P. (2022). To What Extent We Repeat Ourselves? Discovering Daily Activity Patterns Across Mobile App Usage. *IEEE Transactions on Mobile Computing*, *21*(4), 1492–1507. IEEE Transactions on Mobile Computing. https://doi.org/10.1109/TMC.2020.3021987

Li, T., Xia, T., Wang, H., Tu, Z., Tarkoma, S., Han, Z., & Hui, P. (2022). Smartphone App Usage Analysis: Datasets, Methods, and Applications. *IEEE Communications Surveys & Tutorials*, *24*(2), 937–966. https://doi.org/10.1109/COMST.2022.3163176

MapOnTap. (2021). UZH – Digital Society Initiative – Mobility. Retrieved July 21, 2025, from https://mobility.dsi.uzh.ch/project/mapontap/

Muhammad, I., & Yan, Z. (2015). SUPERVISED MACHINE LEARNING APPROACHES: A SURVEY. *ICTACT Journal on Soft Computing*, *05*(03), 946–952. https://doi.org/10.21917/ijsc.2015.0133

Muller, M., Wolf, C. T., Andres, J., Desmond, M., Joshi, N. N., Ashktorab, Z., Sharma, A., Brimijoin, K., Pan, Q., Duesterwald, E., & Dugan, C. (2021). Designing Ground Truth and the Social Life of Labels. *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, 1–16. https://doi.org/10.1145/3411764.3445402

Murel, J. (2024). *What is upsampling? | IBM*. Retrieved September 3, 2025, from https://www.ibm.com/think/topics/upsampling

Otebolaku, A. M., & Andrade, M. T. (2016). User context recognition using smartphone sensors and classification models. *Journal of Network and Computer Applications*, *66*, 33–51. https://doi.org/10.1016/j.jnca.2016.03.013

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, *12*(85), 2825–2830.

Reichenbacher, T. (2004). *Mobile Cartography: Adaptive Visualisation of Geographic Information on Mobile Devices* [Technische Universität München]. https://mediatum.ub.tum.de/?id=601066

Reichenbacher, T., Aliakbarian, M., Ghosh, A., & Fabrikant, S. I. (2022). Tappigraphy: Continuous ambulatory assessment and analysis of in-situ map app use behaviour. *Journal of Location Based Services*, *16*(3), 181–207. https://doi.org/10.1080/17489725.2022.2105410

Roth, R. E., Çöltekin, A., Delazari, L., Filho, H. F., Griffin, A., Hall, A., Korpi, J., Lokka, I., Mendonça, A., Ooms, K., & van Elzakker, C. P. J. M. (2017). User studies in cartography: Opportunities for empirical research on interactive maps and visualizations. *International Journal of Cartography*, *3*(sup1), 61–89. https://doi.org/10.1080/23729333.2017.1288534

Savino, G.-L., Sturdee, M., Rundé, S., Lohmeier, C., Hecht, B., Prandi, C., Nunes, N. J., & Schöning, J. (2021). MapRecorder: Analysing real-world usage of mobile map applications. *Behaviour & Information Technology*, *40*(7), 646–662. https://doi.org/10.1080/0144929X.2020.1714733

Shiffman, S., Stone, A., & Hufford, M. (2008). Ecolocial Momentary Assessment. *Annual Review of Clinical Psychology*, *4*, 1–32. https://doi.org/10.1146/annurev.clinpsy.3.022806.091415

Singh, A., Thakur, N., & Sharma, A. (2016). A review of supervised machine learning algorithms. *2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom)*, 1310–1315. https://ieeexplore.ieee.org/document/7724478/

swisstopo—Apps on Google Play. (n.d.). Retrieved September 24, 2025, from https://play.google.com/store/apps/details?id=ch.admin.swisstopo&hl=en

Taps.ai. (n.d.). Retrieved July 2, 2025, from https://www.taps.ai/login

Tian, Y., Zhou, K., Lalmas, M., & Pelleg, D. (2020). Identifying Tasks from Mobile App Usage Patterns. *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, 2357–2366. https://doi.org/10.1145/3397271.3401441

Xu, Q., Erman, J., Gerber, A., Mao, Z., Pang, J., & Venkataraman, S. (2011). Identifying diverse usage behaviors of smartphone apps. *Proceedings of the 2011 ACM SIGCOMM Conference on Internet Measurement Conference*, 329–344. https://doi.org/10.1145/2068816.2068847

Yang, L., Yuan, M., Wang, W., Zhang, Q., & Zeng, J. (2016). *Apps on the move: A fine-grained analysis of usage behavior of mobile apps*. 1–9. https://doi.org/10.1109/INFOCOM.2016.7524464

Zingaro, D., Bartling, M., & Reichenbacher, T. (2023). Exploring Map App Usage Behaviour Through Touchscreen Interactions (Short Paper) [Application/pdf]. *LIPIcs, Volume 277, GIScience 2023*, *277*, 95:1-95:6. https://doi.org/10.4230/LIPICS.GISCIENCE.2023.95

Zingaro, D., & Reichenbacher, T. (2022). *Exploratory analysis of mobile app usage in relation to distance from home*. https://doi.org/10.5167/UZH-224336

Zingaro, D., Reichenbacher, T., Bartling, M., & Fabrikant, S. I. (2024a). *Exploring the Relation Between Sense of Direction and Spatial Anxiety in Everyday Mobile Map App Use*.

Zingaro, D., Savino, G.-L., Reichenbacher, T., Schöning, J., & Fabrikant, S. I. (2024b). *Tapping into Mobile App Use: What Touch Event Data Can Reveal About Active and Passive In-App Usage Behaviour*. SSRN. https://doi.org/10.2139/ssrn.4768783

Appendix

Table 10: Task list for group A with corresponding interaction state (for group B the same blocks were used in a different order)

SBB Block 1	Interaction State
Look up train connection from here to Solothurn, save this route	Search
Press ticket, choose return and 2nd class (do not actually buy it)	Buy
Check for valid tickets or abonnements, look at the ticket of today	Check
Search the quickest connection from Zurich HB to "Solothurn, Amthausplatz", save this route	Search
Look at the map that shows you were to find the bus and orientate yourself (click walk)	Map
Search the quickest connection from Solothurn to Basel SBB, save this route	Search
Check saved routes and open the one from Solothurn to Basel SBB and check the details	Check
Search best connection from Solothurn to "Voltaplatz" at 1 o'clock tomorrow , save this route	Search
Look at the map that shows you were to find the tram and orientate yourself (click walk)	Map
Search best connection from "Voltaplatz" to Basel SBB at 2 o'clock on Tuesday , save this route	Search
Check saved routes and open the one you just saved (Voltaplatz to Basel SBB) and check the details	Check
Search the quickest connection from Basel SBB to Zurich HB, save this route	Search
Look up train connection from here to Zurich Sihlcity, save this route	Search
Look at the map that shows you how to walk there from the last stop and orientate yourself (click walk)	Map
Look up train connection from here to Zurich HB, save this route	Search
Press ticket, choose return and 2nd class (do not actually buy it)	Buy
Check for valid tickets or abonnements, look at the ticket of today	Check
End of SBB Block 1 -> delete saved routes	

Google Maps Block 2	
Search for Solothurn	Search
Explore Solothurn as if you are planning to have dinner there - what options are there	Map
Search for “Restaurant Tiger” (in Solothurn)	Search
Look at the details - figure out opening hours, look at pictures	Place
Search Basel	Search
Look at the details - pictures etc	Place
Plan a route from Solothurn to Basel using public transport and choose a specific route	Direction
Explore Basel as if you are planning to have dinner there - what options are there	Map
Search the Zara in Basel	Search
Look at the details - opening hours etc.	Place
Search Zürich	Search
Look at the details - pictures etc	Place
Plan a route from Basel to Zurich using the car and start it	Direction
Explore Zurich as if you are planning to have dinner there - what options are there	Map
Search the "Old Inn" (Zurich)	Search
Look at the details - opening hours etc.	Place
Plan a route from your current location to the Old Inn by public transport and click a specific one	Direction
Search the "Sim Sim" (Oerlikon)	Search
Look at the details - opening hours etc.	Place
Plan a route from your current location to the "Sim Sim" by foot and start it	Direction
End of Google Maps Block 2	
Break	
SBB Block 2	
Search the quickest connection from Solothurn to Basel SBB, save this route	Search

Check saved routes and open the one from Solothurn to Basel SBB and check the details	Check
Search best connection from Solothurn to "Voltaplatz" at 1 o'clock tomorrow , save this route	Search
Look at the map that shows you were to find the tram and orientate yourself (click walk)	Map
Look up train connection from here to Zurich Sihlcity, save this route	Search
Look at the map that shows you how to walk there from the last stop and orientate yourself (click walk)	Map
Look up train connection from here to Solothurn, save this route	Search
Press ticket, choose return and 2nd class (do not actually buy it)	Buy
Check for valid tickets or abonnements, look at the ticket of today	Check
Search best connection from "Voltaplatz" to Basel SBB at 2 o'clock on Tuesday , save this route	Search
Check saved routes and open the one you just saved (Voltaplatz to Basel SBB) and check the details	Check
Look up train connection from here to Zurich HB, save this route	Search
Press ticket, choose return and 2nd class (do not actually buy it)	Buy
Check for valid tickets or abonnements, look at the ticket of today	Check
Search quickest connection from Basel to Zurich HB, save this route	Search
Search shortest connection from Zurich HB to "Solothurn, Amthausplatz", save this route	Search
Look at the map that shows you were to find the bus and orientate yourself (click walk)	Map
End of SBB Block 2 -> delete saved routes	
Google Maps Block 1	
Search Basel	Search
Look at the details - pictures etc	Place
Plan a route from Solothurn to Basel using public transport and choose a specific route	Direction
Explore Basel as if you are planning to have dinner there - what options are there	Map

Search the Zara in Basel	Search
Look at the details - opening hours etc.	Place
Search Zürich	Search
Look at the details - pictures etc	Place
Plan a route from Basel to Zurich using the car and start it	Direction
Search for “Restaurant Tiger” (in Solothurn)	Search
Look at the details - figure out opening hours, look at pictures	Place
Search for Solothurn	Search
Explore Solothurn as if you are planning to have dinner there - what options are there	Map
Search the "Old Inn" (in Zurich)	Search
Look at the details - opening hours etc.	Place
Plan a route from your current location to the Old Inn by public transport and click a specific one	Direction
Explore Zurich as if you are planning to have dinner there - what options are there	Map
Search the "Sim Sim" (Oerlikon)	Search
Look at the details - opening hours etc.	Place
Plan a route from your current location to the "Sim Sim" by foot and start it	Direction
End of Google Maps Block 1	

Table 11: Names of the apps included in the app combination public transport

Apps Included in the App Combination Public Transport
SBB Mobile
VBZ Fahrinfo
DB Navigator
ZVV Fahrplan
JR East App
MVG Fahrinfo München
ZVV OneApp
Eurail/Interrail Rail Planner
Trainline
BVG Fahrinfo Berlin
Treni Lite
Ticket BV
Tokyo Metro Map
Wemlin
ProntoTreno
SNCF
Carris

Personal Declaration

I hereby declare that the submitted Thesis is the result of my own, independent work. All external sources are explicitly acknowledged in the Thesis.

Additionally, ChatGPT-4 was used to improve the Python code, while both ChatGPT-4 and Grammarly were used to improve the grammar and flow of the text.

Oliva Schilling, 23.10.2025