



**University of
Zurich^{UZH}**

Indoor-outdoor detection with GPS and IMU data

GEO 511 Master's Thesis

Author

Chenxi Jiang
22-738-272

Supervised by

Dr. Cheng Fu
Prof. Dr. Robert Weibel

Faculty representative

Prof. Dr. Robert Weibel

03.02.2025

Department of Geography, University of Zurich

Abstract

Building upon the research gaps identified in the Mobility, Activity, and Social Interactions Study (MOASIS), which highlighted the need for a more nuanced understanding of mobility patterns among older adults, this study focuses on indoor and outdoor environment detection. Utilizing the “uTrail” device, which integrates GPS and Inertial Measurement Unit (IMU) sensors, I collected data to classify environments into indoor, outdoor, and semi-indoor categories. Data collection was facilitated through the “uTrail” device in conjunction with the custom-developed “Label Logger” software, enabling efficient and accurate data acquisition. Subsequently, a Random Forest model was employed to train and evaluate the classification performance. The feature engineering process was designed to utilize the time-window method, enabling the capture of temporal dependencies. This approach facilitated an in-depth exploration of the optimal time window size and its influence on model performance. Additionally, the impact of different input data configurations on classification accuracy was investigated to assess how variations in data structure affect the model. The analysis further examined whether individual differences and environmental contexts contribute to variations in model performance. This study contributes to the future research of MOASIS and offers a comprehensive framework for environmental classification in real-world mobility studies.

Keywords: Random Forest, Indoor-outdoor detection, semi-indoor, MOASIS, GPS, IMU

Acknowledgments

I would like to express my deepest gratitude to Prof. Dr. Robert Weibel and Dr. Cheng Fu for their continuous support, guidance, encouragement, and invaluable feedback throughout the course of this thesis. Their insights and constructive suggestions have been instrumental in shaping the direction of my research, and for that, I am eternally grateful. Without their constant encouragement and motivation, the completion of this thesis would not have been possible. Their dedication and belief in my work provided the momentum I needed to overcome challenges and stay focused throughout the process.

I would also like to extend my heartfelt appreciation to my parents, whose unwavering support, understanding, and belief in me have been a constant source of motivation throughout my academic journey.

Last but not least, I would like to thank myself for persevering through moments of procrastination and self-doubt, and for having the determination to see this thesis through to completion.

Contents

Abstract.....	1
Acknowledgments	2
List of Tables	6
List of Figures.....	8
1. Introduction.....	9
2. Background	12
2.1 Motivation.....	12
2.1.1 Imprecise Indoor Geolocation	12
2.1.2 Further Research of MOASIS.....	13
2.2 Introduction to GPS and IMU sensors	15
2.2.1 GPS	15
2.2.2 IMU.....	17
2.2.2.1 Accelerometer	18
2.2.2.2 Gyroscope	18
2.2.2.3 Magnetometer	20
2.3 Indoor/outdoor detection with different sensors	20
2.3.1 Indoor-outdoor Detection based on GNSS Sensors.....	20
2.3.2 Indoor-outdoor Detection Based on Inertial Sensors.....	21
2.4 Indoor-outdoor Detection with Machine Learning Methods.....	22
2.4.1 Machine Learning	22
2.4.2 Role and Approaches of Machine Learning in IO Detection	23
2.5 Research Gaps.....	24
2.6 Research Objective and Research Questions	25
3. Methodology	27
3.1 Concept	27
3.1.1 Indoor, Outdoor, and Semi-indoor.....	27
3.1.2 Scenarios	28
3.2 Data Collection	30
3.2.1 Data Collection Workflow	30

3.2.2 “uTrail” Device.....	32
3.2.3 “Label-logger”	33
3.2.3.1 Home Page.....	35
3.2.3.2 Scenario Page.....	37
3.2.3.3 Log Page	38
3.3 Data Preprocessing.....	39
3.4 Implementation of Model Training.....	40
3.4.1 Model Training	40
3.4.1.1 Time Series Split.....	41
3.4.1.2 Cross Validation.....	42
3.4.1.3 Hyperparameter.....	43
3.4.2 Feature Engineering.....	44
3.4.3 Training Strategy	45
3.5 Evaluation	47
3.5.1 Precision, Recall and F1-score.....	47
3.5.2 Shapiro-Wilk Test and One-way ANOVA.....	47
4. Results	49
4.1 Data overview	49
4.2 Hyperparameter Chosen.....	52
4.3 Comparison of train by mix-data with Different Time Windows.....	53
4.3.1 Result of Different Time Windows	53
4.3.2 Result Comparison.....	57
4.3.3 One-way ANOVA	58
4.3.4 Feature Importance	59
4.4 Comparison of Training by Time-series Data	60
4.4.1 Training Results Averaged by Person	61
4.4.2 Training Results Averaged by Scenario	64
4.5 Comparison of Training with and without Semi-indoor.....	69
5. Discussion	72
5.1 Best Time Window Size	72
5.2 The Comparison of Models between Different Training Strategies.....	73
5.2.1 Model Trained with Mix Data and with Time-series Data.....	73
5.2.2 Model Trained with and without Semi-indoor Dataset	74
5.2.3 Best Model.....	74
5.3 The Influencing Factor of Model Performance	75

5.3.1 The Relationship between Time Window Size on Model Training	75
5.3.2 The Relationship between Participant and Model Training	76
5.3.3 The Relationship between Scenario and Model Training.....	76
5.3.4 Feature Importance	77
5.4 Prediction Visualization.....	78
5.5 Comparison with Related Research	82
5.6 Limitation.....	84
6. Conclusion	86
Reference	91
Appendix.....	97
Appendix A.....	97
Appendix B	99
Appendix C	105
Personal declaration	106

List of Tables

Table 1 Model features	45
Table 2 Scenario Overview.....	49
Table 3 Typical places of each environment type	51
Table 4 Data ratio of each environment.....	52
Table 5 Confusion matrix when lag range =0.....	54
Table 6 Classification report when lag range =0	54
Table 7 Confusion matrix when lag range =3.....	54
Table 8 Classification report when lag range =3	54
Table 9 Confusion matrix when lag range =5.....	55
Table 10 Classification report when lag range =5	55
Table 11 Confusion Matrix when lag range =10	55
Table 12 Classification report when lag range=10	55
Table 13 Confusion matrix when lag range =15.....	56
Table 14 Classification report when lag range =15	56
Table 15 Confusion matrix when lag range=20.....	56
Table 16 Classification report when lag range=20	56
Table 17 Test set precision, recall, and F1-score when lag range (0-20)	57
Table 18 Accuracy and training time of lag range (0-20).....	58
Table 19 Histogram of accuracy of lag range (0-20).....	58
Table 20 Shapiro-Wilk test when lag range (0-20).....	59
Table 21 One-way ANOVA result when lag range (0-20)	59
Table 22 Top 5 important features of lag range (0-20).....	60
Table 23 Model performance averaged by person.....	61
Table 24 Shapiro-Wilk test of precision averaged by person	62
Table 25 Shapiro-Wilk test recall averaged by person	62
Table 26 Shapiro-Wilk test F1-score averaged by person	63
Table 27 One-way ANOVA result of precision, recall, F1-score averaged by person	63
Table 28 Top 5 importance features of each person	64

Table 29 Model performance averaged by scenario	65
Table 30 Shapiro-Wilk test of precision averaged by scenario	66
Table 31 Shapiro-Wilk test of recall averaged by scenario	66
Table 32 Shapiro-Wilk test of F1-score averaged by scenario	66
Table 33 One-way ANOVA result of precision, recall, F1-score averaged by scenario	66
Table 34 Top 5 importance features of each scenario	67
Table 35 Confusion matrix of a good performance example in City Center Scenario	68
Table 36 Classification report of a good performance example in City Center Scenario	68
Table 37 Confusion matrix of a bad performance example in Irchel Campus Scenario	68
Table 38 Classification report of a bad performance example in Irchel Campus Scenario	68
Table 39 Confusion matrix of mix-data training without semi-indoor	69
Table 40 Classification report of mix-data training without semi-indoor	69
Table 41 Model performance averaged by person without semi-indoor	70
Table 42 Model performance averaged by scenario without semi-indoor	70

List of Figures

Figure 1 GPS Indoor Trajectory Example	10
Figure 2 Data Collection Workflow	32
Figure 3 “Label Logger” Overview	34
Figure 4 “Label Logger” Home Page	35
Figure 5 “Label Logger” Scenario Page	37
Figure 6 “Label Logger” Log Page.....	38
Figure 7 Time series cross-validation illustration.....	42
Figure 8 Outdoor scenes	50
Figure 9 Semi-indoor scenes.....	50
Figure 10 Indoor scenes.....	50
Figure 11 Comparison of Two Cases of GPS Points in Irchel Tram Station	51
Figure 12 Histogram of test set precision, recall and F1-score when lag range (0-20)	57
Figure 13 Histogram of model performance averaged by person.....	61
Figure 14 Histogram of model performance averaged by scenario	65
Figure 15 Histogram of model performance averaged by person without semi-indoor	70
Figure 16 Histogram of model performance averaged by scenario without semi-indoor	71
Figure 17 Map visualization example 1	79
Figure 18 Map visualization example 2.....	79
Figure 19 Map visualization example 3.....	80
Figure 20 Map visualization example 4.....	81
Figure 21 Map visualization example 5.....	81
Figure 22 Map visualization example 6.....	82

1. Introduction

The location context change is essential information for the further modeling like activity tracking, routine analysis, and vehicle navigation. Especially for older individuals, location context recognition provides additional spatial facts about their daily mobile habits for older adults' health research and is instrumental and valuable in emergency response. Nowadays, the location is obtainable with the broad application of Global Navigation Satellite System (GNSS). However, signal attenuation occurs in any material medium (Lachapelle, 2004). The signal attenuation and non-homogeneities reflection indoors vary for different materials, generating big errors in indoor signals (Lachapelle, 2004). Therefore, geolocation is imprecise, even chaotic, within the outdoor-indoor transition and indoor environments, which poses a challenge for Indoor-outdoor Detection (IOD). Since the proposal of the concept of IOD, there has been a lot of influential literature on the use of external sensors to conduct IOD. For example, in 2012, an indoor-outdoor detection tool named IODetector was proposed using three lightweight, integrated sensors in the smartphone, i.e., light sensor, cellular module, and magnetism sensor, to detect indoor and outdoor environments fast (Zhou et al., 2018). As the faulty and abnormal behaviors of the GNSS positioning system are connected with the environment transition, taking advantage of this attribute and retrieving extra information from GNSS receivers is also a groundbreaking solution. (Zhu et al., 2019) extracted spatial geometry distribution, time sequence, and statistical features from GNSS measurements in Android smartphones and employed an ensemble model to innovate a high-accuracy strategy. Moreover, the authors take Signal-to-Noise Ratio (SNR) as the GNSS performance metric to generate IOD and reach an overall 95% accuracy. Thus, GNSS is still an unignorable and crucial foundational data for IOD.

This thesis is an extensive study of the Mobility, Activity, and Social Interactions Study (MOASIS) research project. The goal of MOASIS is to give an insight into older adults' daily activity profile and the association between their life, health, and well-being (Röcke et al., 2023). Mobility is an essential part of daily activity, but the relationship between older adults' health, well-being, and mobility still requires further exploration from diverse perspectives

(Röcke et al., 2023). IOD presents a view of spatial activity, which shows the pattern of activity across different environments.



Figure 1 GPS Indoor Trajectory Example

The current challenge of MOASIS study about mobility context is GPS geolocation reading scatter chaotically when getting indoors. For instance, when an older adult returns home, GPS points always fall a buffer zone around the house (as Figure 1), making it challenging to determine their exact arrival time. Furthermore, it typically takes several minutes for the GPS to reacquire a stable satellite signal. Under such circumstances, it becomes difficult to discern whether the individual is inside the house or working in the garden. To address the problem, except for indoor and outdoor environments, a semi-indoor environment should be considered. Semi-indoor is designated to mean activities conducted in spaces that are partially enclosed, such as a covered patio or a greenhouse. This concept could provide extensive study and facilitate the analysis of activities like commuting, gardening work, and park workouts.

To support MOASIS mobility research, this thesis aims to provide an innovative method to distinguish the environment from MOASIS data. The basis of the detection is to capture regular performance using a supervised data learning method to identify and categorize them as indoor, semi-indoor, and outdoor by learning regular behavior. Regarding data source, this thesis follows the framework of MOASIS project, leveraging data acquired from a custom-designed device called “uTrail” which contains a Global Positioning System (GPS) receiver and an

inertial measurement unit (IMU). The label for training is recorded by a self-developed application, “Label Logger”.

The research for this thesis was carried out within the framework of the research project “MOASIS”, carried out jointly between the UZH Healthy Longevity Center (HLC) and the Geographic Information Systems (GIS) Unit at the Department of Geography, University of Zurich (UZH). Data collection was completed with the support of the Geographic Information System Unit.

2. Background

2.1 Motivation

2.1.1 Imprecise Indoor Geolocation

Global Navigation Satellite Systems (GNSS), including GPS, Beidou, and Galileo, provide real-time positional and temporal data, enabling accurate navigation and localization. GNSS applications range from emergency services to disaster relief, tracking, and even virtual reality. While GNSS performs effectively outdoors, its reliability indoors is significantly reduced due to signal attenuation, multipath effects, and obstructions caused by structures (Puricer & Kovar, 2007). These challenges result in degraded localization accuracy, limited satellite visibility, and high noise levels (Lachapelle, 2004). Signal attenuation within buildings can reach up to 2.9 dB per meter of structure, necessitating receivers with high sensitivity to detect weak signals indoors (Xu et al., 2015).

To address these issues, techniques such as Assisted-GPS (A-GPS) and massive parallel correlation have been developed. A-GPS improves acquisition speed and sensitivity by narrowing the search space using satellite orbit data and Doppler shifts (Van Diggelen, 2002). High-sensitivity receivers further enhance performance through extended correlations, although this increases acquisition time and complexity. These innovations enable GNSS to function in partially sheltered environments, but signal degradation indoors still limits its standalone application.

In indoor environments, alternative technologies such as Wi-Fi, Bluetooth, infrared, and ultrasound are commonly employed. Wi-Fi positioning, based on Received Signal Strength (RSS) or fingerprinting, offers reasonable accuracy but requires extensive infrastructure and frequent updates (Mallik et al., 2023). Similarly, Bluetooth and infrared have limited range and coverage, while ultrasound systems achieve high accuracy but suffer from noise and interference. Inertial navigation systems (INS), which use accelerometers and gyroscopes, are

effective for short-term tracking but face cumulative error over time. Pseudolite systems, offering precise localization with GPS-like signals, provide another solution but require specialized equipment and installation (Xu et al., 2015).

Indoor-outdoor (I/O) detection serves as the foundation for seamless navigation, enabling systems to transition between GPS-based outdoor localization and indoor methods. By determining the user's environment, I/O detection activates appropriate technologies, optimizing energy use and ensuring continuous service. For example, it can disable GPS indoors to conserve power or improve Wi-Fi performance by recognizing indoor signal patterns (Bui et al., 2020). I/O detection also enhances user experience by adapting device settings, such as brightness and notifications, to environmental conditions.

I/O detection systems leverage various sensors and techniques. Light sensors distinguish environments based on ambient light intensity, while magnetic field sensors detect unique interference patterns indoors. Cellular signal analysis evaluates signal strength and variability to identify indoor or outdoor conditions. Hybrid approaches, like IODetector, combine light, cellular, and magnetic data for improved accuracy and energy efficiency (Zhou et al., 2018). Machine learning methods further enhance these systems by analyzing complex sensor patterns and adapting to dynamic environments, offering robust and scalable solutions.

2.1.2 Further Research of MOASIS

Daily activities are fundamental for older adults to achieve well-being and life satisfaction, as emphasized in aging theories. The Mobility, Activity, and Social Interactions Study (MOASIS) is an interdisciplinary project that aims to gain insights into the daily lives of healthy older adults over an extended period across diverse activity domains (Martin et al., 2015; Michel & Sadana, 2017). The study investigates the contextual determinants of the relationships between health, daily activities, and well-being by tracking older adults' spatial activities—examining where they are and what they do.

To achieve this, MOASIS employs a custom-designed mobile sensor, the uTrail, which collects data on spatial, physical, and social activities. The uTrail combines GPS tracking, a 3-axis accelerometer, a 3-axis magnetometer, and a MEMS microphone (Röcke et al., 2023). GPS data tracks daily spatial mobility, while the accelerometer and magnetometer measure velocity changes, orientation, and physical activity intensity. Over a 30-day observation period, data

2. Background

from 150 community-dwelling adults aged 65 and older were gathered, creating a robust dataset of their day-to-day routines and interactions.

To complement the objective data, MOASIS integrates smartphone-based experience sampling methods. Participants report their affect, stress levels, and perceived activities and complete cognitive tasks, such as an ambulatory working memory exercise. This multimodal approach creates detailed within-person profiles of activities, experiences, and contexts, which lifespan psychologists, gerontologists, and geographical information scientists analyze collaboratively. These analyses aim to identify patterns that influence health, mobility, and well-being, providing a foundation for designing targeted interventions to enhance cognitive and emotional functioning and overall quality of life for older adults (Röcke et al., 2023).

The project has significantly advanced our understanding of the daily lives of older adults by combining diverse ambulatory assessment methods such as passive sensing, self-reports, and cognitive testing. It offers insights into spatial, physical, and social activities and their associations with well-being, health, and cognitive performance. By maintaining high compliance rates and causing minimal disruption to participants' routines, MOASIS has collected a rich dataset that supports advanced analyses, including time-series modeling, machine learning, and multimodal data integration (Li et al., 2020; Paraschiv-Ionescu et al., 2012).

MOASIS's 30-day observation period uniquely captures inter- and intra-individual differences in behavior, enabling analyses that explore activity patterns and their implications for healthy aging. By leveraging longitudinal data, the project also contributes to global efforts, such as the WHO's Decade of Healthy Aging (2021–2030), to promote interventions for improved mobility, social engagement, and well-being. These findings form the basis for cross-cultural studies, with adaptations of the MOASIS framework already underway in Hong Kong and Mexico City (World Health Organization, 2020).

Despite its success in analyzing outdoor activities using reliable GPS data, MOASIS highlights the need to address the challenges of tracking activities in indoor and semi-indoor environments, where GNSS signals become unstable and chaotic. These environments require advanced tracking technologies to accurately localize participants and monitor transitions between indoor and outdoor spaces.

Inertial navigation provides a promising solution, with IMUs (Inertial Measurement Units) offering self-contained tracking by measuring angular velocity and linear acceleration to

determine position and orientation. IMU data can complement GPS, enhancing monitoring and localization in GNSS-compromised settings like indoor or semi-indoor environments (Röcke et al., 2023). This approach allows for a more complete understanding of activities in diverse contexts.

Recent research demonstrates the success of machine learning (ML) and deep learning (DL) in improving indoor-outdoor (IO) detection and localization (Li et al., 2020). By integrating ML and DL techniques with the rich MOASIS dataset, the project can develop precise methodologies for tracking mobility patterns in these challenging environments. These advancements will enable detailed analyses of activity transitions and their implications for health and mobility in aging populations.

Ultimately, extending MOASIS to refine IO detection addresses an important gap, leveraging its methodological innovations to explore mobility and health in environments where traditional localization approaches fall short.

2.2 Introduction to GPS and IMU sensors

This thesis utilizes data collected by the uTrail device, a custom-built mobile sensor used in the MOASIS project, to perform indoor-outdoor (IO) detection. The uTrail combines GPS and Inertial Measurement Unit (IMU) data, including a 3-axis accelerometer and 3-axis magnetometer, to track spatial mobility and physical activities with high precision.

2.2.1 GPS

The Global Navigation Satellite System (GNSS) has emerged as a vital technology for enabling real-time activity monitoring on a global scale. Its rapid growth initially gained momentum with the development of commercial applications, driven by advancements in navigation satellites and related equipment. The GNSS is defined as the group of all SNSs and their augmentations. There are 195 countries across the globe, but very few countries host their own navigation system through a specific satellite (Jiang et al., 2013). Globally four countries host navigation systems: GPS (US), GLObal NAVigation Satellite System (GLONASS of Russia), Galileo (EU), and BeiDou (China).

2. Background

The Global Positioning System (GPS) is a satellite-based navigation technology that provides location, velocity, and time synchronization data to GPS receivers on Earth. Developed initially for military use by the United States Department of Defense, GPS has become a cornerstone technology in civilian applications due to its high precision and global coverage. The system relies on a constellation of satellites transmitting signals that GPS receivers decode to calculate their position in real-time.

GPS is an indispensable tool with diverse applications across civilian and military domains. In civilian use, GPS supports transportation in aviation, maritime, shipping, and rail and road systems, enhancing safety and efficiency. It aids daily activities such as geotagging, carpools, navigation for the visually impaired, fitness monitoring, and emergency assistance. Outdoor activities like hiking, skiing, paragliding, and skydiving also benefit from GPS precision (Jacobson, 2017). Furthermore, it plays a critical role in agriculture (e.g., precision farming), disaster management, supply chain logistics, geospatial data collection, and scientific research, including environmental monitoring, tectonic studies, and communication network synchronization.

In the military, GPS offers higher precision through dual-frequency systems, enabling operations such as reconnaissance, surveillance, and navigation in unfamiliar territories. It supports GPS-guided missiles, tracking enemy movements, and battlefield logistics, including supply delivery and precise target navigation. Its versatility and accuracy make GPS a vital technology for modern society in both civilian and defense sectors (Kumar et al., 2021).

The Global Positioning System (GPS) calculates a receiver's position by determining the time and location of signals transmitted by satellites. GPS satellites are equipped with highly stable atomic clocks that are synchronized with each other and with ground receivers. Any time offset in ground receivers is corrected daily to maintain accuracy. Similarly, the satellite positions are continuously monitored and controlled with great precision. Each satellite transmits a radio signal containing the current time and its location, allowing GPS receivers to receive signals passively without needing to send a request.

The distance between a satellite and a GPS receiver is determined by calculating the delay between the transmitted and received signals, using the constant speed of radio waves. By measuring distances from multiple satellites, a GPS receiver calculates its precise position and its deviation from real-time. To achieve accurate positioning, the receiver requires signals from

at least four satellites, solving for four unknown quantities in the positioning equation (e.g., latitude, longitude, altitude, and time offset) (Roewe et al., 2018).

The GPS satellite system consists of a constellation of 24 to 32 satellites in medium Earth orbit. Initially, the GPS operated with 24 satellites distributed across three circular orbital planes, each containing eight satellites (Tinder, 2007). The system was later updated to six orbital planes with four satellites in each plane. The satellites have an orbital period of nearly 12 hours (11 hours and 58 minutes), allowing them to pass over the same locations daily. This configuration ensures that at least six satellites are visible from any point on Earth's surface, providing the necessary coverage for continuous global positioning (Bui et al., 2020).

2.2.2 IMU

An inertial measurement unit (IMU) is an electronic device that measures and reports body force, angular velocity, and, in some cases, the orientation of a body by combining accelerometers, gyroscopes, and occasionally magnetometers (Wikipedia, 2025).

The development of inertial measurement units (IMUs) began in the 1930s, primarily for aircraft navigation and other large-scale applications. Early IMUs were limited by size, cost, and power consumption, making them unsuitable for smaller devices or consumer applications. Therefore, the application of IMUs in consumer products requires the sensors to be small and lightweight to fit seamlessly into the product design while maintaining good mobility. The most commonly used IMU technology in the consumer industry is the MEMS (Micro-Electro-Mechanical Systems) sensor.

Early IMUs combined two types of sensors: accelerometers, which measure inertial acceleration, and gyroscopes, which measure angular rotation. Both sensors typically operate with three degrees of freedom, measuring along three axes. Advancements in IMU technology introduced magnetometers, which measure magnetic bearing direction. Gyroscopes are prone to drift errors over extended periods, while accelerometers are sensitive to translational acceleration or rapid rotations in any direction. To address these limitations, advancements in IMU technology have introduced magnetometers, which measure magnetic bearing direction. By combining magnetometer data with gyroscope readings, overall accuracy and reliability are significantly improved.

The degree of freedom (DOF) refers to the number of independent parameters that a system can measure or control. An IMU typically consists of an accelerometer, a gyroscope, and a magnetometer—commonly configured as tri-axial sensors. This arrangement allows measurements across three axes, resulting in a total of 9 DOF.

Regarding response rate, the standard range for most current applications is between 1 Hz and 50 Hz. However, advanced applications, such as vehicle navigation, may require a higher response rate of up to 200 Hz to ensure data is captured quickly and accurately in real-time (Ahmad et al., 2013).

2.2.2.1 Accelerometer

An accelerometer is a device that measures the proper acceleration of an object (Tinder, 2007). Proper acceleration refers to the physical acceleration experienced by an object relative to a free-fall inertial frame (Rindler, 2013). It is the acceleration measured by an accelerometer attached to the object and is distinct from coordinate acceleration, which depends on the choice of a reference frame.

Accelerometers are widely used in industry, consumer products, and scientific applications. Highly sensitive accelerometers play a key role in inertial navigation systems for aircraft and missiles. In unmanned aerial vehicles, they assist in stabilizing flight. Micromachined micro-electromechanical systems (MEMS) accelerometers are integrated into handheld electronic devices, such as smartphones, cameras, and video game controllers, to detect movement and orientation. In industrial settings, accelerometers monitor vibration in machinery, while seismometers, which are specialized accelerometers, detect ground movements, including earthquakes (Wikipedia, 2025).

2.2.2.2 Gyroscope

A gyroscope is a device used to measure or maintain orientation and angular velocity. It typically consists of a spinning wheel or disc with an axis of rotation (spin axis) that can freely assume any orientation. When spinning, the orientation of the axis remains unaffected by tilting or rotation of its mounting, following the conservation of angular momentum (Wikipedia, 2025).

Many classes of gyroscopes exist, depending on the operating physical principle and the involved technology, such as microchip-packaged Micro-Electro-Mechanical System (MEMS)

2. Background

gyroscopes found in electronic devices, solid-state ring laser gyroscopes, fiber optic gyroscopes, and highly sensitive quantum gyroscopes.

MEMS gyroscopes are motion sensors that detect and measure the angular motion of an object. They measure the rate of rotation around one, two, or three axes. Initially developed for high-cost military applications, MEMS gyroscopes are now widely used in low-cost commercial applications, including consumer electronics for automotive, defense, industrial, and medical purposes. The growing demand for mobile devices has further driven the expansion of the MEMS gyroscope market. Additionally, the cost of MEMS gyroscopes is expected to decrease significantly in the coming years, leading to broader adoption of these devices (Passaro et al., 2017).

Gyroscopes are used in various applications, including inertial navigation systems, such as those in the Hubble Space Telescope or the steel hull of a submerged submarine. Due to their precision, they are also employed in gyrotheodolites for maintaining direction in tunnel mining. Gyroscopes are integral to gyrocompasses, which supplement or replace magnetic compasses in ships, aircraft, spacecraft, and other vehicles. Additionally, they assist in stability for bicycles, motorcycles, and ships or serve as components in inertial guidance systems (Wikipedia, 2024).

2.2.2.3 Magnetometer

A magnetometer is a device that measures the magnetic field, including its strength and orientation. Some magnetometers measure magnetic dipole moments, which involve closed loops of electric current or pairs of poles. Various types of magnetometers are used to measure the direction, strength, or relative variations of a magnetic field at a specific location. One common example is a compass, which detects the direction of the surrounding magnetic field, typically the Earth's magnetic field (Wikipedia, 2025).

Magnetometers are widely used for measuring the Earth's magnetic field, detecting magnetic anomalies, and determining the magnetic properties of materials. They serve as heading references in aircraft systems and are used militarily in magnetic mines to detect submarines, with sensitive models often classified as military technology. Magnetometers can also function as metal detectors, capable of detecting ferrous metals at much greater distances than conventional metal detectors. For instance, they can detect large objects like cars over 10 meters away. With advancements in miniaturization, magnetometers are now integrated into low-cost circuits, making them increasingly common as compact compasses in MEMS-based devices (Wikipedia, 2025).

2.3 Indoor/outdoor detection with different sensors

2.3.1 Indoor-outdoor Detection based on GNSS Sensors

The advancement of Global Navigation Satellite Systems (GNSS) has introduced abundant location information, enabling more precise positioning and expanding opportunities for various applications. Hsu et al. (2017) explored methods to detect and mitigate faulty GNSS signals in urban canyon areas using consistency checks. Their approach highlights the challenges posed by urban environments where signal reflection and obstruction lead to inaccuracies. Similarly, Tabatabaei & Mosavi (2019) demonstrated the benefits of integrating multiple GNSS systems, such as GPS and GLONASS, to enable information exchange across channels, improving overall positioning accuracy. Zhou et al. (2018) further developed a GNSS cooperative positioning framework, emphasizing collaboration among systems to enhance performance.

Despite these advancements, many correction methods that rely on GNSS require additional receivers or specialized equipment, limiting their applicability in everyday scenarios. For instance, such solutions may not be feasible for casual or wearable device users due to cost and complexity. Moreover, when transitioning to indoor or sheltered environments, GNSS positioning performance suffers significantly. Issues such as reduced signal intensity, altered frequencies, and limited satellite constellations contribute to decreased accuracy.

To address these challenges, incorporating additional GNSS-derived information as features for classification has gained popularity as a method for improving indoor-outdoor (IO) detection. For example, Bui et al. (2020) leveraged GPS redundancy and the presence of solid obstacles as key characteristics for machine learning classification, achieving an impressive overall accuracy of 97%. Zhu et al. (2019) took a similar approach by extracting 36 features from categories such as spatial geometry distribution, time sequence, and statistical properties, demonstrating the potential of feature-rich data to improve detection.

2.3.2 Indoor-outdoor Detection Based on Inertial Sensors

Beyond solely relying on GPS sensors for data collection, a growing trend involves utilizing multiple sensors to address indoor-outdoor (IO) detection challenges. Wearable devices, such as smartphones and smartwatches, have become essential tools in daily life and are equipped with a variety of sensors, including light sensors, inertial measurement units (IMUs), GPS, Bluetooth, and cellular connectivity. These sensors provide practical and innovative methods for tackling the IO detection problem.

Zhu et al. (2019) proposed an approach that combines data from a light sensor, magnetic sensor, and GNSS for IO detection. Their system applies a pedestrian dead reckoning algorithm using inertial sensors for indoor environments, while GNSS and magnetic sensors are employed for outdoor environments. Similarly, Rajak et al. (2021) developed an IO detection framework that leverages smartphone sensors such as ambient light, magnetic field, and GPS as features for a machine-learning model. Their grid-based framework distinguishes between outdoor, semi-indoor, indoor, and deep indoor environments. However, this approach requires substantial computational resources and a large volume of input data, making it computationally intensive.

Among the variety of sensors available, inertial sensors, such as magnetometers, gyroscopes, and accelerometers, play a critical role in real-time movement detection. These sensors are particularly valuable for continuous monitoring, especially for older adults and differently-

abled individuals navigating unfamiliar environments. Özdemir & Barshan (2014) demonstrated the potential of wearable motion inertial sensors by developing an automated fall detection system for older adults, applicable in both indoor and outdoor settings. Their design schema offers valuable insights for IO detection systems.

Another noteworthy study by Belhajem et al. (2018) explored the integration of GPS and an Inertial Navigation System (INS) with an Extended Kalman Filter (EKF) and Support Vector Machine (SVM) methods. This approach effectively addressed challenges such as signal masking and multipath issues, common in urban or enclosed environments. Additionally, Okamoto & Chen (2015) highlighted the significance of satellite signal distribution changes, demonstrating their strong correlation with users' movement directions. This finding enables accurate location tracking even in partially sheltered areas.

Ashraf et al. (2018) investigated the use of ambient magnetic field data from smartphones for IO detection and achieved an accuracy of 85.30%. Their study highlights the potential of using single-sensor data for this purpose. While GPS features have shown high accuracy in IO detection, and research on inertial sensors is extensive, there remains a significant gap in the integration of GPS with accelerometers and magnetometers for comprehensive and robust IO detection.

2.4 Indoor-outdoor Detection with Machine Learning Methods

Traditional approaches for indoor-outdoor (IO) detection rely heavily on GPS data, which performs well outdoors but encounters significant limitations indoors due to signal obstructions, reflections, and multipath effects. To overcome these challenges, machine learning methods have emerged as a powerful solution, capable of integrating data from multiple sensors and learning complex patterns that traditional models struggle to capture.

2.4.1 Machine Learning

Machine Learning (ML) is a branch of artificial intelligence (AI) that focuses on building systems capable of learning and improving from experience without being explicitly instructed. Instead of following predefined rules, ML models identify patterns and relationships within

data to make predictions or decisions. The process involves training on datasets, validating performance, and deploying models for real-world applications.

ML systems rely on structured and unstructured data, algorithms such as support vector machines and neural networks, and critical features derived from data. Learning paradigms include supervised learning (using labeled data), unsupervised learning (identifying patterns in unlabeled data), and reinforcement learning (learning through interaction and feedback).

In indoor-outdoor (IO) detection, ML effectively integrates multi-sensor data, such as GPS and IMU readings, to enable accurate environmental classification, overcoming the limitations of traditional methods.

2.4.2 Role and Approaches of Machine Learning in IO Detection

Machine learning offers a flexible framework for processing diverse sensor data and making accurate IO classifications. By leveraging algorithms capable of identifying patterns in GPS signals, inertial measurements, and environmental data, machine learning can provide robust detection even in challenging environments, such as urban canyons or fully enclosed spaces.

Key advantages of machine learning in IO detection include multi-sensor integration, feature extraction, and scalability. For example, it can combine data from GPS, inertial sensors, Wi-Fi, Bluetooth, and environmental sensors for holistic decision-making and utilizes advanced techniques to identify the most relevant features, such as signal strength, motion patterns, and magnetic anomalies. Besides, it adapts to different hardware platforms, from wearable devices to smartphones, with minimal calibration requirements.

Supervised models are trained on labeled datasets that include features like signal intensity, motion patterns, and ambient sensor readings alongside their corresponding indoor or outdoor labels. Algorithms such as Support Vector Machines (SVM), Random Forest, Gradient Boosting Machines, and Neural Networks are commonly used for IO detection.

When labeled data is unavailable, clustering methods such as K-Means and DBSCAN can group similar data points, identifying transitions between indoor and outdoor clusters. These methods can also highlight semi-indoor environments or detect anomalies.

Convolutional Neural Networks (CNNs) and Recurrent Neural Networks (RNNs) are increasingly being used for feature-rich IO detection tasks. RNNs, in particular, excel in sequential data analysis, making them suitable for time-series sensor data.

2.5 Research Gaps

Although GPS technology is highly effective for tracking outdoor activities, its performance significantly declines in indoor environments due to signal obstructions caused by buildings and other structures. This limitation results in erratic and unreliable location data, commonly referred to as “GPS drift”, making it difficult to accurately distinguish between indoor and outdoor activities or to track transitions between these environments. For the MOASIS project, which primarily relies on GPS data to monitor daily mobility, this limitation creates a significant challenge in capturing a complete picture of mobility patterns, especially for older adults who frequently transition between indoor, semi-indoor, and outdoor spaces.

Additionally, the current focus of the MOASIS project is largely on outdoor activity monitoring, leaving limited understanding of mobility behaviors in indoor and semi-indoor environments. To achieve a more comprehensive analysis of mobility, it is crucial to incorporate additional data sources that can complement GPS and improve the accuracy of activity detection across various settings.

Recent advancements in inertial measurement unit (IMU) technology, including accelerometers, gyroscopes, and magnetometers, have demonstrated strong potential for improving indoor-outdoor detection. However, the MOASIS project has yet to fully explore the integration of IMU data for this purpose. By utilizing data available through the uTrail platform and applying machine learning classification methods, there is an opportunity to enhance the detection of mobility patterns and improve the reliability of indoor-outdoor transition tracking.

Therefore, the primary research gaps include:

- The limited accuracy of GPS for indoor positioning and its inability to reliably track transitions between indoor, semi-indoor, and outdoor spaces due to GPS drift.
- The insufficient focus on mobility patterns in indoor and semi-indoor environments within the MOASIS project led to an incomplete understanding of older adults' spatial behaviors.
- The underutilization of IMU data and the lack of integration with GPS data through machine learning methods improve classification and detection accuracy.

This study seeks to address these gaps by combining GPS and IMU data with machine learning techniques, enabling more accurate detection of mobility patterns across diverse environments and contributing to a more comprehensive understanding of daily mobility in older populations.

2.6 Research Objective and Research Questions

The goal of this master's thesis is to develop an effective method for indoor-outdoor detection using GPS and IMU data obtained from the MOASIS project. In the MOASIS project, GPS logs were primarily utilized to monitor daily spatial mobility, while accelerometers and magnetometers were employed to detect physical activity types. Although GPS performs reliably as a standalone system in outdoor environments, enabling MOASIS to conduct extensive research on older adults' outdoor activities. Its effectiveness diminishes indoors due to signal interference caused by construction materials. This limitation leads to chaotic GPS patterns, complicating the distinction between indoor and mobile trajectories.

The inability to accurately track older adults' transitions between indoor, semi-indoor, and outdoor environments presents a significant obstacle to understanding their daily activities comprehensively. For the purpose of this study:

- Indoor activities are defined as those conducted within fully enclosed spaces, such as buildings or rooms.
- Outdoor activities occur in open-air environments, unimpeded by physical enclosures.
- Semi-indoor activities take place in partially enclosed areas, such as covered patios or greenhouses.

These classifications are critical for understanding the nuances of movement and activity patterns, especially as older adults may frequently transition between these spaces throughout their day.

Given the excellent performance of IMU data in indoor-outdoor detection, I hypothesize that integrating GPS and IMU data can significantly enhance the accuracy of activity tracking within the MOASIS framework. By combining the strengths of these systems, the detection of transitions between indoor, semi-indoor, and outdoor environments could be improved, providing a more robust analysis of older adults' spatial mobility and activity patterns.

However, further research is required to validate this hypothesis and refine the methodologies for achieving reliable indoor-outdoor detection.

To guide this research, the following research questions will serve as focal points throughout this master's project:

- How can GPS and IMU data be optimally combined to improve the accuracy of indoor-outdoor detection?
- What specific features of GPS and IMU data are most effective in distinguishing between indoor, outdoor, and semi-indoor environments?
- Do individual, scenario and 'semi-indoor' definition affect model performance?

By addressing these questions, this study aims to contribute to the broader research on spatial mobility and provide valuable insights into improving tracking technologies for older populations.

3. Methodology

3.1 Concept

3.1.1 Indoor, Outdoor, and Semi-indoor

In general, indoor environments are defined as fully enclosed spaces, such as buildings or rooms, where conditions are controlled, and exposure to external environmental factors is minimized. These spaces are designed to regulate variables such as temperature, lighting, and ventilation, creating a stable environment independent of outdoor influences. Conversely, outdoor environments are open-air settings, unimpeded by physical enclosures, where activities are directly subject to natural elements such as weather, sunlight, and airflow. These distinctions provide a foundational framework in research for categorizing spatial contexts and analyzing their impact on human activity and behavior.

In this thesis, the concept of semi-indoor environments is introduced to address a transitional category of spaces that are partially enclosed, such as covered patios, greenhouses, and transit vehicles. This category bridges the gap between fully enclosed indoor and fully open outdoor environments. The inclusion of semi-indoor spaces aims to explore whether these areas exhibit unique characteristics, distinct from those observed in traditional indoor or outdoor settings. Specifically, the study investigates how partial shelter and exposure influence the interaction between individuals and their surroundings, with implications for understanding mixed-use or hybrid environments.

Indoor activities refer to those conducted in fully enclosed spaces, such as buildings or rooms, where physical boundaries like walls and ceilings create a controlled environment insulated from external elements. This definition highlights the purpose of such spaces in ensuring stability and predictability in activity settings.

Outdoor activities occur in open-air environments, characterized by the absence of physical enclosures. These settings enable direct interaction with natural conditions, emphasizing the influence of environmental variability on human experiences and activities.

Semi-indoor activities take place in partially enclosed spaces that blend features of indoor and outdoor environments. Examples include areas with partial roofs or walls, such as greenhouses and covered transit stops. These spaces provide some degree of protection while still allowing exposure to external conditions, offering a unique interplay between control and openness.

The definitions established in this thesis are tailored to facilitate a clear and accurate distinction between these spatial contexts, aligning with the study's objective to analyze how environmental characteristics influence human activity and interaction.

3.1.2 Scenarios

Selecting scenes for data collection based on indoor, outdoor, and semi-indoor definitions ensures comprehensive representation of different spatial environments and their unique characteristics. The choice of scenes reflects typical human activities in various environments, balancing realism with the need for controlled data collection. Below is a summary of the selection process and representative scenes for each category.

Indoor scenes are chosen to capture activities conducted in fully enclosed spaces, where environmental factors such as temperature, lighting, and noise are regulated. Typical examples include:

- Teaching buildings: Classrooms, libraries, and offices, where individuals engage in study or work.
- Shopping malls: Retail stores and supermarkets, representing everyday activities in controlled public spaces.
- Transit vehicles: Trains and trams, enclosed environments with regulated conditions during transit.

Outdoor scenes are selected to reflect activities in open-air environments, unimpeded by physical enclosures, and exposed to natural elements such as sunlight and weather. Examples include:

- Parks and courtyards: Spaces where people walk or relax amidst natural surroundings.

- City streets: Urban areas such as Bahnhofstrasse, where outdoor walking and public interaction occur.
- Open squares: Central gathering points, such as shopping mall, used for transit or public events.

Semi-indoor scenes are identified to investigate partially enclosed spaces that combine indoor and outdoor characteristics. These spaces provide partial shelter while allowing some exposure to external conditions. Examples include:

- Covered patios or eaves: Transitional spaces offering protection but open to the surroundings.
- Transit stations or stops: Areas under glass roofs or platforms with partial enclosure.
- Atrium or community centers: Locations with significant openness yet offering shelter.

The actual scenes for data collection were carefully designed based in Zurich, Switzerland, considering the diversity of indoor, outdoor, and semi-indoor environments. A total of seven predefined scenarios were developed, ensuring a balanced representation of spatial contexts to align with the study's objectives. Each scenario provides a practical framework for examining environmental characteristics and their impact on activities.

The seven scenarios include diverse settings such as university campuses, public transportation hubs, shopping districts, and residential areas. Each scenario integrates indoor, outdoor, and semi-indoor environments, offering a comprehensive view of different spatial contexts. Each scenario consists of 5 to 7 specific tasks or locations, ensuring participants experience a variety of environments. Data collection for each scenario was designed to be completed in under 1.5 hours, ensuring feasibility for participants, detailed scenario see Appendix A.

1. Irchel Campus Scenario: Combines university buildings, green spaces, and semi-enclosed courtyards.
2. Train Station Scenario: Captures transitions from outdoor public areas to enclosed shopping malls and transit spaces.
3. Main Campus Scenario: Focuses on academic, urban, and transit-related environments.
4. Garden Scenario: Explores residential and recreational settings with transitional spaces like eaves and patios.
5. Irchel Campus Scenario 2: Adds emphasis on informal gathering spaces like cafeterias and bars under semi-enclosures.
6. Central Scenario: Highlights transit hubs and blended environments such as glass-roofed

stations and tram rides.

7. City Center Shopping Scenario: Represents urban retail and leisure spaces, integrating streets, shops, and public transport.

In addition to the predefined scenarios, participants have the flexibility to create their own custom scenarios. This customization enables the study to adapt to individual or contextual differences while adhering to specific requirements:

- Inclusion of indoor, outdoor, and semi-indoor environments: Each custom scenario must include indoor, outdoor, and semi-indoor settings to ensure consistency with the study's framework.
- Minimum stay duration in each scene: Participants are required to remain in each location for at least 5 minutes before transitioning to the next space. This ensures adequate data collection and observation within each environment.

The predefined scenarios serve as practical templates, while the flexibility to design custom scenarios ensures the study remains adaptable to varying conditions and user preferences. This approach ensures robust and comprehensive data collection.

3.2 Data Collection

3.2.1 Data Collection Workflow

The data collection process involved 12 participants. All affiliated with the University of Zurich's GIScience group. Participants were responsible for collecting data using two tools: the "uTrail" device and the "Label-logger" application.

Prior to participating, all individuals signed an informed consent form, ensuring they fully understood the purpose, scope, and methods of data collection. The data collection and usage procedures strictly adhered to ethical guidelines and data protection regulations to safeguard participant privacy and confidentiality.

Each participant attended a brief training session and designed a tutorial to understand the use of the "uTrail" device and the "Label-logger" application (tutorial see Appendix B). This session provided instructions on wearing the device, using the app to label environments, and following the data collection protocol.

After the training, participants were provided with a “uTrail” device, which was pre-configured and set to default ON mode. They were instructed to download and install the “Label-logger” app on their smartphones to facilitate environment labeling during data collection.

Participants were required to complete 3 to 5 scenarios during the data collection process. They could select from 7 predefined scenarios or create a custom scenario based on their preferences. Both scenario selection and custom scenario creation were conducted using the “Label-logger” app, ensuring flexibility and adherence to study protocols. Each scenario was selected no more than 4 times across all participants, ensuring a balanced sample size for each scenario.

Before starting data collection, participants were instructed to wait in an open space to allow the “uTrail” device to acquire a reliable GPS signal and ensure the UTC time synchronization. This step was critical for maintaining accuracy in spatial and temporal data alignment.

Participants wore the “uTrail” device at a designated fixed position on their body (e.g., waist or shoulder) to record movement and environmental data continuously.

Using the “Label-logger” app, participants manually labeled their environment as indoor, outdoor, or semi-indoor based on their location and activity.

Participants followed the tasks outlined in their chosen or custom scenarios, remaining in each location for at least 5 minutes to ensure sufficient data collection. The predefined scenarios guided participants through a range of spatial environments, while custom scenarios allowed for personalized exploration within the study’s framework.

The data collected by the “uTrail” device and the “Label-logger” application were synchronized using timestamps. This ensured that environmental labels recorded via the app could be accurately matched with location and movement data from the “uTrail” device, enabling comprehensive analysis of spatial context and activity patterns. The workflow is shown as Figure 2.

Data Collection Workflow

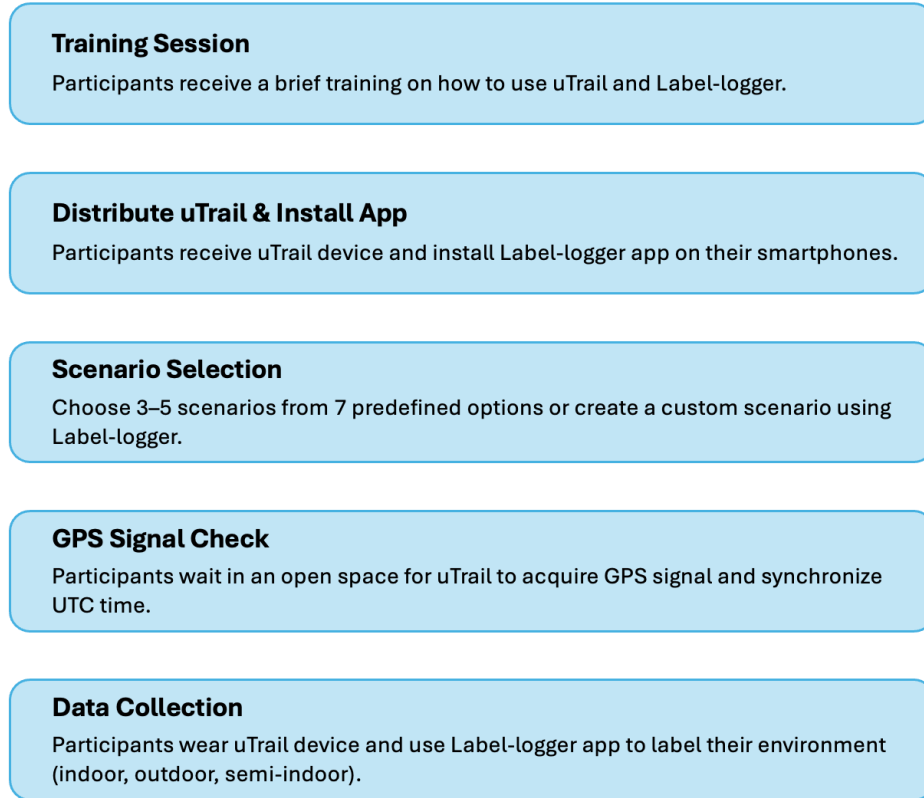


Figure 2 Data Collection Workflow

3.2.2 “uTrail” Device

“uTrail” device is a compact, wearable device specifically designed for data collection, equipped with an audio sensor, a GPS sensor (uBlox UC530M) configured with a 1 Hz sampling rate, and an accelerometer (ST Microelectronics LSM303D) operating at a 50 Hz sampling rate. The accelerometer captures data across three magnetic field channels and three acceleration channels (Allahbakhshi et al., 2020).

The recommended placement of the “uTrail” device typically aligns with the most common accelerometer positioning, which is on the waist or hip. This placement, being close to the body trunk, is widely considered optimal for accurately capturing human movement (Liao et al., 2015). However, as this study does not require precise movement orientation data but instead prioritizes reliable GPS signal reception, participants were advised to carry the “uTrail” device in an external pocket of their clothing, such as a pants pocket, or in an outer pocket of a personal bag. This placement ensures both convenience and the consistent acquisition of GPS signals during data collection.

The “uTrail” device is a light, buttonless wearable. Its operation, including powering on and off, is entirely managed through the accompanying “uTrailConfigure” software. This software, compatible with Windows operating systems, serves as the central hub for device management, offering a wide range of functionalities. Users can use it to activate or deactivate the device, download collected data, process it for analysis, and perform cleanup operations to prepare the device for future use. One of the key features of the “uTrailConfigure” software is its ability to customize the device’s sampling settings to suit specific research requirements. The IMU frequency can be adjusted to 50 Hz, 100 Hz, or 200 Hz, depending on the level of detail needed for motion tracking. For GPS functionality, the activity interval can be configured as low as 1 second, ensuring high-resolution spatial data collection. This flexibility makes the “uTrail” device adaptable for a variety of applications, from precise movement analysis to broader environmental tracking.

The “uTrail” device is charged using a Type-C charging cable, ensuring compatibility with modern charging standards. A full charge allows the device to operate continuously for approximately one day, making it suitable for extended data collection sessions. Additionally, every time the Type-C charging cable is connected or disconnected, the device automatically creates a new data file. This feature simplifies the differentiation of data collected during different periods, enhancing the organization and traceability of the collected datasets. The automatic segmentation of data also brings added convenience to the subsequent data processing and analysis stages.

3.2.3 “Label-logger”

The data for this research was collected using the “uTrail” device, with support from the GIScience group, alongside the “Label-logger” application, which was specifically designed for this study to facilitate environment labeling.

3. Methodology

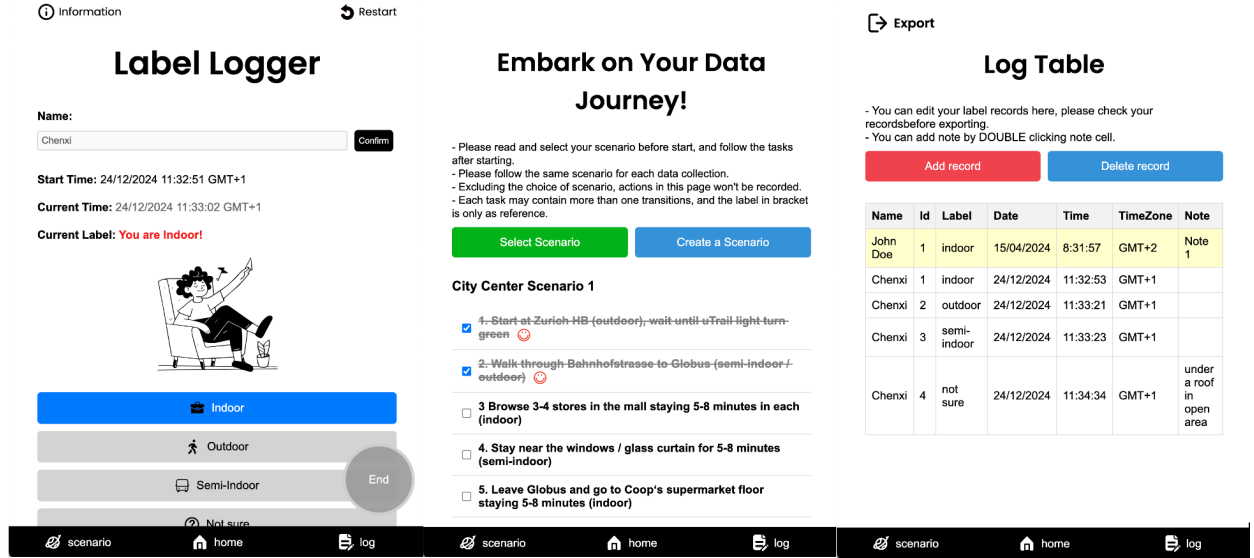


Figure 3 “Label Logger” Overview

As shown in Figure 3, the “Label-logger” application serves as a labeling tool for the “uTrail” device, enabling participants to annotate their environment during data collection. The primary goal of the application is to associate environmental labels with the data recorded by the “uTrail” device. This research categorizes the environment into three types: indoor, outdoor, and semi-indoor. Participants were required to specify their environment type at the start of data collection and update the label whenever they transitioned between environments.

Label-logger is available as both a web-based application and an Android version. For the purpose of this thesis, the web-based application was predominantly used due to its convenience and ease of access. However, the Android application, developed later using the median.co online platform, offers additional functionality, such as offline usage through caching, making it more versatile for fieldwork.

After starting the Application, upon launching the application, participants entered and confirmed their name on the “Home” page. Participants are navigated to the “Scenario” page, where they could view or create their scenarios. The scenario provided a general guide on where to go and what to do but allowed flexibility in execution.

Once ready, participants pressed the “Start” button on the “Home” page to begin data collection. Participants wore the “uTrail” device, followed the scenario tasks, and updated the environment label on the application whenever their environment changed.

At the end of each session, participants accessed the “Log” page and exported the log as an Excel file. This file contained the labeled data required for further analysis.

The “Label-logger” application was designed with machine learning objectives in mind, as the study requires labeled training and validation datasets for subsequent model development. The web application was primarily developed using JavaScript for ease of deployment and compatibility across devices. The Android version was configured on the median.co platform, which allowed rapid adaptation to mobile usage while supporting offline functionality through caching. By combining these tools, the application ensured seamless and efficient data labeling, providing a robust foundation for the study’s data collection and machine learning pipeline.

3.2.3.1 Home Page

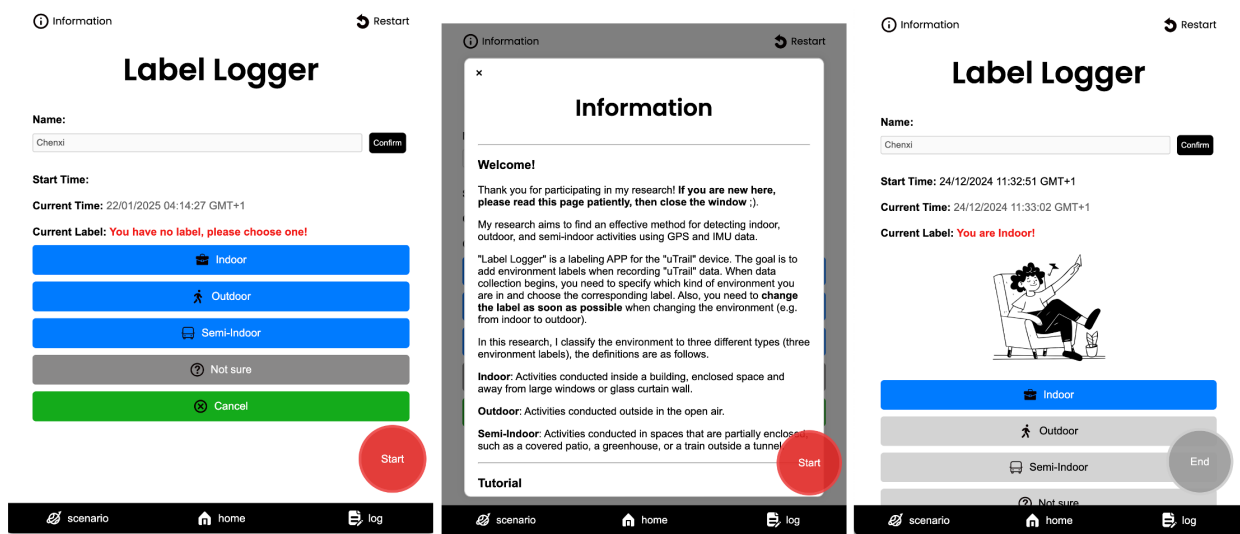


Figure 4 “Label Logger” Home Page

As shown in Figure 4, the Home Page of the Label Logger application serves as the primary interface for participants to begin and manage their data collection process. Its features are designed for ease of use and flexibility in labeling environments during the study.

The “Information” button on the Home Page opens an information page containing a brief introduction to the research and a concise tutorial. This page provides a quick reference guide for participants, ensuring clarity in using the application.

Participants are required to input their name (or a code name, as per the consent form) on the Home Page. This name ensures data traceability while maintaining anonymity. Participants are advised not to change their name during a single data collection session to ensure consistency in the records.

The most important feature of the Home Page is the label selection function, which allows participants to record their environment type. Participants can switch labels as they change environments by simply selecting the new label.

To stop recording the current label, participants can click on the label again to end it. If a label is selected by mistake, the “Cancel” button can be used to remove the erroneous record. However, using the cancel function will erase the associated data, so participants are encouraged to promptly switch to the correct label instead.

The “Start” button initiates the data collection process. The “End” button terminates the current data collection session. The “Restart” button, located in the upper-right corner, clears all recorded data, allowing participants to begin a new session.

3.2.3.2 Scenario Page

Embark on Your Data Journey!

- Please read and select your scenario before start, and follow the tasks after starting.
 - Please follow the same scenario for each data collection.
 - Excluding the choice of scenario, actions in this page won't be recorded.
 - Each task may contain more than one transitions, and the label in bracket is only as reference.

Select Scenario
Create a Scenario

City Center Scenario 1

- ☒ 1- Start at Zurich HB (outdoor); wait until uTrail light turn green 🚶
- ☒ 2- Walk through Bahnhofstrasse to Globus (semi-indoor/-outdoor) 🚶
- ☐ 3 Browse 3-4 stores in the mall staying 5-8 minutes in each (indoor)
- ☐ 4. Stay near the windows / glass curtain for 5-8 minutes (semi-indoor)
- ☐ 5. Leave Globus and go to Coop's supermarket floor staying 5-8 minutes (indoor)

Embark on Your Data Journey!

- Please read and select your scenario before start, and follow the tasks after starting.
 - Please follow the same scenario for each data collection.
 - Excluding the choice of scenario, actions in this page won't be recorded.
 - Each task may contain more than one transitions, and the label in bracket is only as reference.

Select Scenario
Cancel creating

Add task
Save Scenario

scenario
home
log

scenario
home
log

Figure 5 “Label Logger” Scenario Page

As shown in Figure 5, the Scenario Page is designed to guide participants through their data collection process by providing pre-defined or customizable scenario plans.

Participants can choose from 7 predefined scenarios or create their own scenario. Each scenario outlines specific tasks or locations to visit during data collection. By selecting a scenario, participants gain a general idea of where to go and what to do. However, participants are allowed flexibility and are not required to follow the scenario strictly.

Participants can create their own scenarios using the “Create a Scenario” function. When designing a scenario, the following protocol must be followed:

- The scenario should include indoor, outdoor, and semi-indoor environments.
- Participants should stay in each location for at least 5 minutes before moving to the next.

The scenario must have a title and clearly defined tasks. Participants can quit the creation process at any time by clicking the “Cancel Creating” button.

The Scenario Page functions as a to-do list, where participants can mark tasks as completed. This ensures participants remain organized during the data collection process.

3.2.3.3 Log Page

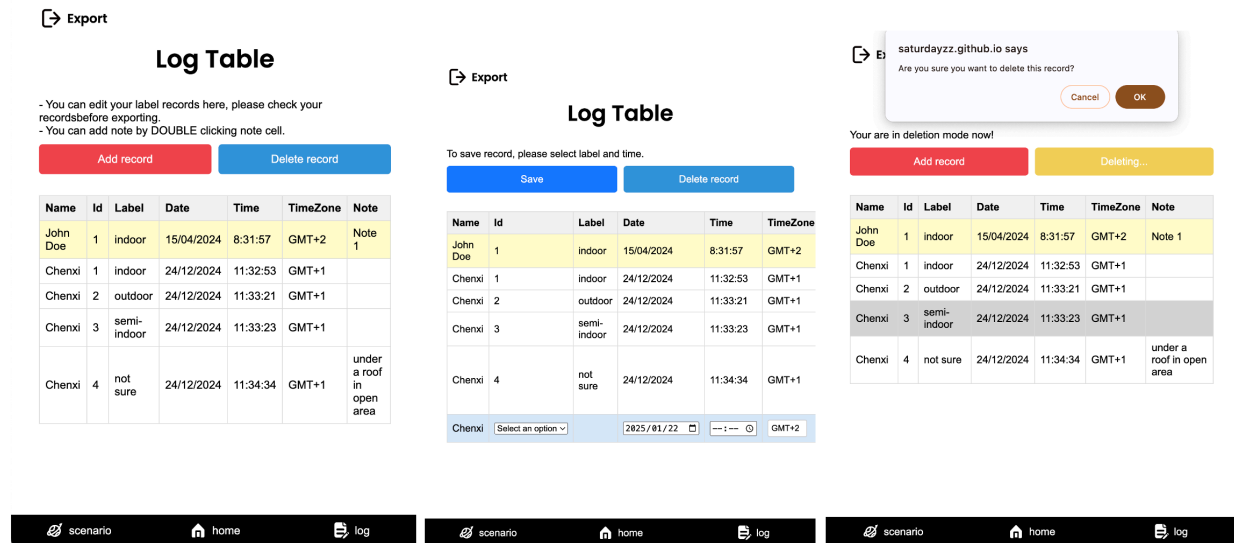


Figure 6 “Label Logger” Log Page

As shown in Figure 6, the Log Page is a comprehensive interface for managing recorded labels. It allows participants to review, edit, and export their label records for analysis.

The log table displays all recorded labels, color-coded for clarity:

- Yellow: The first record, serving as an example.
- White: Records created during data collection on the Home Page.
- Blue: Records manually added using the “Add record” button.

Participants can add new records by clicking the “Add record” button, filling in required fields (Environment ID, Date, Time, and a description if using the “Not Sure” label), and saving the entry.

Incorrect records can be deleted using the “Delete record” button. Deleted records will appear grey before final confirmation to avoid accidental deletions.

At the end of each data collection session, participants must export their label log. Clicking the “Export” button in the upper-left corner generates a CSV file containing the recorded data. The exported file can be downloaded and shared via email for further analysis. These features ensure that participants can effectively manage and validate their recorded labels, providing high-quality data for the study.

3.3 Data Preprocessing

After collecting data from the “uTrail” device, the data was exported using the “uTrailConfigure” software. During the export process, the “uTrailConfigure” software converts the raw data stream format (stored as .dat files) into more accessible and structured file formats, including CSV, KML, and TXT files. These exported files contain essential information, including GPS and IMU data, as well as the configuration settings of the “uTrail” device.

This conversion step is crucial as it transforms the raw and unstructured data into formats compatible with standard data analysis tools, facilitating further processing and interpretation of the collected information. By preserving both the measurement data (GPS and IMU) and the device configuration details, the exported files provide a comprehensive dataset for subsequent analysis. In addition to the exported data files, corresponding label files are collected separately by the user via the “Label-logger” application. The label files are downloaded in CSV format through the application and are typically sent via email for integration with the exported data.

Once all the files have been successfully downloaded, the first step involves verifying the format and correctness of the label files. This process begins by cross-referencing the label file against notes recorded by participants to ensure that the labels (e.g., indoor, outdoor, or semi-indoor) have been appropriately assigned. Given that the dataset consists of only 50 files, this step is carried out manually to ensure accuracy. Each record in the label file is carefully inspected, and any incorrectly assigned labels are manually corrected to reflect the actual scenario. This manual adjustment ensures that the dataset maintains a high level of reliability and consistency.

After verifying and correcting the label files, the next step is to match the label files with the corresponding “uTrail” data files. The matching process is primarily based on the participant’s name and the data collection timestamps. Any scenario labels that were modified during the previous step in the label files are also updated in the corresponding “uTrail” files to ensure consistency across the dataset.

This matching process is automated using Python scripts to handle the data integration efficiently. However, the procedure comes with two main challenges. The first challenge lies in ensuring uniformity in the timestamp formats across both label and “uTrail” files, as discrepancies in time formats can disrupt the matching process. The second challenge is the

computational overhead involved in reading and processing large-scale data, which can be time-consuming.

One significant challenge in this process is the large volume of data generated by the “uTrail” device, particularly the IMU data, which is recorded at a frequency of 50 Hz (i.e., 50 data points per second). The high frequency of IMU data results in large file sizes, making single-threaded data processing exceedingly time-consuming.

To address this, a multi-threaded strategy is implemented in the Python code to improve efficiency. By leveraging parallel processing, the data is divided into smaller chunks that are processed simultaneously across multiple threads or cores. This approach significantly reduces the time required to read and process the large datasets, especially when handling high-frequency IMU data.

3.4 Implementation of Model Training

3.4.1 Model Training

Random Forest is a widely used ensemble learning algorithm for classification and regression tasks. It operates by constructing multiple decision trees during training and combining their outputs, either by majority voting (for classification) or averaging (for regression), to enhance accuracy and robustness. By using random subsets of the data and selecting features randomly at each split, Random Forest reduces overfitting and ensures diversity among trees, making it effective for handling high-dimensional and complex datasets. Its non-parametric nature allows it to model nonlinear relationships, and its ability to calculate feature importance provides interpretability, making it suitable for classifying environments such as indoor, outdoor, and semi-indoor in this study.

For this study, the `RandomForestClassifier` from the `sklearn.ensemble` module is used in Python, a widely used machine learning library that provides efficient implementations of various algorithms, including Random Forest. The `RandomForestClassifier` allows for flexible configuration of hyperparameters such as the number of trees (`n_estimators`), maximum tree depth (`max_depth`), and feature selection strategies to optimize performance. Initially, a small-scale sample dataset was used for training, and hyperparameters were fine-tuned based on training accuracy. Grid search and cross-validation techniques were applied to determine the

optimal values for key parameters, including `n_estimators`, which defines the number of decision trees in the forest; `max_depth`, which controls the maximum depth of each tree to prevent overfitting; `min_samples_split`, which specifies the minimum number of samples required to split an internal node; and `min_samples_leaf`, which sets the minimum number of samples required in a leaf node. These hyperparameters were iteratively adjusted to balance model accuracy and generalization.

The grid search process assesses model performance using multiple evaluation metrics, including F1-score, recall, precision, and accuracy. However, the final model selection is based on the F1-score, which is optimized in the grid search process. The F1-score is chosen as the primary evaluation metric because it represents a harmonic mean of precision and recall, balancing both false positives and false negatives. This is particularly relevant when dealing with imbalanced data or when misclassification of one class is more consequential than another.

3.4.1.1 Time Series Split

In contrast to traditional data splitting methods such as K-Fold cross-validation, which randomly partitions data into training and testing sets, time series data inherently possess temporal dependencies that must be preserved. Random shuffling of such data can lead to data leakage, where future information inadvertently influences the model during training.

To address this, the Time Series Split method is employed, which adheres to the following principles:

1. **Preservation of Temporal Order:** The data is partitioned such that earlier observations are used for training, while later observations are reserved for testing, maintaining the natural chronological sequence.
2. **Progressive Expansion of the Training Set:** With each split, the training set is incrementally expanded to include more historical data, while the testing set always consists of data points that chronologically follow the training set.

In this thesis, the dataset is divided into five sequential training and testing combinations to ensure robust model evaluation while strictly preventing any potential data leakage from future observations.

3.4.1.2 Cross Validation

Cross-validation is a technique used to evaluate the performance of machine learning models by partitioning the dataset into multiple subsets. The model is repeatedly trained and tested on different data splits, reducing its dependency on a specific partition and providing a more reliable assessment of its generalization ability.

Among various cross-validation methods, k-fold cross-validation is commonly used. It randomly divides the dataset into k equally sized subsets, where each subset serves as the test set once, while the remaining k-1 subsets are used for training. However, this approach is not suitable for time series data due to its random shuffling, which can lead to data leakage by allowing future information to influence model training.

In contrast, time series data exhibits strong temporal dependencies. In such cases, it is critical to ensure that future data does not affect model training. Therefore, traditional k-fold cross-validation is inappropriate for time series tasks.

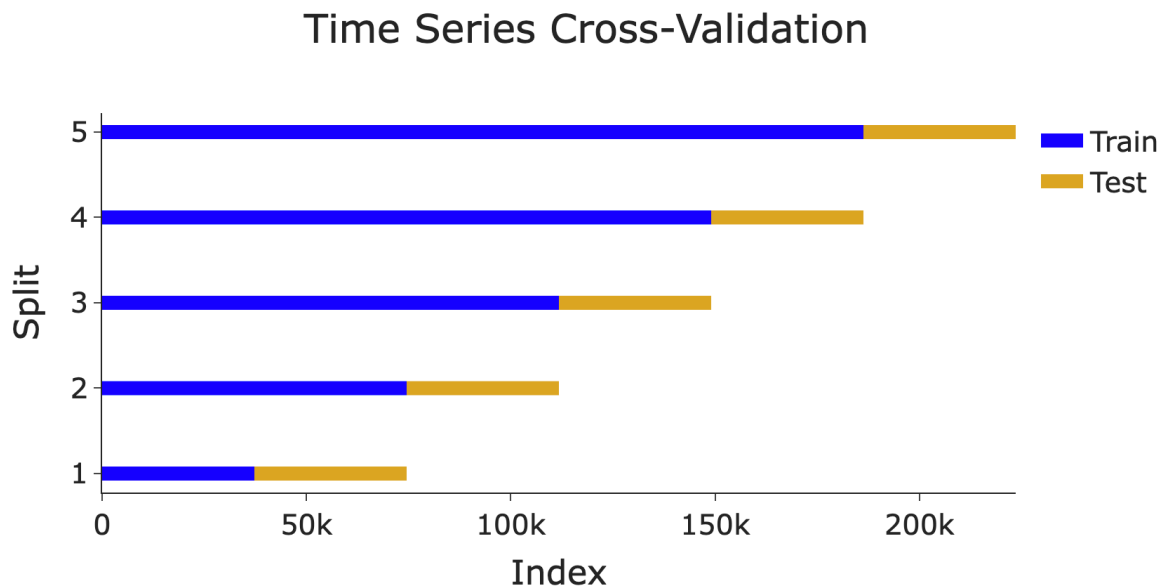


Figure 7 Time series cross-validation illustration

To address this, time series cross-validation maintains the chronological order of the data. As shown in Figure 7, the key idea is to adopt either a rolling (sliding window) or expanding window strategy. In each split, the training set progressively expands, incorporating more historical data, while the test set always consists of data points that immediately follow the training set. This approach simulates real-world forecasting scenarios, where models are trained on past observations to predict future outcomes as shown in Figure 7.

3.4.1.3 Hyperparameter

Random Forest includes several hyperparameters influencing model performance, generalization, and computational efficiency. Among these, the most critical parameters are the number of trees, the maximum depth of each tree, and the number of features considered at each split. In this study, hyperparameter selection is based on the three most influential parameters.

The number of trees in the ensemble impacts model stability and predictive performance by reducing variance. Increasing this value generally enhances accuracy but leads to diminishing returns beyond a certain threshold while significantly increasing computational time. Given this tradeoff, 50, 100, 200, and 300 trees are selected to examine the balance between performance and efficiency. The maximum tree depth controls the complexity of individual trees, where deeper trees capture intricate patterns but may overfit, reducing generalization. To assess this impact, none, 5, 10, 15, 20, and 30 are tested, with unrestricted depth allowing full tree growth and explicit constraints facilitating regularization. The number of features per split determines the subset of features considered at each decision point. Introducing randomness in this selection prevents dominant features from being repeatedly chosen, improving model diversity and reducing overfitting. The commonly used sqrt and log2 values are tested, as they effectively balance accuracy and variance reduction in classification tasks. Given the dataset size of approximately 200,000 records, these hyperparameter choices ensure a thorough evaluation of model performance while maintaining computational feasibility, with smaller training subsets enabling faster experimentation and more efficient hyperparameter tuning.

To optimize these hyperparameters, grid search is employed, a systematic method for hyperparameter tuning that exhaustively evaluates all possible parameter combinations within a predefined search space. This approach enables an objective comparison of model performance under different configurations, ensuring that the selected hyperparameters achieve the best balance between accuracy and generalization.

The selection of hyperparameters for grid search is conducted under two conditions. In the first case, when training with mixed data, a subset of the dataset is used, specifically data collected from one scenario per participant, covering all 12 participants. This subset ensures a representative sample for tuning hyperparameters while maintaining computational feasibility. In the second case, when training with data from a single scenario, the dataset is smaller;

therefore, grid search is performed on samples from 12 different scenarios. This approach allows for the selection of hyperparameters that are broadly applicable across most samples.

3.4.2 Feature Engineering

Feature engineering in this study focused on transforming raw data from the “uTrail” device and “Label-logger” application into meaningful features for training the model. Temporal features such as time of day and duration, spatial features derived from GPS data like speed and altitude changes, and sensor features such as acceleration variance and signal energy were generated. Additionally, environmental transitions between indoor, outdoor, and semi-indoor were included to provide contextual information. Data cleaning ensured consistency by removing duplicate or erroneous records, and feature selection based on Random Forest’s importance ranking helped refine the dataset, ensuring the model’s efficiency and accuracy.

Feature engineering played a critical role in transforming raw data from the “uTrail” device and “Label-logger” application into meaningful features for model training. Given the time-dependent nature of the data, temporal relationships were incorporated using a time windowing approach. This method ensures that each record’s feature matrix includes contextual information from surrounding time steps. The optimal time window size was determined empirically based on training results. In the training process, different window sizes were referred to as lag ranges, and the optimal lag range was determined empirically based on training by mix data results.

The selected features included as Table 1:

3. Methodology

Feature	Description	Unit
GPS Features		
Latitude (latitude)	Geographic coordinate specifying the north-south position.	Degrees
Longitude (longitude)	Geographic coordinate specifying the east-west position.	Degrees
Altitude (m) (altitude)	Elevation above sea level in meters.	Meters (m)
Horizontal Accuracy (m) (horAcc)	Estimated error in horizontal positioning.	Meters (m)
Vertical Accuracy (m) (vertAcc)	Estimated error in vertical positioning.	Meters (m)
Ground Speed (m/s) (groundSpeed)	Speed of movement relative to the ground.	Meters/second (m/s)
Speed Accuracy (m/s) (speedAcc)	Estimated accuracy of the speed measurement.	Meters/second (m/s)
Number of Satellites (nrSat)	The number of GPS satellites used in positioning.	Count
IMU Features		
Acceleration X (mg) (accX)	Linear acceleration along the X-axis in milligravity.	Milligravity (mg)
Acceleration Y (mg) (accY)	Linear acceleration along the Y-axis in milligravity.	Milligravity (mg)
Acceleration Z (mg) (accZ)	Linear acceleration along the Z-axis in milligravity.	Milligravity (mg)
Magnetometer X (mGauss) (magX)	Magnetic field strength along the X-axis in milligauss.	Milligauss (mGauss)
Magnetometer Y (mGauss) (magY)	Magnetic field strength along the Y-axis in milligauss.	Milligauss (mGauss)
Magnetometer Z (mGauss) (magZ)	Magnetic field strength along the Z-axis in milligauss.	Milligauss (mGauss)
Gyroscope X (mdps) (gyroX)	Angular velocity around the X-axis in millidegrees per second.	Millidegrees/second (mdps)
Gyroscope Y (mdps) (gyroY)	Angular velocity around the Y-axis in millidegrees per second.	Millidegrees/second (mdps)
Gyroscope Z (mdps) (gyroZ)	Angular velocity around the Z-axis in millidegrees per second.	Millidegrees/second (mdps)
Environmental Features		
Pressure (Pa) (pressure)	Atmospheric pressure in Pascals.	Pascals (Pa)
Temperature (°C) (temperature)	Ambient temperature in degrees Celsius.	Degrees Celsius (°C)
Humidity (%) (humidity)	Percentage of relative humidity.	Percentage (%)

Table 1 Model features

In the feature engineering process, additional statistical features were derived from the IMU data, which consists of measurements from three types of 3-axis sensors—accelerometer (accX(mg), accY(mg), accZ(mg)), magnetometer (magX(mgauss), magY(mgauss), magZ(mgauss)), and gyroscope (gyroX(mdps), gyroY(mdps), gyroZ(mdps)). For each of these features, a set of aggregation functions was applied, including mean, standard deviation, minimum, maximum, variance, mean of absolute values, mean square, and value range. These aggregated features capture the statistical characteristics of the IMU sensor data, enhancing the model's ability to learn temporal patterns and variations effectively.

3.4.3 Training Strategy

In this study, we employed a Random Forest classifier to investigate how different training strategies influence the performance of the environment classification model. The dataset consists of time-series sensor data collected from multiple participants, with labels

corresponding to three environmental conditions: indoor, outdoor, and semi-indoor. To explore the impact of different training approaches, we conducted four distinct training strategies: Train by Person (time-series), Train with Mixed Data including Semi-Indoor, Train with Different Time-Window including Semi-Indoor, and Train with Different Time-Window excluding Semi-Indoor.

The Train by Person (time-series) approach involved training the model using data from a single participant while leaving out data from other participants. This method evaluates how well the model generalizes across different individuals by testing on unseen participants. Since time-series sensor data varies significantly between individuals due to differences in movement patterns, sensor noise, and environmental conditions, this approach is essential for understanding the extent of person-specific biases in the model.

The Train with Mixed Data including Semi-Indoor method takes a different approach by aggregating data from all participants into a single training dataset. Unlike the previous method, where individual differences could lead to biased results, this approach ensures that the model is exposed to a diverse range of sensor patterns and environmental variations. By including semi-indoor environments in the training process, the model learns to differentiate between all three environmental categories, potentially improving its robustness when dealing with ambiguous or transitional environments that exhibit characteristics of both indoor and outdoor settings.

To further explore the impact of time-series features, we employed the Train with Different Time-Window including Semi-Indoor strategy, where a time-windowing technique was applied to structure the dataset. Instead of treating each sensor reading as an independent instance, a sliding window was used to aggregate contextual information from surrounding time steps. The size of the time window was varied to analyze how different window lengths affect classification performance. This approach is particularly useful for capturing temporal dependencies in sensor data, as environmental transitions often exhibit gradual changes rather than abrupt shifts.

Finally, the Train with Different Time-Window excluding Semi-Indoor strategy followed the same time-windowing approach but removed semi-indoor labels, restricting the model to classify only indoor and outdoor environments. The goal of this experiment was to assess how well the model performs when distinguishing between clearly defined environments without the complexity introduced by semi-indoor transitions. By comparing the results of this method

with those obtained when including semi-indoor labels, we can evaluate the extent to which the presence of transitional states affects the model's classification accuracy and whether the removal of semi-indoor data leads to clearer decision boundaries in the feature space.

Through these four experimental setups, we systematically analyzed how participant variability, dataset diversity, time-window selection, and label structure impact the classification performance of the Random Forest model. The insights gained from these experiments provide a deeper understanding of the challenges associated with environment classification based on time-series sensor data and inform best practices for designing robust models for real-world applications.

3.5 Evaluation

3.5.1 Precision, Recall and F1-score

Precision, recall, and F1-score are fundamental evaluation metrics in classification tasks, particularly when dealing with imbalanced datasets where accuracy alone may not provide a reliable assessment of model performance. Precision quantifies the proportion of correctly predicted positive cases among all cases classified as positive, making it particularly relevant in scenarios where false positives carry significant consequences, such as fraud detection or spam filtering. Recall, also known as sensitivity or the true positive rate, measures the proportion of actual positive cases that are correctly identified by the model, which is crucial in applications where missing a positive instance could lead to severe repercussions, such as medical diagnosis. Since precision and recall often present a trade-off, the F1-score is introduced as their harmonic mean, offering a balanced metric that ensures both false positives and false negatives are considered. This makes the F1-score particularly useful when working with imbalanced datasets, where achieving a balance between precision and recall is essential for robust model evaluation.

3.5.2 Shapiro-Wilk Test and One-way ANOVA

statistical analysis, the Shapiro-Wilk test and one-way ANOVA are commonly employed to assess data distribution and compare multiple groups to determine whether their differences

are statistically significant. The Shapiro-Wilk test is used to evaluate whether a dataset follows a normal distribution, making it an essential preliminary step before applying parametric tests that assume normality. It is particularly sensitive in detecting deviations from normality, even in small sample sizes, and is often used to validate the assumptions required for tests such as the t-test or ANOVA.

Once the assumption of normality is confirmed, the one-way ANOVA (Analysis of Variance) can be applied to compare the means of three or more independent groups. This parametric test determines whether at least one group's mean significantly differs from the others by analyzing the variance within and between groups. One-way ANOVA is widely used in experimental studies to compare the performance of multiple classification models or to evaluate the effects of different conditions on measured outcomes, such as accuracy or F1-score. When ANOVA indicates a significant difference, post-hoc tests (e.g., Tukey's HSD) are typically conducted to identify which specific groups differ from each other.

In this thesis, the Shapiro-Wilk test and one-way ANOVA were applied to systematically evaluate whether the observed differences in classification performance of the Random Forest model, influenced by participant variability, dataset diversity, time-window selection, and label structure, are statistically significant. The Shapiro-Wilk test was first used to assess the normality of the performance metrics, such as accuracy and F1-score, ensuring the validity of subsequent parametric tests. Upon confirming normality, one-way ANOVA was conducted to compare the classification performance across different experimental conditions. This approach allowed us to determine whether the variations in model performance were due to random fluctuations or represented statistically significant differences attributable to the experimental factors.

4. Results

4.1 Data overview

The dataset used in this study consists of 51 recorded scenarios collected from 12 participants. These scenarios include both the 7 predefined scenarios and custom scenarios created by participants. The distribution of scenario types is presented in Table 2.

	Scenario Type	Count
Defined	City Center Scenario 1	5
	City Center Scenario 2	5
	Tram 7 Scenario	6
	Train Station Scenario	2
	Central Campus Scenario	2
	Irchel Campus Scenario 1	8
	Irchel Campus Scenario 2	7
	Single Family Scenario	2
User-defined	Commute	6
	Single Family	1
	Shopping Tour	2
	Free	5

Table 2 Scenario Overview

For the predefined scenarios, the average data collection time was approximately 40 minutes per session. In contrast, custom scenarios, which were tailored to participants' daily routines, tended to have longer durations, ranging between 1 to 4 hours. This variation in duration reflects the nature of custom scenarios, where participants integrated the data collection process into their regular activities rather than following a structured experimental setup.

As part of the data preprocessing outlined in Chapter 3, a thorough label verification process was conducted. The vast majority of labels were correctly assigned by participants. However, due to the transitional nature of the semi-indoor category, some uncertainty arose in distinguishing semi-indoor from outdoor environments. One notable example was tram stations, which posed classification challenges. Some tram stations had extensive overhead coverings,

4. Results

while others had minimal structures, leading participants to perceive them differently in terms of exposure to environmental elements. The structural differences were apparent, the impact on GPS signals was generally minimal. To maintain labeling consistency, all tram stations were uniformly categorized as semi-indoor, aligning with the definition established in this thesis.



Figure 8 Outdoor scenes



Figure 9 Semi-indoor scenes

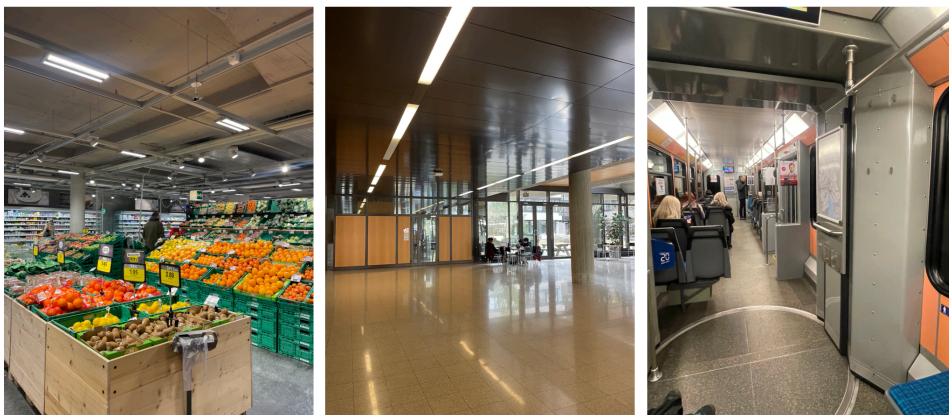


Figure 10 Indoor scenes

4. Results

Table 3 presents a list of typical scenes corresponding to indoor, outdoor, and semi-indoor environments, while Figure 8 to Figure 10 provide real-world images of these settings. The dataset contains a wide range of labeled locations, with indoor spaces including classrooms, offices, and shopping malls, outdoor locations covering parks, open streets, and squares, and semi-indoor spaces encompassing transit stations, covered walkways, and sheltered public areas.

Environment Type	Typical Place
Indoor	Tunnel, underground, classroom, shops, office room, train station
Outdoor	Irchel park, city streets, open squares, courtyard
Semi-indoor	Medizinerhof, tram station, tram, car, next to window, Niederdorfstrasse

Table 3 Typical places of each environment type

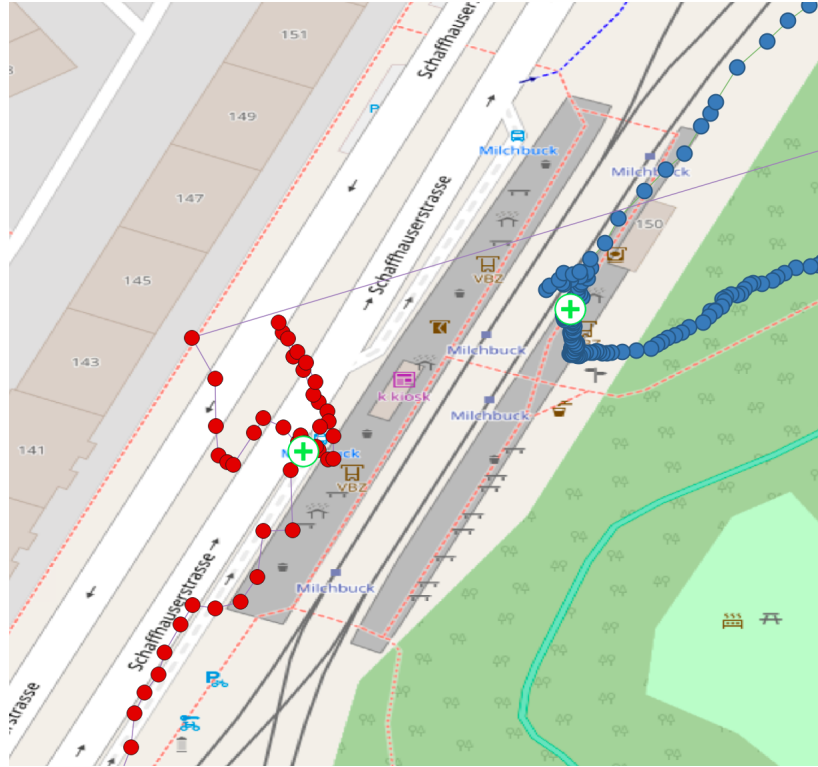


Figure 11 Comparison of Two Cases of GPS Points in Irchel Tram Station

As shown in Figure 11, even at the same station, some GPS data points are accurate while others are not, making classification challenging. To address this inconsistency, the station is uniformly categorized as semi-indoor.

Given the GPS sampling rate of one data entry per second, a large volume of spatial data was recorded. The total distribution of indoor, outdoor, and semi-indoor data points is summarized in Table 4, reflecting the relative proportions of each environment type within the dataset.

4. Results

Environment Type	Proportion	Count
Indoor	29.11%	61931
Outdoor	20.26%	43114
Semi-indoor	50.63%	107739

Table 4 Data ratio of each environment

As shown in Table 4, the data distribution shows that semi-indoor environments have the highest number of recorded instances, accounting for 50.63% (107,739 records). Indoor environments represent 29.11% of the data, with 61,931 recorded instances, while outdoor environments make up 20.26% of the data, totaling 43,114 records.

4.2 Hyperparameter Chosen

For the train by mixed data approach, given the dataset size of approximately 200,000 records, the grid search process was designed to optimize model performance while maintaining computational efficiency. The hyperparameter values selected for tuning included the number of trees set to 100, 200, and 300, the maximum tree depth set to None, 10, 20, and 30; and the number of features tested with sqrt and log2, both of which are commonly used for classification tasks.

In contrast, for the train by scenario and person approach, the dataset size is significantly smaller, with most subsets containing around 3,000 records. Therefore, the hyperparameter values were adjusted accordingly, with the number of trees set to 50, 100, and 200; the maximum tree depth set to None, 5, 10, and 15, and the number of features tested with sqrt and log2.

The evaluation criterion for selecting the optimal hyperparameter combination was the F1-score, as the dataset is imbalanced (as shown in Table 3). The F1-score, which balances precision and recall, ensures that the model is optimized for both false positives and false negatives, making it a more reliable metric for classification under class imbalance conditions.

Based on the grid search results, the optimal hyperparameters for the train by mixed data approach were identified as follows: the number of trees set to 200, the maximum tree depth set to 20, and the number of features set to sqrt. For the train by scenario and person approach, the best hyperparameter combination included the number of trees set to 200, the maximum tree depth set to 5, and the number of features set to log2.

These configurations were selected to balance model performance and computational efficiency while addressing the unique characteristics of each training strategy.

4.3 Comparison of train by mix-data with Different Time Windows

In this section, the comparison of the train by mixed data approach is presented. This approach involves merging data from all 51 scenarios into a single training dataset, ensuring comprehensive coverage of diverse environments. Importantly, the data sequence within each scenario is preserved, meaning that the temporal order of the time-series data is not shuffled.

4.3.1 Result of Different Time Windows

The performance of the model was evaluated across different lag ranges to investigate the impact of temporal context on classification accuracy. In this study, lag range refers to the time window size used during the training phase, representing the temporal span of data considered for feature extraction. Specifically, the lag range parameter defines the time window within which features are generated, incorporating data from both the preceding and following time steps relative to each data point. This approach allows the model to capture temporal dependencies and contextual information from surrounding data points.

To explore the effect of temporal context, the time window (lag range) was set to 0 s, 3 s, 5 s, 10 s, 15 s, and 20 s. Each lag range represents the inclusion of data within the specified time frame before and after the target data point, enabling the model to learn from both past and future observations within the defined window.

Tables 5 to 16 present the confusion matrices and classification reports for models trained with different time window sizes (lag ranges). These results include the precision, recall, and F1-score for each class—indoor, outdoor, and semi-indoor—as well as the macro average and weighted average metrics for precision, recall, and F1-score.

The macro average represents the arithmetic mean of the precision, recall, and F1-score calculated for each class independently. It treats all classes equally, regardless of their size.

4. Results

In contrast, the weighted average takes into account the support (i.e., the number of instances) of each class when calculating the average. This means that classes with more instances have a greater influence on the final metric.

		Predicted		
		Indoor	Outdoor	Semi-indoor
Actual	Indoor	13183	482	3794
	Outdoor	710	8960	2380
	Semi-indoor	1539	1187	10322

Table 5 Confusion matrix when lag range =0

Class	Precision	Recall	F1-score	Support
Indoor	0.85	0.76	0.8	17,459
Outdoor	0.84	0.74	0.79	12,050
Semi-indoor	0.63	0.79	0.7	13,048
Macro Avg	0.77	0.76	0.76	42,557
Weighted Avg	0.78	0.76	0.77	42,557
Accuracy		0.76		42,557

Table 6 Classification report when lag range =0

		Predicted		
		Indoor	Outdoor	Semi-indoor
Actual	Indoor	13265	557	3637
	Outdoor	621	9188	2241
	Semi-indoor	1557	1143	10348

Table 7 Confusion matrix when lag range =3

Class	Precision	Recall	F1-score	Support
Indoor	0.86	0.76	0.81	17,459
Outdoor	0.84	0.76	0.8	12,050
Semi-indoor	0.64	0.79	0.71	13,048
Macro Avg	0.78	0.77	0.77	42,557
Weighted Avg	0.79	0.77	0.77	42,557
Accuracy		0.77		42,557

Table 8 Classification report when lag range =3

4. Results

		Predicted		
		Indoor	Outdoor	Semi-indoor
Actual	Indoor	13272	625	3562
	Outdoor	616	9259	2175
	Semi-indoor	1268	1132	10648

Table 9 Confusion matrix when lag range =5

Class	Precision	Recall	F1-score	Support
Indoor	0.88	0.76	0.81	17,459
Outdoor	0.84	0.77	0.8	12,050
Semi-indoor	0.65	0.82	0.72	13,048
Macro Avg	0.79	0.78	0.78	42,557
Weighted Avg	0.8	0.78	0.78	42,557
Accuracy		0.78		42,557

Table 10 Classification report when lag range =5

		Predicted		
		Indoor	Outdoor	Semi-indoor
Actual	Indoor	13372	599	3488
	Outdoor	635	9179	2236
	Semi-indoor	1316	1071	10661

Table 11 Confusion Matrix when lag range =10

Class	Precision	Recall	F1-score	Support
Indoor	0.87	0.77	0.82	17,459
Outdoor	0.85	0.76	0.8	12,050
Semi-indoor	0.65	0.82	0.72	13,048
Macro Avg	0.79	0.78	0.78	42,557
Weighted Avg	0.8	0.78	0.78	42,557
Accuracy		0.78		42,557

Table 12 Classification report when lag range=10

4. Results

		Predicted		
		Indoor	Outdoor	Semi-indoor
Actual	Indoor	13346	612	3501
	Outdoor	702	9037	2311
	Semi-indoor	1441	999	10608

Table 13 Confusion matrix when lag range =15

Class	Precision	Recall	F1-score	Support
Indoor	0.86	0.76	0.81	17,459
Outdoor	0.85	0.75	0.8	12,050
Semi-indoor	0.65	0.81	0.72	13,048
Macro Avg	0.79	0.78	0.78	42,557
Weighted Avg	0.79	0.78	0.78	42,557
Accuracy		0.78		42,557

Table 14 Classification report when lag range =15

		Predicted		
		Indoor	Outdoor	Semi-indoor
Actual	Indoor	13353	585	3521
	Outdoor	647	9014	2389
	Semi-indoor	1555	876	10617

Table 15 Confusion matrix when lag range=20

Class	Precision	Recall	F1-score	Support
Indoor	0.86	0.76	0.81	17,459
Outdoor	0.86	0.75	0.8	12,050
Semi-indoor	0.64	0.81	0.72	13,048
Macro Avg	0.79	0.78	0.78	42,557
Weighted Avg	0.79	0.78	0.78	42,557
Accuracy		0.78		42,557

Table 16 Classification report when lag range=20

4.3.2 Result Comparison

Table 17 presents the precision, recall, and F1-score of the test set for each lag range. To enhance the visibility of performance differences across different lag ranges, Figure 12 visualizes these results using bar charts, making it easier to compare the model's performance under varying temporal contexts.

Table 18 displays the average accuracy and runtime of the test set for each lag range, providing insights into both the model's predictive performance and computational efficiency. Similarly, Table 19 visualizes these metrics, highlighting the trade-offs between accuracy and processing time across different lag ranges. This comparison helps identify the optimal lag range that balances classification performance and computational cost.

Lag Range	Test Set Precision	Test Set Recall	Test Set F1-score
0	0.78	0.76	0.77
3	0.79	0.77	0.77
5	0.8	0.78	0.78
10	0.8	0.78	0.78
15	0.79	0.78	0.78
20	0.79	0.78	0.78

Table 17 Test set precision, recall, and F1-score when lag range (0-20)

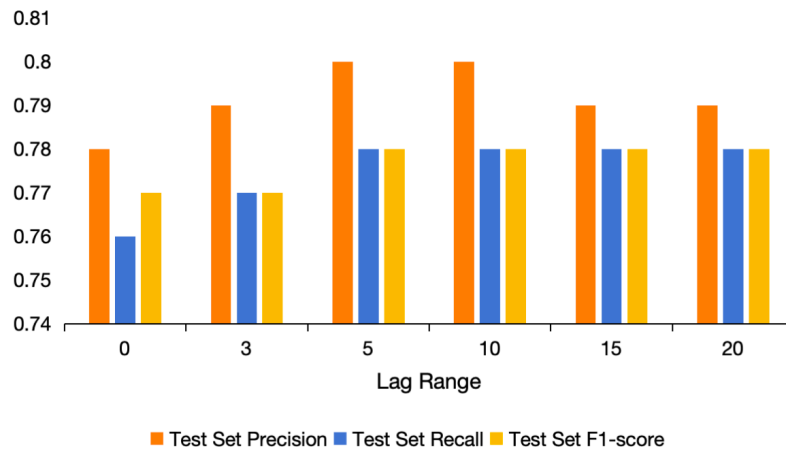


Figure 12 Histogram of test set precision, recall and F1-score when lag range (0-20)

4. Results

Lag Range	Test Set Accuracy	Fold Average Accuracy	Training Time
0	0.76	0.72	0:02:12
3	0.77	0.73	0:07:15
5	0.78	0.74	0:10:51
10	0.78	0.74	0:14:05
15	0.78	0.75	0:16:30
20	0.78	0.75	1:30:24

Table 18 Accuracy and training time of lag range (0-20)

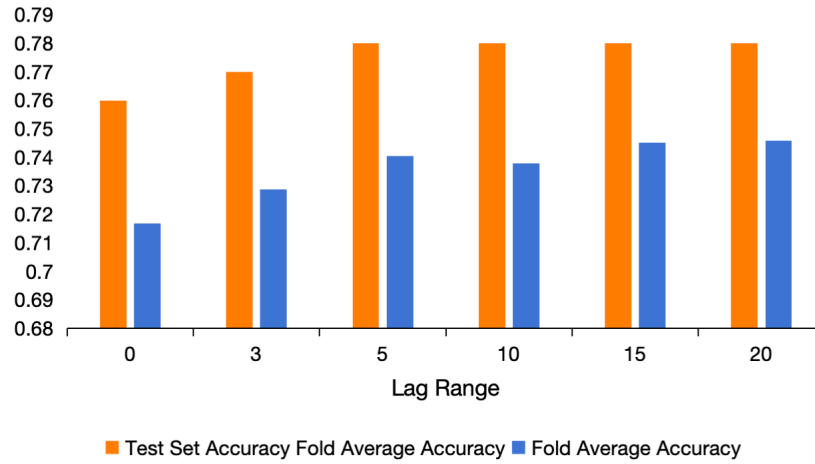


Table 19 Histogram of accuracy of lag range (0-20)

4.3.3 One-way ANOVA

After training the model with different lag ranges, a one-way ANOVA test was conducted to evaluate whether the lag range has a significant impact on the model's accuracy. Prior to the ANOVA test, the Shapiro-Wilk test was performed to assess the normality of the accuracy data, as normality is a key assumption for conducting ANOVA. The Shapiro-Wilk test evaluates whether the data distribution significantly deviates from a normal distribution, with a p-value greater than 0.05 indicating that the data can be considered normally distributed.

As shown in Table 20, the results of the Shapiro-Wilk test indicate that the accuracy data across different lag ranges follow a normal distribution, satisfying the condition for proceeding with the one-way ANOVA test. The ANOVA test results, presented in Table 21, reveal whether the differences in model accuracy across various lag ranges are statistically significant.

4. Results

Lag Range	Metric	Statistic	p-value	Normal Distribution
0	Precision	0.9658	0.863	Yes
	Recall	0.9826	0.9635	Yes
	F1-Score	0.9773	0.9371	Yes
3	Precision	0.9316	0.5927	Yes
	Recall	0.9846	0.9719	Yes
	F1-Score	0.983	0.9655	Yes
5	Precision	0.9388	0.6498	Yes
	Recall	0.9905	0.9904	Yes
	F1-Score	0.9784	0.9435	Yes
10	Precision	0.9063	0.4126	Yes
	Recall	0.9853	0.9748	Yes
	F1-Score	0.9814	0.9584	Yes
15	Precision	0.913	0.4568	Yes
	Recall	0.9909	0.9913	Yes
	F1-Score	0.9885	0.9851	Yes
20	Precision	0.8868	0.3019	Yes
	Recall	0.9909	0.9913	Yes
	F1-Score	0.9838	0.9689	Yes

Table 20 Shapiro-Wilk test when lag range (0-20)

Metric	F-statistic	p-value	Conclusion
Precision	0.0435	0.9988	Not significant
Recall	0.027	0.9996	Not significant
F1-Score	0.0141	0.9999	Not significant

Table 21 One-way ANOVA result when lag range (0-20)

4.3.4 Feature Importance

Table 22 presents the top 5 most important features for the Random Forest model across different lag ranges. The feature importance is determined based on the mean decrease in impurity (MDI), also known as Gini importance which measures how much each feature contributes to reducing the impurity in the decision trees. Features with higher importance scores indicate a greater impact on the model's decision-making process.

4. Results

Lag Range	Top5 Important Features
0	1. speedAccuracy(m/s) 2. pressure(Pa) 3. accZ(mg)_mean_square 4. nrSat 5. groundSpeed(m/s)
3	1. speedAccuracy(m/s) 2. speedAccuracy(m/s)_lag1 3. accY(mg)_max_lag-1 4. speedAccuracy(m/s)_lag-2 5. speedAccuracy(m/s)_lag3
5	1. speedAccuracy(m/s) 2. speedAccuracy(m/s)_lag3 3. accZ(mg)_mean_abs_lag1 4. accZ(mg)_mean_abs_lag-3 5. speedAccuracy(m/s)_lag-1
10	1. speedAccuracy(m/s)_lag-1 2. speedAccuracy(m/s)_lag1 3. accZ(mg)_mean_abs_lag-3 4. accZ(mg)_mean_abs_lag-4 5. speedAccuracy(m/s)_lag2
15	1. accZ(mg)_mean_abs_lag-3 2. speedAccuracy(m/s) 3. speedAccuracy(m/s)_lag1 4. accZ(mg)_mean_abs_lag-9 5. accZ(mg)_mean_abs_lag-4
20	1. speedAccuracy(m/s)_lag-2 2. speedAccuracy(m/s)_lag2 3. accZ(mg)_mean_abs_lag-6 4. accZ(mg)_mean_abs_lag-1 5. speedAccuracy(m/s)_lag-1

Table 22 Top 5 important features of lag range (0-20)

4.4 Comparison of Training by Time-series Data

Time-series data refers to data collected in a sequential, time-ordered manner, where the temporal continuity of observations is crucial for capturing trends and dependencies over time. In this analysis, each scenario's data was used as a separate training dataset to preserve this temporal continuity. The model's performance was evaluated based on the training results from all scenarios, with accuracy analyzed from two perspectives: across different participants and across various scenarios. This approach helps to determine whether the model's performance

4. Results

is influenced by individual-specific patterns or the conditions under which the data was collected.

4.4.1 Training Results Averaged by Person

The training results for all scenarios are provided in Appendix C. Table 23 presents the model's performance metrics grouped by different participants, including precision, recall, F1-score, and accuracy on the test set, as well as the average accuracy across all folds in cross-validation. To better illustrate these results, Figure 13 visualizes the performance metrics using bar charts, highlighting observable differences in model performance across individuals.

Participant	Test Set accuracy	Test Set precision	Test Set Recall	Test Set F1-Score	Fold Average Accuracy
1	0.74	0.77	0.75	0.73	0.75
2	0.71	0.85	0.81	0.81	0.81
3	0.80	0.73	0.71	0.69	0.71
4	0.75	0.86	0.82	0.82	0.82
5	0.35	0.73	0.68	0.66	0.68
6	0.75	0.78	0.77	0.75	0.77
7	0.97	0.78	0.75	0.74	0.75
8	0.55	0.84	0.84	0.83	0.84
9	0.65	0.78	0.74	0.73	0.74
10	0.40	0.73	0.63	0.62	0.63
11	0.75	0.78	0.79	0.76	0.79
12	0.82	0.71	0.73	0.71	0.73
Average	0.68	0.79	0.76	0.74	0.76

Table 23 Model performance averaged by person

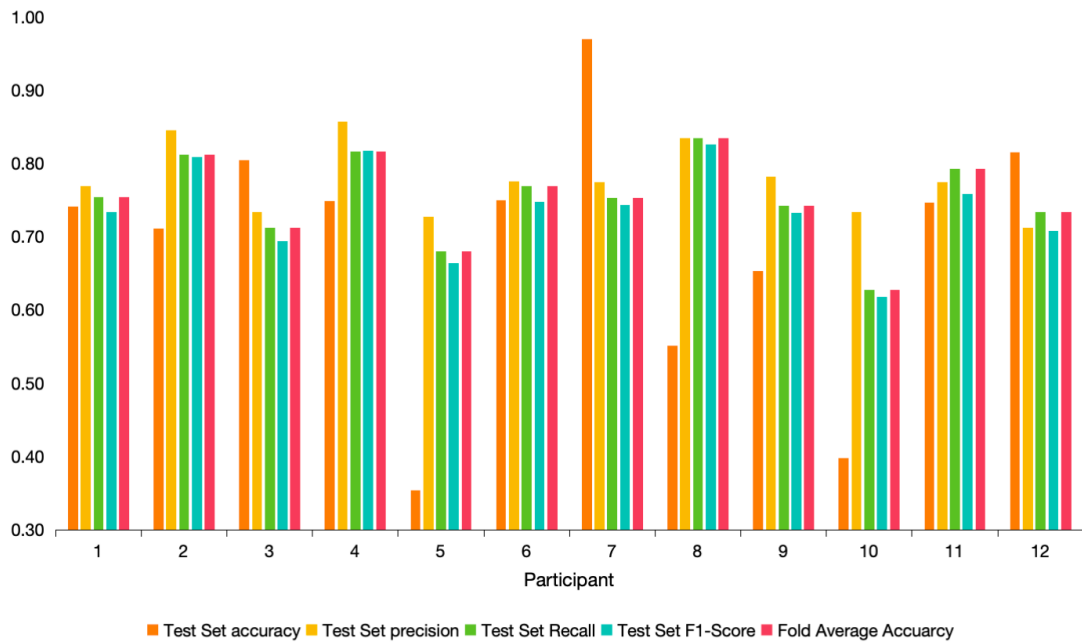


Figure 13 Histogram of model performance averaged by person

4. Results

Tables 24 to 26 present the results of the normality tests, consistent with the previous analysis. The Shapiro-Wilk test results indicate that the majority of the data follow a normal distribution, satisfying the assumptions required for conducting a one-way ANOVA test. Based on this, a one-way ANOVA was performed to evaluate whether participant variability has a significant impact on the model's performance. The ANOVA results, shown in Table 27, reveal that the effect of participants on the model's classification performance is not statistically significant, suggesting that individual differences do not lead to substantial variations in the model's accuracy, precision, recall, or F1-score. It is important to note that Participant 8 was excluded from the analysis due to insufficient data.

Participant	p-value	Normally Distributed
1	0.97	Yes
2	0.96	Yes
3	0.18	Yes
4	0.62	Yes
5	0.90	Yes
6	0.30	Yes
7	0.55	Yes
9	0.98	Yes
10	0.19	Yes
11	0.15	Yes
12	0.11	Yes
8		Skipped

Table 24 Shapiro-Wilk test of precision averaged by person

Participant	p-value	Normally Distributed
1	0.31	Yes
2	0.05	Yes
3	0.63	Yes
4	0.75	Yes
5	0.86	Yes
6	0.61	Yes
7	0.32	Yes
9	0.16	Yes
10	0.72	Yes
11	0.56	Yes
12	0.22	Yes
8		Skipped

Table 25 Shapiro-Wilk test recall averaged by person

4. Results

Participant	p-value	Normally Distributed
1	0.42	Yes
2	0.36	Yes
3	0.97	Yes
4	0.62	Yes
5	0.85	Yes
6	0.09	Yes
7	0.01	No
9	0.62	Yes
10	0.81	Yes
11	0.08	Yes
12	0.18	Yes
8		Skipped

Table 26 Shapiro-Wilk test F1-score averaged by person

Metric	F-statistic	p-value	Conclusion
Precision	0.94	0.51	Not significant
Recall	0.74	0.70	Not significant
F1-Score	0.69	0.74	Not significant

Table 27 One-way ANOVA result of precision, recall, F1-score averaged by person

Table 28 presents the top five most important features for each participant, aggregated across all scenarios. These features were identified based on their overall contribution to the model's performance, reflecting the key factors that influence classification accuracy for each individual. This analysis provides insights into the dominant patterns in the data that are most relevant to participant-specific model performance.

4. Results

Participant	Top5 Important Features		
1	1. nrSat_lag1	7	1. pressure(Pa)_lag-1
	2. nrSat_lag2		2. accX(mg)_std
	3. speedAccuracy(m/s)_lag3		3. pressure(Pa)_lag-4
	4. longitude		4. accX(mg)_mean_abs_lag1
	5. pressure(Pa)_lag-1		5. pressure(Pa)_lag-2
2	1. nrSat_lag1	8	1. nrSat_lag4
	2. nrSat_lag2		2. nrSat_lag2
	3. speedAccuracy(m/s)_lag3		3. nrSat_lag1
	4. humidity(percentage)_lag3		4. pressure(Pa)_lag-1
	5. longitude		5. pressure(Pa)_lag-4
3	1. latitude_lag3	9	1. nrSat_lag1
	2. accZ(mg)_max		2. nrSat_lag2
	3. humidity(percentage)_lag-5		3. accZ(mg)_max
	4. accX(mg)_std		4. longitude
	5. humidity(percentage)		5. speedAccuracy(m/s)_lag3
4	1. nrSat_lag1	10	1. nrSat_lag1
	2. speedAccuracy(m/s)_lag3		2. nrSat_lag2
	3. accY(mg)_std_lag1		3. speedAccuracy(m/s)_lag3
	4. longitude		4. longitude
	5. temperature(C)_lag-2		5. latitude_lag-3
5	1. nrSat_lag1	11	1. nrSat_lag2
	2. nrSat_lag2		2. nrSat_lag1
	3. speedAccuracy(m/s)_lag3		3. speedAccuracy(m/s)_lag3
	4. latitude_lag-3		4. longitude
	5. accZ(mg)_mean_abs_lag-2		5. accX(mg)_mean_lag2
6	1. pressure(Pa)_lag-1	12	1. nrSat_lag2
	2. accX(mg)_std		2. nrSat_lag1
	3. pressure(Pa)_lag-4		3. speedAccuracy(m/s)_lag3
	4. accX(mg)_mean_abs_lag1		4. longitude
	5. pressure(Pa)_lag-2		5. pressure(Pa)_lag-4

Table 28 Top 5 importance features of each person

4.4.2 Training Results Averaged by Scenario

Table 29 presents the model's performance across different scenarios, including test set accuracy, precision, recall, F1-score, and fold average accuracy. The results show variations in performance, with accuracy ranging from 0.42 to 0.89 and corresponding differences in other metrics. Figure 14 visualizes these results using bar charts, clearly highlighting the substantial differences in model performance across scenarios.

4. Results

Scenario	Test Set accuracy	Test Set precision	Test Set Recall	Test Set F1-Score	Fold Average Accuracy
City Center Scenario 1	0.80	0.84	0.86	0.84	0.86
City Center Scenario 2	0.89	0.84	0.84	0.83	0.84
Irchel Campus Scenario 1	0.84	0.74	0.67	0.67	0.67
Irchel Campus Scenario 2	0.59	0.62	0.56	0.53	0.56
Central Campus Scenario	0.42	0.81	0.70	0.71	0.70
Single Family Scenario	0.50	0.85	0.87	0.85	0.87
Train Station Scenario	0.77	0.83	0.85	0.81	0.85
Tram 7 Scenario	0.65	0.88	0.84	0.83	0.84
Free	0.60	0.79	0.76	0.75	0.76
Average	0.68	0.79	0.76	0.74	0.76

Table 29 Model performance averaged by scenario

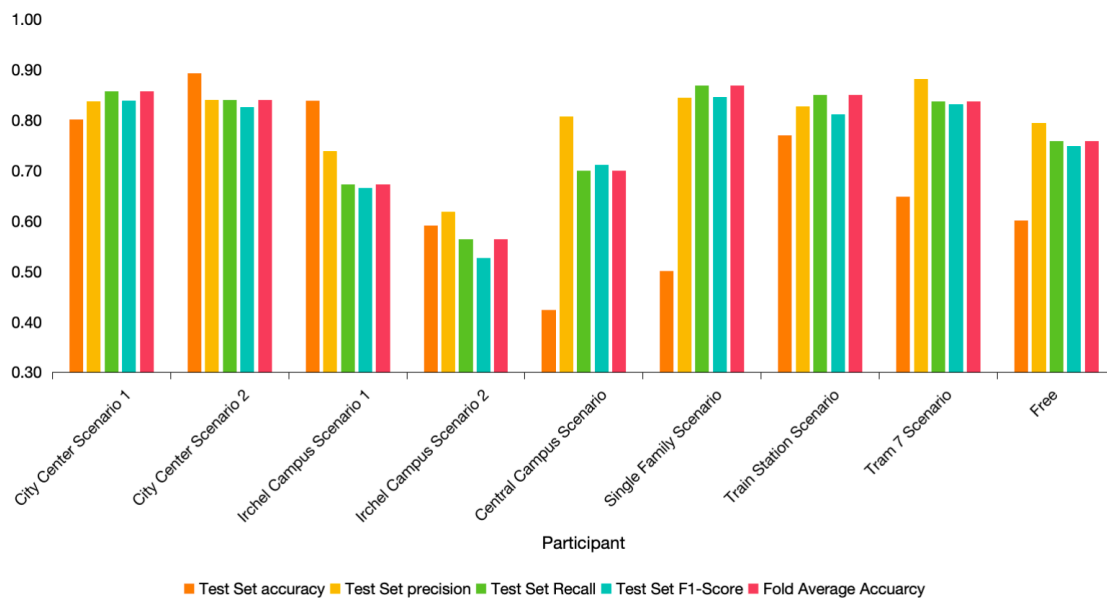


Figure 14 Histogram of model performance averaged by scenario

Tables 30 to 32 present the normality test results for scenario data, indicating that the majority of the data follow a normal distribution, thus meeting the assumptions required for conducting a one-way ANOVA test. Based on this, a one-way ANOVA was performed to assess whether different scenarios significantly impact the model's performance. The ANOVA results, shown in Table 33, reveal statistically significant differences in precision, recall, and F1-score, with F-statistics of 4.47, 5.46, and 5.20, respectively, and p-values all below 0.01. This suggests that the model's performance varies across scenarios, likely due to scenario-specific factors. It is important to note that scenarios with insufficient data were excluded from the analysis.

4. Results

Scenario	p-value	Normally Distributed
City Center Scenario 1	0.91	Yes
City Center Scenario 2	0.05	Yes
Irchel Campus Scenario 1	0.46	Yes
Irchel Campus Scenario 2	0.24	Yes
Tram 7 Scenario	0.19	Yes
Free	0.36	Yes
Central Campus Scenario		Skipped
Single Family Scenario		Skipped
Train Station Scenario		Skipped

Table 30 Shapiro-Wilk test of precision averaged by scenario

Scenario	p-value	Normally Distributed
City Center Scenario 1	0.09	Yes
City Center Scenario 2	0.08	Yes
Irchel Campus Scenario 1	0.58	Yes
Irchel Campus Scenario 2	0.77	Yes
Tram 7 Scenario	0.25	Yes
Free	0.48	Yes
Central Campus Scenario		Skipped
Single Family Scenario		Skipped
Train Station Scenario		Skipped

Table 31 Shapiro-Wilk test of recall averaged by scenario

Scenario	p-value	Normally Distributed
City Center Scenario 1	0.63	Yes
City Center Scenario 2	0.25	Yes
Irchel Campus Scenario 1	0.54	Yes
Irchel Campus Scenario 2	0.83	Yes
Tram 7 Scenario	0.17	Yes
Free	0.38	Yes
Central Campus Scenario		Skipped
Single Family Scenario		Skipped
Train Station Scenario		Skipped

Table 32 Shapiro-Wilk test of F1-score averaged by scenario

Metric	F-statistic	p-value	Conclusion
Precision	4.47	0.00	Significant
Recall	5.46	0.00	Significant
F1-Score	5.20	0.00	Significant

Table 33 One-way ANOVA result of precision, recall, F1-score averaged by scenario

4. Results

Table 34 presents the top five most important features for each scenario, considering data from all participants. These features were selected based on their contribution to the model's predictive performance, highlighting the key factors that influence classification accuracy within different environmental contexts. This analysis helps identify scenario-specific patterns and the features that the model relies on most in varying conditions.

Scenario	Top 5 Important Features		
City Center Scenario 1	1. nrSat_lag2	Single Family Scenario	1. accX(mg)_mean_lag2
	2. nrSat_lag1		2. accX(mg)_min
	3. longitude		3. accX(mg)_mean_lag1
	4. speedAccuracy(m/s)_lag3		4. accX(mg)_mean_lag-1
	5. temperature(C)_lag5		5. accX(mg)_mean_lag5
City Center Scenario 2	1. nrSat_lag1	Train Station Scenario	1. altitude(m)_lag1
	2. nrSat_lag2		2. altitude(m)_lag4
	3. nrSat_lag4		3. latitude_lag-3
	4. speedAccuracy(m/s)_lag3		4. accY(mg)_std_lag2
	5. nrSat_lag4		5. accZ(mg)_std_lag1
Irchel Campus Scenario 1	1. nrSat_lag2	Tram 7 Scenario	1. nrSat_lag1
	2. nrSat_lag1		2. nrSat_lag2
	3. nrSat		3. speedAccuracy(m/s)_lag3
	4. nrSat_lag4		4. nrSat
	5. longitude		5. nrSat_lag4
Irchel Campus Scenario 2	1. latitude_lag-3	Free Scenario	1. nrSat_lag2
	2. latitude_lag1		2. nrSat_lag1
	3. latitude_lag-5		3. nrSat
	4. latitude_lag3		4. nrSat_lag4
	5. latitude_lag4		5. speedAccuracy(m/s)_lag3
Central Campus Scenario	1. pressure(Pa)_lag-1		
	2. humidity(percentage)		
	3. pressure(Pa)_lag1		
	4. pressure(Pa)		
	5. nrSat_lag1		

Table 34 Top 5 importance features of each scenario

Based on the analysis in Table 29, City Center Scenario 2 and Irchel Campus Scenario 1 demonstrate the best overall performance among all scenarios. To further investigate the model's behavior in these scenarios, the models of the best-performing and worst-performing participants within each scenario were selected. Their corresponding confusion matrices and classification reports are presented in Tables 35 to 38.

These tables provide detailed insights into the model's precision, recall, and F1-score for each environment type. It is important to note that, since the test data only covers a subset of

4. Results

scenarios, it is normal for one or two environment types to be missing in certain cases. This reflects the variability in data distribution and highlights the influence of scenario-specific conditions on the classification performance.

		Predicted		
		Indoor	Outdoor	Semi-indoor
Actual	Indoor	0	0	0
	Outdoor	0	334	1
	Semi-indoor	2	0	169

Table 35 Confusion matrix of a good performance example in City Center Scenario

Class	Precision	Recall	F1-score	Support
Indoor	0	0	0	0
Outdoor	1	1	1	335
Semi-indoor	0.99	0.99	0.99	171
Macro Avg	0.66	0.66	0.66	506
Weighted Avg	1	0.99	1	506
Accuracy		0.99		506

Table 36 Classification report of a good performance example in City Center Scenario

		Predicted	
		Indoor	Semi-indoor
Actual	Indoor	37	294
	Semi-indoor	4	31

Table 37 Confusion matrix of a bad performance example in Irchel Campus Scenario

Class	Precision	Recall	F1-score	Support
Indoor	0.9	0.11	0.2	331
Semi-indoor	0.1	0.89	0.17	35
Macro Avg	0.5	0.5	0.19	366
Weighted Avg	0.83	0.19	0.2	366
Accuracy		0.19		366

Table 38 Classification report of a bad performance example in Irchel Campus Scenario

4.5 Comparison of Training with and without Semi-indoor

The decision to exclude the semi-indoor category was based on its potentially ambiguous and controversial definition, which may introduce uncertainty into the classification process. To explore whether removing this category could significantly improve model performance, classification experiments were conducted both with and without the semi-indoor data.

Tables 39 and 40 present the confusion matrix and classification report for the model trained without semi-indoor included. It shows the precision, recall, F1-score, and support for each class (indoor and outdoor), along with the overall accuracy, macro average, and weighted average scores.

		Predicted	
		Indoor	Outdoor
Actual	Indoor	12042	946
	Outdoor	950	7071

Table 39 Confusion matrix of mix-data training without semi-indoor

Class	Precision	Recall	F1-score	Support
Indoor	0.93	0.93	0.93	12988
Outdoor	0.88	0.88	0.88	8021
Macro Avg	0.9	0.9	0.9	21009
Weighted Avg	0.91	0.91	0.91	21009
Accuracy		0.91		21009

Table 40 Classification report of mix-data training without semi-indoor

Table 41 and Figure 15 display the participant-wise performance metrics, including test set accuracy, precision, recall, F1-score, and fold average accuracy. This table summarizes the

4. Results

model's performance across different participants when trained without semi-indoor data.

Participant	Test Set accuracy	Test Set precision	Test Set Recall	Test Set F1-Score	Fold Average Accuracy
1	0.86	0.85	0.86	0.86	0.86
2	0.95	1	0.96	0.95	0.95
3	0.93	0.77	0.91	0.93	0.92
4	0.9	0.86	0.88	0.9	0.89
5	0.88	0.98	0.9	0.88	0.88
6	0.9	0.91	0.89	0.9	0.88
7	0.9	0.9	0.89	0.9	0.88
8	0.82	1	0.8	0.82	0.81
9	0.93	0.99	0.94	0.93	0.92
10	0.77	1	0.8	0.77	0.76
11	0.83	0.97	0.84	0.83	0.82
12	0.82	1	0.8	0.81	0.8
Average	0.88	0.93	0.88	0.88	0.87

Table 41 Model performance averaged by person without semi-indoor

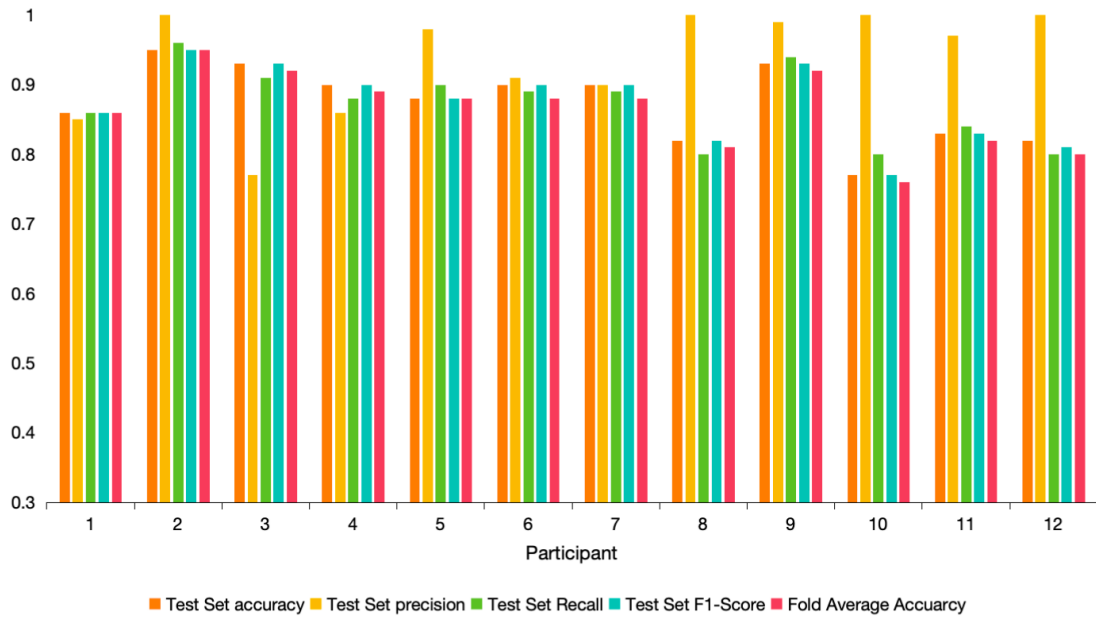


Figure 15 Histogram of model performance averaged by person without semi-indoor

Scenario	Test Set accuracy	Test Set precision	Test Set Recall	Test Set F1-Score	Fold Average Accuracy
City Center Scenario 1	0.95	1	0.98	0.95	0.95
City Center Scenario 2	0.94	0.99	0.98	0.94	0.94
Irchel Campus Scenario 1	0.86	1	0.86	0.86	0.85
Irchel Campus Scenario 2	0.86	1	0.86	0.86	0.85
Central Campus Scenario	0.99	1	1	0.99	0.99
Single Family Scenario	0.76	0.94	0.78	0.76	0.74
Train Station Scenario	0.88	1	0.9	0.88	0.88
Tram 7 Scenario	0.88	0.91	0.88	0.88	0.88
Free	0.84	0.89	0.88	0.84	0.84
Average	0.88	0.93	0.88	0.88	0.87

Table 42 Model performance averaged by scenario without semi-indoor

4. Results

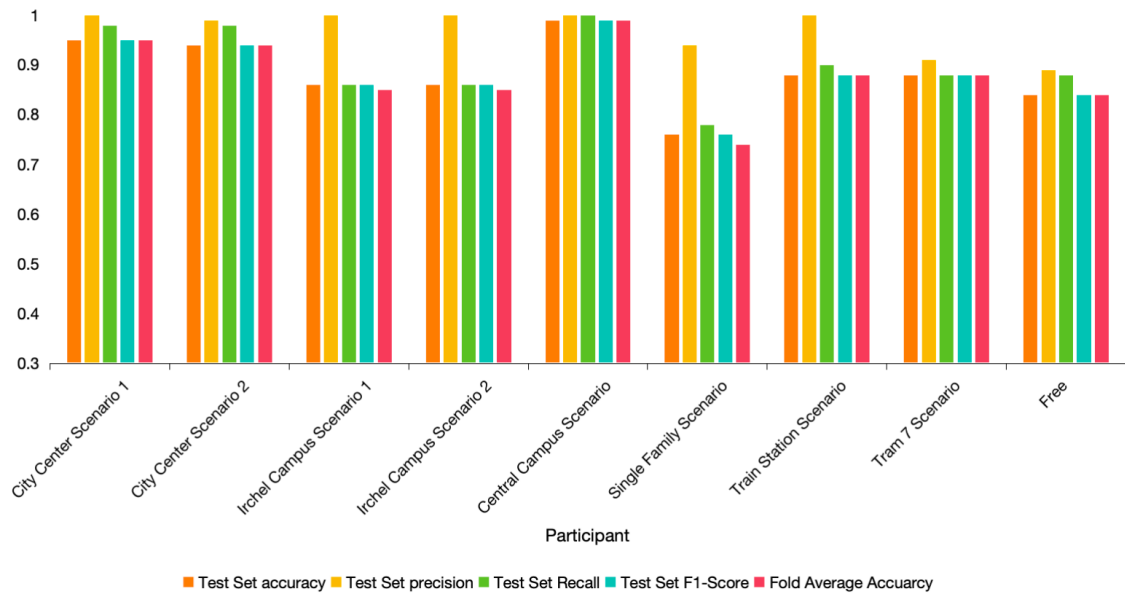


Figure 16 Histogram of model performance averaged by scenario without semi-indoor

Table 42 and Figure 16 present the corresponding performance metrics when the model is trained without the semi-indoor category. This allows for a direct comparison to assess the impact of excluding the semi-indoor class on model accuracy and robustness.

5. Discussion

5.1 Best Time Window Size

From Tables 5 to 16, the model's performance remains stable across different lags (0s, 3s, 5s, 10s, 15s, and 20s), with minimal variation in accuracy, precision, recall, and F1-scores, indicating that lag time has little impact on predictive capability. For class-wise performance, the indoor class shows high precision (0.85-0.88) but lower recall (0.76), suggesting it correctly identifies most indoor instances but misses some. The outdoor class achieves balanced precision and recall (0.84-0.86 and 0.74-0.77 respectively), while the semi-indoor class has lower precision (0.63-0.65) but higher recall (0.79-0.82), indicating it captures most semi-indoor instances but misclassifies some from other classes. Both macro and weighted averages are consistent across lags, reflecting balanced performance across all classes without bias.

Table 17 and Table 18 show model's performance across different lag ranges demonstrates consistent and stable results in terms of precision, recall, F1-score, and accuracy. The test set accuracy ranges from 0.76 to 0.78, with the highest accuracy achieved at lag ranges of 5, 10, 15, and 20. The F1-scores and recall metrics also show minimal variation, indicating that the model is robust across all tested lag ranges. However, there is a noticeable increase in training time as the lag range expands, with the longest training time occurring at lag 20 (1 hour and 30 minutes), compared to lag 0 (2 minutes and 12 seconds). Precision, recall, and F1-scores improve slightly as the lag range increases, with the best performance observed at lag 5 and beyond. Accuracy peaks at 0.78 for lag ranges 5, 10, 15, and 20, showing no significant improvement beyond lag 5. Training time increases significantly with larger lag ranges, and the fold average accuracy improves slightly but remains marginal.

The optimal lag range is lag 5, as it achieves the highest test set accuracy (0.78) and F1-score (0.78) while maintaining a reasonable training time (10 minutes and 51 seconds). Larger lag ranges do not provide significant performance improvements but require substantially more computational resources and time. If computational efficiency is a priority, lag 0 could be

considered, but lag 5 offers a better balance between performance and efficiency. In conclusion, lag 5 is recommended for its optimal balance of accuracy, computational efficiency, and training time.

Besides, as in Table 22, in the lag 5 window, speedAccuracy(m/s) and its lagged features (e.g., lag3, lag-1) consistently contribute to the model's predictions, indicating that lag 5 effectively captures key information related to the target variable. Its balanced inclusion of both past and future lagged features enhances prediction accuracy. Compared to other windows, lag 0 relies solely on current data, missing temporal context; lag 3 introduces limited lagged features, insufficient for long-term dependencies; while larger windows like lag 10, 15, and 20 increase computational cost without significant performance gains. Overall, lag 5 offers the best balance of feature importance, diversity, and efficiency.

5.2 The Comparison of Models between Different Training Strategies

5.2.1 Model Trained with Mix Data and with Time-series Data

The mixed data model demonstrates superior performance compared to the scenario data model, achieving a higher overall accuracy of 0.78 versus 0.68. While the scenario data model shows slightly higher precision (0.79 compared to the mixed data model's 0.80 weighted average), the mixed data model excels in recall (0.78 weighted average vs. 0.76) and F1-score (0.78 weighted average vs. 0.74), indicating a better balance between precision and recall. Additionally, the mixed data model performs more consistently across all classes, particularly in the semi-indoor category, where it achieves a higher recall of 0.82. This suggests that the mixed data model generalizes better to unseen data, likely due to the diversity of the mixed training dataset, which captures a wider range of scenarios and conditions. In contrast, the scenario data model, despite its marginally higher precision, falls short in overall accuracy and class-wise performance. Therefore, the mixed data model is the better choice for this classification task, offering stronger overall performance and robustness.

5.2.2 Model Trained with and without Semi-indoor Dataset

The comparison between the multi-class classification model (including the semi-indoor category) and the binary classification model (excluding semi-indoor) reveals significant differences in performance, driven by the inherent complexity of the tasks and the nature of the classes involved. The binary classification model achieves a substantially higher accuracy of 0.91, with strong class-wise performance for both Indoor (F1-score = 0.93) and Outdoor (F1-score = 0.88) categories. In contrast, the multi-class model achieves a lower accuracy of 0.78, with the semi-indoor class exhibiting the weakest performance (F1-score = 0.72). This disparity highlights the challenges introduced by the inclusion of the semi-indoor category, which is inherently ambiguous and shares characteristics with both Indoor and Outdoor environments.

The binary classification model benefits from the simplicity of distinguishing between only two distinct classes, allowing the Random Forest algorithm to create clear and effective decision boundaries. This results in higher precision, recall, and F1-scores for both Indoor and Outdoor classes. On the other hand, the multi-class model must handle the additional complexity of separating three classes, including the less distinct semi-indoor category. This increases the likelihood of misclassification, particularly for instances that lie near the decision boundaries between semi-indoor and the other two classes. The lower precision (0.65) and higher recall (0.82) for the semi-indoor class suggest that while the model captures most true positives, it frequently misclassifies semi-indoor instances as either Indoor or Outdoor.

The performance gap between the two models can be attributed to the increased complexity of the multi-class task and the ambiguous nature of the semi-indoor category. Random Forest models, while effective in binary classification, face greater challenges in multi-class settings where decision boundaries become more intricate and overlapping. The removal of the semi-indoor class in the binary classification task simplifies the problem, enabling the model to achieve higher overall performance. Thus, the binary classification model demonstrates superior performance, while the multi-class model struggles due to the added complexity and ambiguity introduced by the semi-indoor category.

5.2.3 Best Model

When comparing the performance of the models with and without the semi-indoor category, it is evident that the binary classification model (without semi-indoor) outperforms the multi-

class model (with semi-indoor) in terms of overall accuracy, precision, recall, and F1-score. The binary classification model achieves an accuracy of 0.91 and demonstrates strong, balanced performance across both Indoor and Outdoor classes, with F1-scores of 0.93 and 0.88, respectively. In contrast, the multi-class model achieves a lower accuracy of 0.78, with the semi-indoor class being the primary source of difficulty, as reflected in its lower F1-score of 0.72. This performance gap is largely due to the increased complexity and ambiguity introduced by the semi-indoor category, which shares characteristics with both Indoor and Outdoor environments, making it harder for the model to establish clear decision boundaries.

Overall, the binary classification strategy (without semi-indoor) is more effective for this task, as it simplifies the problem and allows the model to focus on distinguishing between two more distinct categories. While the multi-class model provides the ability to classify semi-indoor instances, its lower performance and the inherent ambiguity of the semi-indoor class make it less reliable. Therefore further efforts to improve the multi-class model—such as refining the definition of the semi-indoor class, collecting more representative data, or exploring more advanced modeling techniques—may be necessary to enhance its performance.

5.3 The Influencing Factor of Model Performance

5.3.1 The Relationship between Time Window Size on Model Training

The results of the one-way ANOVA test, as shown in Table 21, indicate that the choice of lag range does not have a statistically significant impact on model performance metrics, including precision, recall, and F1-score. The F-statistic values for all three metrics are extremely low (0.0435 for precision, 0.027 for recall, and 0.0141 for F1-score), suggesting minimal variance between different lag ranges. Additionally, the p-values are substantially higher than the conventional significance threshold of 0.05 (0.9988, 0.9996, and 0.9999, respectively), confirming that the observed differences are not statistically significant. Therefore, it can be concluded that variations in lag range do not meaningfully affect the model's classification performance.

5.3.2 The Relationship between Participant and Model Training

From Table 23, The overall performance of the model shows relatively strong precision (0.79) and recall (0.76), with an F1-score of 0.74, indicating a balanced capability to minimize both false positives and false negatives. However, the average test set accuracy (0.68) is lower than the fold average accuracy from cross-validation (0.76), suggesting potential performance variability across different test sets. Participant-wise analysis reveals significant differences: Participant 7 achieves the highest accuracy (0.97) with strong precision and recall, while Participants 4 and 2 also perform well with high F1-scores (0.82 and 0.81, respectively). Moderate performance is observed for Participants 1, 6, 11, and 12, with accuracy between 0.71 and 0.82. In contrast, Participants 5 and 10 exhibit low accuracy (0.35 and 0.40) and F1-scores (0.66 and 0.62), with low recall indicating difficulty in identifying true positives despite relatively high precision. Participant 8 shows high precision and recall (0.84 each) but low accuracy (0.55), suggesting possible class imbalance. Overall, precision tends to be higher than recall, reflecting a trade-off favoring fewer false positives, while significant variability across participants highlights potential sensitivity to individual differences and data distribution shifts.

The one-way ANOVA results in Table 27 indicate that differences in precision, recall, and F1-score across participants are not statistically significant, with p-values of 0.51, 0.70, and 0.74, respectively. This suggests that performance variability among participants may result from random chance rather than systematic differences. However, the notably low accuracy and recall for participants like 5 and 10 highlight the need for further investigation. The lack of statistical significance may be due to small sample sizes or high variance. While the model shows balanced overall performance, targeted data analysis or personalized modeling could help address variability and enhance robustness.

5.3.3 The Relationship between Scenario and Model Training

In Table 33, the model's performance varies significantly across different scenarios, with an average accuracy of 0.68 and an F1-score of 0.74, indicating a reasonable balance between precision (0.79) and recall (0.76). The top-performing scenarios include City Center Scenario 2 (accuracy: 0.89, F1-score: 0.83) and City Center Scenario 1 (accuracy: 0.80, F1-score: 0.84), where the model demonstrates strong and consistent performance. Moderate performance is observed in scenarios like Train Station Scenario (accuracy: 0.77, F1-score: 0.81) and Tram 7

Scenario (accuracy: 0.65, F1-score: 0.83), where precision remains high but accuracy is lower. In contrast, Irchel Campus Scenario 2 (accuracy: 0.59, F1-score: 0.53) and Central Campus Scenario (accuracy: 0.42, F1-score: 0.71) show notably lower performance, with reduced recall indicating challenges in identifying true positives. The Single Family Scenario (accuracy: 0.50, F1-score: 0.85) exhibits high precision and recall but low accuracy, suggesting potential class imbalance.

The one-way ANOVA results for different scenarios indicate statistically significant differences in precision, recall, and F1-score, with F-statistics of 4.47, 5.46, and 5.20, respectively, and p-values all below 0.01. These results suggest that the model's performance varies significantly across scenarios, likely due to differences in environmental conditions, data collection settings, or scenario-specific characteristics. Unlike the participant-wise analysis, where variability was not statistically significant, these findings highlight the influence of scenario-specific factors on model performance. Further analysis is needed to identify the key factors contributing to these differences and to improve the model's robustness across diverse scenarios.

5.3.4 Feature Importance

In Table 22, Table 28 and Table 34, the analysis of feature importance across different participants and scenarios reveals several key insights into the factors driving model predictions. Features such as `nrSat_lag1` and `nrSat_lag2` consistently appear as important across multiple participants and scenarios, indicating that satellite-related data, likely linked to positioning accuracy or environmental conditions, plays a critical role in the model's performance. Additionally, `speedAccuracy(m/s)_lag3` is frequently highlighted, suggesting that historical speed accuracy data significantly influences predictions.

Scenario-specific patterns also emerge. For instance, in the City Center Scenario, acceleration-related features like `accX(mg)_mean_lag2` and `accX(mg)_min` are prominent, likely reflecting the complex traffic dynamics and frequent acceleration changes typical of urban environments. In contrast, the Irchel Campus Scenario emphasizes latitude-related features such as `latitude_lag-3` and `latitude_lag1`, possibly due to the specific geographic layout of the campus. The Central Campus Scenario shows a strong reliance on meteorological features like `pressure(Pa)_lag-1` and `humidity(percentage)`, indicating that local microclimate conditions significantly impact predictions in this area.

Participant-specific variations are also evident. Some participants show higher sensitivity to features like `pressure(Pa)_lag-1` and `accX(mg)_std`, while others prioritize `humidity(percentage)_lag3` or `temperature(C)_lag5`. These differences may reflect variations in individual behavior patterns or localized environmental conditions.

Overall, the analysis underscores the importance of considering both universal and context-specific features in model design. Features related to satellite data and speed accuracy are universally important, while acceleration, geographic, and meteorological features gain prominence in specific scenarios. To optimize model performance, it is advisable to prioritize universally important features like `nrSat_lag1` and `nrSat_lag2`, while also incorporating scenario-specific features such as acceleration data for urban settings or latitude and weather-related features for campus environments. This tailored approach can enhance the model's accuracy and generalizability across different contexts.

5.4 Prediction Visualization

This section presents the visualization of prediction results to facilitate a detailed comparison between the predicted labels and the actual labels. Figures 17 to 22 illustrate the prediction outcomes for six selected scenarios, each representing different levels of model performance, ranging from high precision to lower precision. These scenarios were chosen from the dataset trained with time series data to highlight how the model performs under various conditions.

In the visualizations, matching circle colors indicate correct predictions, while discrepancies in colors signify misclassifications. The use of map-based visualization provides a clear spatial context, allowing for an intuitive understanding of where the model performs well and where frequent misclassifications occur. This approach helps identify patterns in prediction accuracy, such as areas where environmental transitions (for example, between indoor and semi-indoor spaces) might lead to confusion. By analyzing these visualizations, it becomes easier to discuss specific locations or environmental conditions that challenge the model, offering insights into potential areas for further improvement.

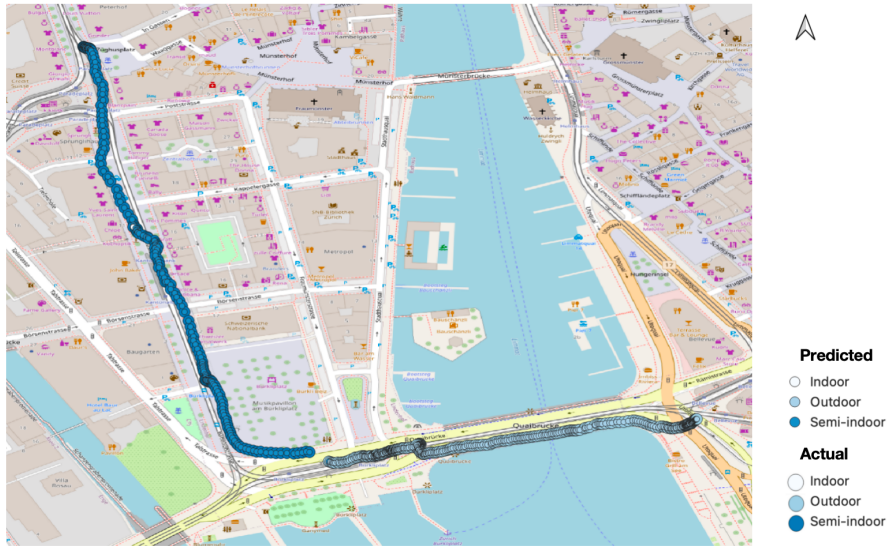


Figure 17 Map visualization example 1

For this City Center Scenario 2 (Figure 17), the model achieved an accuracy of 0.9941, with a weighted average precision of 1.00, recall of 0.99, and an F1-score of 1.00, based on 506 samples in the test set. This indicates excellent overall performance, with the model accurately distinguishing between different environmental contexts. Notably, the model performed exceptionally well in predicting rapid transitions between semi-indoor and outdoor environments, such as during boarding and alighting from trams.

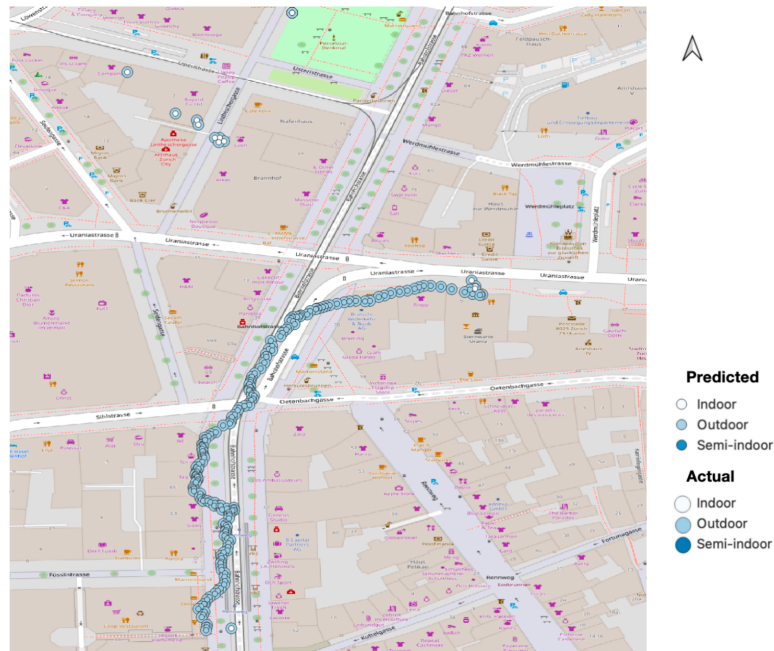


Figure 18 Map visualization example 2

5. Discussion

For this City Center Scenario 1 (Figure 18), the model achieved an accuracy of 0.9241, with a weighted average precision of 0.93, recall of 0.92, and an F1-score of 0.92, based on 514 samples in the test set. This reflects strong overall performance in classifying different environments. However, there were instances of delayed predictions during transitions between indoor and outdoor environments, indicating that the model may require additional adjustments to improve responsiveness in rapidly changing contexts.



Figure 19 Map visualization example 3

For this Tram 7 Scenario (Figure 19), the model achieved an accuracy of 0.9810, with a weighted average precision, recall, and F1-score all at 0.98, based on 473 samples in the test set. This indicates excellent overall performance, with the model demonstrating a high level of accuracy in predicting environmental transitions. Notably, the scene transitions between indoor, outdoor, and semi-indoor environments were predicted with high precision here.

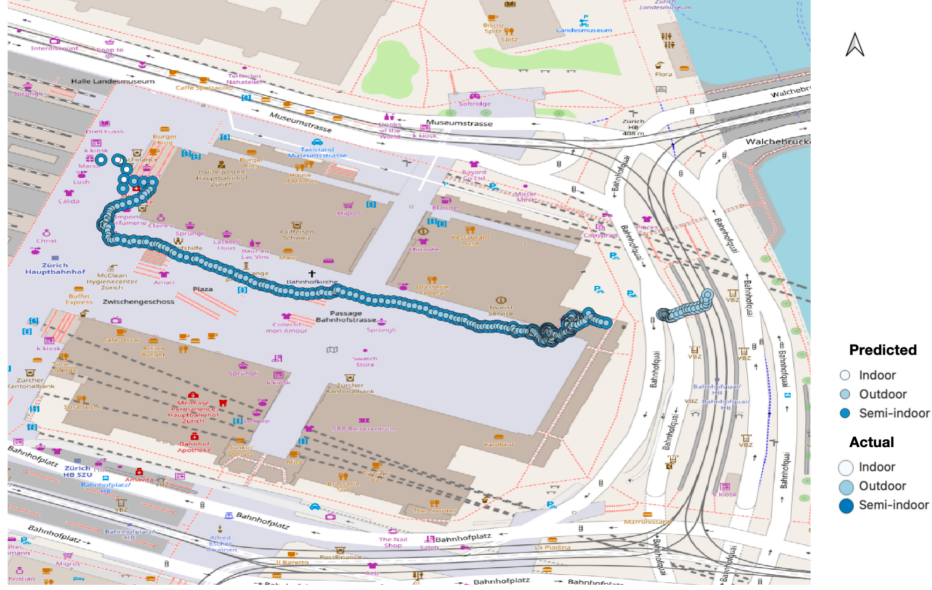


Figure 20 Map visualization example 4

For the Train Station Scenario (Figure 20), the model achieved a low accuracy of 0.1381, with a weighted average precision of 0.05, recall of 0.14, and an F1-score of 0.07, based on 391 samples in the test set. This indicates poor overall performance, particularly in accurately classifying environmental transitions. The model struggled significantly with detecting transitions from outdoor to indoor environments, as these changes were often not correctly predicted.

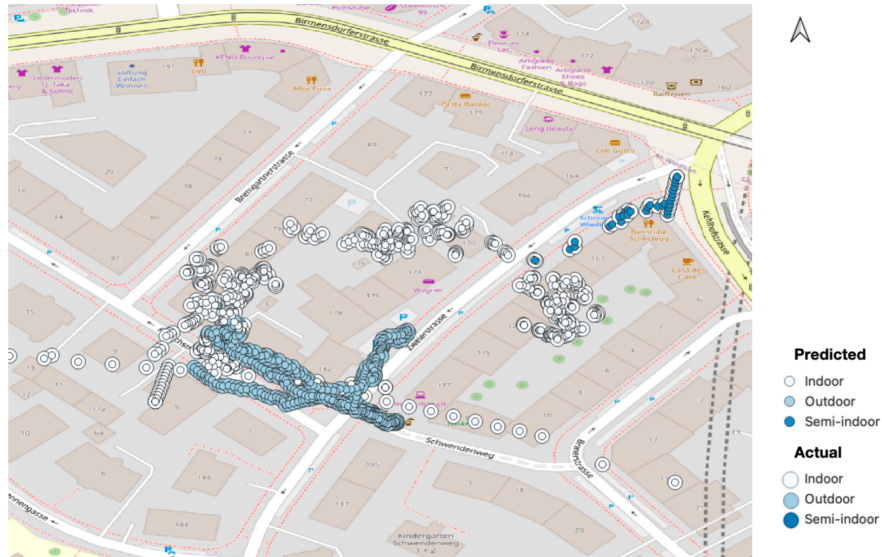


Figure 21 Map visualization example 5

For this Free Scenario (Figure 21), the model achieved an accuracy of 0.7421, with a weighted average precision of 1.00, recall of 0.74, and an F1-score of 0.85, based on 1683 samples in

the test set. While the overall performance is relatively good, the model exhibited notable misclassifications, particularly with semi-indoor environments being predicted as indoor. This suggests that the model struggles to differentiate between these two categories.

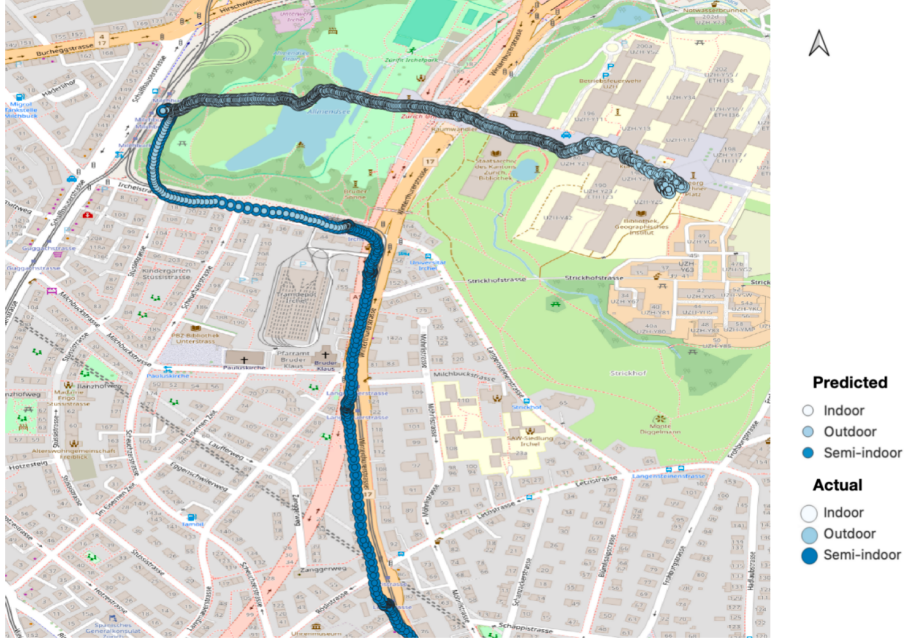


Figure 22 Map visualization example 6

For this Free Scenario (Figure 22), the model achieved an accuracy of 0.9173, with a weighted average precision of 0.93, recall of 0.92, and an F1-score of 0.92, based on 1379 samples in the test set. While the overall performance is strong, there were noticeable prediction errors during transitions between outdoor and semi-indoor environments, which represent the primary source of misclassification in this scenario. This suggests that the model faces challenges in distinguishing between these two environments, likely due to the subtle differences in environmental features during rapid transitions.

5.5 Comparison with Related Research

Several studies have explored similar topics, utilizing GPS and IMU data to classify environmental contexts or human activities. While these studies share common ground in terms of data sources, there are notable differences in data collection methods, preprocessing techniques, and model selection and training strategies. These variations significantly influence the final classification performance. Differences in sensor placement, sampling rates, and the

diversity of collected scenarios affect data quality and feature extraction. Moreover, the choice of machine learning models introduces variability in how patterns are learned and generalized. As a result, these methodological distinctions contribute to differences in classification accuracy and robustness, making it essential to contextualize the findings of this thesis within the broader landscape of related research.

Comparing this thesis with the paper (Bui et al., 2020) the key differences lie in the methodology, feature selection, and resulting performance. This thesis utilizes mixed data from various scenarios, employing a Random Forest model for both multi-class classification (indoor, outdoor, semi-indoor) and binary classification (indoor, outdoor). The multi-class classification achieves an accuracy of 78%, with the semi-indoor category presenting challenges due to its ambiguous nature, resulting in a lower F1-score of 0.72. In contrast, the binary classification achieves a higher accuracy of 91%, indicating that removing the semi-indoor category simplifies the task and improves overall performance.

The referenced paper focuses exclusively on binary classification, using GPS signal strength and the number of visible satellites as primary features, achieving a significantly higher accuracy of 97%. This superior performance can be attributed to the simplicity of the classification task, the effective use of GPS-based features, and a robust preprocessing pipeline that filters and prioritizes strong GPS signals. Additionally, the referenced paper evaluates multiple machine learning models, with K-Nearest Neighbors (KNN) emerging as the best performer due to its effectiveness in handling structured GPS data.

The performance gap between the two studies can be largely explained by the complexity introduced by the semi-indoor category in this thesis, which increases classification ambiguity and reduces accuracy. In contrast, the binary classification approach in the referenced paper benefits from optimized preprocessing and clearer decision boundaries, leading to higher accuracy. Overall, while the referenced paper's streamlined methodology and focus on GPS-specific features yield superior performance, this thesis offers a more nuanced classification framework, addressing complex environmental contexts at the cost of reduced accuracy due to the inclusion of the semi-indoor category.

This thesis is compared with the paper (Zhu et al., 2019) which also focuses on indoor/outdoor (IO) classification using GNSS measurement data. The key differences between the two studies lie in data collection strategies, feature extraction methods, and model complexity, all of which contribute to variations in classification performance.

In terms of data collection, the referenced paper incorporates a wide range of complex environments, including campuses, shopping malls, restaurants, office buildings, pedestrian streets, overpasses, and residential areas. It utilizes data from multiple Android smartphones, enhancing data diversity and generalizability. In contrast, this thesis focuses on specific scenarios such as urban centers, campuses, and transportation hubs, with potentially less environmental variability and fewer device types. This difference in data diversity may limit the model’s robustness when applied to unseen environments.

Regarding methodology, the referenced paper employs a comprehensive feature extraction process, incorporating spatial geometric distribution, temporal sequence features, and statistical characteristics derived from GNSS data. It further enhances performance through the use of sliding windows of varying sizes to capture temporal dynamics effectively. Additionally, the paper integrates multiple machine learning models, including Random Forest, SVM, AdaBoost, XGBoost, and LightGBM, combined with a stacked ensemble approach and Hidden Markov Models (HMM) for state inference. In contrast, this thesis relies primarily on Random Forest without ensemble techniques or temporal smoothing via HMM, which may reduce classification robustness, particularly during transition states.

The performance gap can be attributed to several factors. The referenced paper’s use of diverse data sources, sophisticated feature engineering, and complex model ensembles enhances both accuracy and generalizability. In comparison, this thesis’s more focused dataset and simpler model architecture may limit its adaptability to new environments.

5.6 Limitation

- **Hardware Limitations of the “uTrail” Device:** the “uTrail” device used for data collection lacks Assisted GPS (A-GPS) functionality, which typically enhances signal acquisition, especially in indoor environments. As a result, when transitioning indoors, the GPS signal becomes unstable or even completely obstructed, leading to gaps or inaccuracies in location data. This limitation affects the precision of spatial features derived from GPS, which could impact the overall performance of the classification model.
- **Model Simplicity:** this thesis employs the Random Forest algorithm for classification, which, while robust and interpretable, is relatively simplistic compared to state-of-the-

art machine learning models such as gradient boosting methods or deep learning architectures. The simplicity of the model may limit its ability to capture complex patterns in the data, potentially contributing to lower classification accuracy compared to more advanced models.

- **Data Generalizability and Participant Bias:** although the study is designed to support the objectives of the MOASIS project, the collected dataset consists primarily of samples from young participants. This demographic bias raises concerns about the model's generalizability to MOASIS data, which may involve a more diverse age range. As a result, the model's performance on MOASIS data remains untested and may lack robustness when applied to different populations.
- **Ambiguity in Scenario Instructions and Labeling Consistency:** the scenario instructions provided to participants may have been insufficiently clear, leading to inconsistent interpretations, particularly regarding the definition of semi-indoor environments. Since participants may have had varying understandings of semi-indoor spaces, the labeling process might lack precision, introducing label noise that could degrade model performance.
- **Limited Focus on Spatial Features in Feature Engineering:** during the feature engineering phase, the focus was primarily on temporal features, while spatial characteristics were relatively overlooked. For example, incorporating features such as movement speed derived from GPS coordinates or analyzing spatial patterns could potentially improve classification accuracy. The underutilization of spatial data may have limited the model's ability to fully capture environmental transitions, affecting its predictive capabilities.

6. Conclusion

This thesis developed a Random Forest-based environmental classification model to investigate how different spatial contexts—indoor, outdoor, and semi-indoor environments—can be accurately identified using sensor and GPS data. The model was trained on data collected from diverse real-world scenarios, capturing participants’ activities in various environments to reflect realistic mobility patterns. This study not only focused on distinguishing between traditional indoor and outdoor spaces but also introduced the semi-indoor category, addressing environments that exhibit mixed characteristics, such as covered walkways, transit stations, and partially enclosed public spaces.

The model integrates temporal and spatial features to enhance classification performance, with a particular focus on the temporal dynamics of sensor data. Additionally, the classification results were evaluated using multiple performance metrics, with a specific emphasis on the F1-score to address the challenges posed by imbalanced data. Beyond model development, this thesis contributes to the MOASIS project by applying the model to the MOASIS dataset, using visualization techniques to assess predictive performance in the absence of labeled ground truth. This approach not only demonstrates the model’s generalization capabilities but also provides a foundation for further refinement based on real-world applications.

Building on these outcomes, the following sections address the key research questions that guided this study.

- How can GPS and IMU data be optimally combined to improve the accuracy of indoor-outdoor detection?

The optimal combination of GPS and IMU data for indoor-outdoor detection relies on leveraging the complementary strengths of both data sources. GPS data provides valuable spatial and signal-related information, such as the number of visible satellites and signal accuracy, which are strong indicators of environmental conditions. IMU data, on the other hand, captures dynamic motion patterns, including acceleration, orientation, and pressure changes, which help identify subtle transitions between different environments that GPS alone may miss.

This combination allows the model to detect environmental transitions more robustly, even in scenarios where GPS signals are weak or inconsistent, such as semi-indoor areas. The feature engineering process further enhances performance by applying time window techniques, which help capture temporal dependencies and improve the model's ability to recognize patterns over time.

- What specific features of GPS and IMU data are most effective in distinguishing between indoor, outdoor, and semi-indoor environments?

The analysis of feature importance reveals several GPS and IMU features that are particularly effective in distinguishing between indoor, outdoor, and semi-indoor environments. For GPS data, `nrSat_lag1` and `nrSat_lag2`—representing the number of visible satellites—consistently emerge as critical features across various scenarios. These features reflect signal availability, with lower satellite counts typically indicating indoor conditions due to signal obstruction. Additionally, `speedAccuracy(m/s)_lag3`, which measures historical speed accuracy, plays an important role in identifying areas with fluctuating GPS reliability, often associated with semi-indoor spaces.

For IMU data, acceleration-related features such as `accX(mg)_mean_lag2` and `accX(mg)_min` are highly effective, particularly in complex urban environments. These features capture sudden changes in movement patterns, such as stops and starts when transitioning between indoor and outdoor spaces. In specific scenarios like campus environments, latitude-related features (e.g., `latitude_lag-3`) become prominent due to the geographic layout affecting movement dynamics. Additionally, meteorological features such as `pressure(Pa)_lag-1` and `humidity(percentage)` contribute to distinguishing between environments, as they reflect microclimate variations often present in semi-indoor spaces.

- Do individual, scenario, and “semi-indoor” definition affect model performance?

Impact of Individual Differences: The participant-wise analysis reveals noticeable variability in model performance across different individuals. For example, Participant 7 achieved the highest accuracy (0.97) with strong precision and recall, while Participants 5 and 10 showed much lower accuracy (0.35 and 0.40) and F1-scores (0.66 and 0.62), indicating difficulties in identifying true positives despite relatively high precision. However, the one-way ANOVA results show that these differences are not statistically significant (p-values: 0.51 for precision, 0.70 for recall, and 0.74 for F1-score). This suggests that the observed variability may result

from random fluctuations or differences in sample sizes rather than systematic individual-specific effects. Nonetheless, the poor performance in certain participants highlights the potential influence of data distribution shifts or individual behavioral patterns, which warrants further investigation.

Impact of Scenario Variations: In contrast to individual differences, scenario-specific factors have a statistically significant impact on model performance. The model performs strongly in scenarios like City Center Scenario 2 (accuracy: 0.89, F1-score: 0.83) and City Center Scenario 1 (accuracy: 0.80, F1-score: 0.84), while struggling in others such as Irchel Campus Scenario 2 (accuracy: 0.59, F1-score: 0.53) and Central Campus Scenario (accuracy: 0.42, F1-score: 0.71). The one-way ANOVA test confirms these differences are statistically significant (p -values < 0.01 for precision, recall, and F1-score), suggesting that environmental conditions, data collection contexts, and scenario-specific characteristics—such as GPS signal stability, physical obstructions, or crowd density—substantially influence model performance. This indicates the need for scenario-specific model adjustments or feature engineering to enhance robustness across diverse environments.

Impact of the ‘Semi-Indoor’ definition: While not directly analyzed with statistical tests, the definition and classification of semi-indoor environments appear to be a critical factor affecting model performance. Semi-indoor spaces, which exhibit characteristics of both indoor and outdoor environments, introduce ambiguity that challenges the model’s ability to clearly differentiate between classes. Misclassifications often occur during transitions between outdoor and semi-indoor areas, as observed in specific scenarios like the Train Station Scenario, where transitions were not correctly predicted. The complexity in defining semi-indoor spaces may lead to inconsistent labeling and difficulties in feature representation, impacting both precision and recall in these cases.

For the future work, to address the instability of GPS signals in indoor environments, future work could explore the use of devices equipped with Assisted GPS (A-GPS) or hybrid positioning systems that combine GPS with Wi-Fi, Bluetooth beacons, or inertial sensors. These technologies can significantly improve signal acquisition indoors, reducing data gaps and enhancing the accuracy of spatial features, especially in challenging environments such as underground spaces or densely built-up areas. Improved positioning technologies would contribute to more reliable data collection, ultimately enhancing the robustness of environmental classification models.

While Random Forest provided a solid baseline for classification, future studies could investigate the application of more sophisticated algorithms. Techniques such as Gradient Boosting Machines (e.g., XGBoost, LightGBM), Support Vector Machines (SVMs), or deep learning models like Recurrent Neural Networks (RNNs) and Convolutional Neural Networks (CNNs) could potentially capture more complex patterns in spatiotemporal data. These models are capable of learning intricate relationships and dynamic temporal dependencies, which may lead to improved classification accuracy. Additionally, exploring ensemble learning methods that combine different models might further enhance performance by leveraging the strengths of multiple algorithms.

To improve the model's robustness and ensure its applicability to the MOASIS project, future data collection efforts should focus on gathering more data from older adults, as this demographic is underrepresented in the current dataset. Expanding the participant pool to include a broader age range will help the model better capture the variability in activity patterns and environmental interactions associated with aging. Additionally, the model should be applied directly to the MOASIS dataset to evaluate its performance in real-world conditions. Although the MOASIS data lacks labeled ground truth, visualization techniques can be employed to assess the model's predictions, identifying patterns and potential misclassifications. This process will not only provide insights into the model's generalization capabilities but also serve as a foundation for iterative model refinement based on observed discrepancies, ultimately enhancing its performance for MOASIS-related applications.

Future work should focus on refining the scenario instructions to minimize ambiguities, especially in the classification of semi-indoor environments. Developing a standardized labeling protocol, supported by visual guidelines or interactive training sessions, can help ensure more consistent interpretations among participants. Additionally, implementing real-time validation mechanisms in data collection apps could prompt participants to verify or correct their labels during the data collection process, thereby reducing label noise and improving the reliability of the ground truth data.

To improve model performance, future studies should expand the feature engineering process to incorporate spatial dynamics alongside temporal data. Features such as movement speed, acceleration patterns, changes in direction, and GPS signal strength variations could provide richer contextual information. For example, calculating instantaneous speed from GPS coordinates or identifying stop-and-go movement patterns could help differentiate between indoor, outdoor, and semi-indoor environments. Additionally, leveraging spatial clustering

6. Conclusion

algorithms to detect patterns in mobility behavior could further enhance classification accuracy by capturing subtle spatial transitions.

Reference

Ahmad, N., Ghazilla, R. A. R., Khairi, N. M., & Kasi, V. (2013). Reviews on Various Inertial Measurement Unit (IMU) Sensor Applications. *International Journal of Signal Processing Systems*, 256–262. <https://doi.org/10.12720/ijspss.1.2.256-262>

Akail, N., Moralesl, L. Y., & Murase, H. (2018). Reliability Estimation of Vehicle Localization Result. *2018 IEEE Intelligent Vehicles Symposium (IV)*, 740–747. <https://doi.org/10.1109/IVS.2018.8500625>

Allahbakhshi, H., Conrow, L., Naimi, B., & Weibel, R. (2020). Using Accelerometer and GPS Data for Real-Life Physical Activity Type Detection. *Sensors*, 20(3), 588. <https://doi.org/10.3390/s20030588>

Ashraf, I., Hur, S., Park, S., & Park, Y. (2020). DeepLocate: Smartphone Based Indoor Localization with a Deep Neural Network Ensemble Classifier. *Sensors*, 20(1), Article 1. <https://doi.org/10.3390/s20010133>

Ashraf, I., Hur, S., & Park, Y. (2018). MagIO: Magnetic Field Strength Based Indoor-Outdoor Detection with a Commercial Smartphone. *Micromachines*, 9(10), Article 10. <https://doi.org/10.3390/mi9100534>

Belhajem, I., Maissa, Y. B., & Tamtaoui, A. (2018). Improving Vehicle Localization in a Smart City with Low Cost Sensor Networks and Support Vector Machines. *Mobile Networks and Applications*, 23(4), 854–863. <https://doi.org/10.1007/s11036-017-0879-9>

Bui, V., Le, N. T., Vu, T. L., Nguyen, V. H., & Jang, Y. M. (2020). GPS-Based Indoor/Outdoor Detection Scheme Using Machine Learning Techniques. *Applied Sciences*, 10(2), 500. <https://doi.org/10.3390/app10020500>

Collin, J., Mezentsev, O., & Lachapelle, G. (2003). INDOOR POSITIONING SYSTEM USING ACCELEROMETRY AND HIGH ACCURACY HEADING SENSORS.

- Dedes, G., & Dempster, A. G. (2005). *Indoor GPS positioning—Challenges and opportunities*. *VTC-2005-Fall. 2005 IEEE 62nd Vehicular Technology Conference, 2005.*, 1, 412–415. <https://doi.org/10.1109/VETECF.2005.1557943>
- Esmaili Kelishomi, A., Garmabaki, A. H. S., Bahaghighat, M., & Dong, J. (2019). *Mobile User Indoor-Outdoor Detection through Physical Daily Activities*. *Sensors*, 19(3), Article 3. <https://doi.org/10.3390/s19030511>
- Hsu, L.-T., Tokura, H., Kubo, N., Gu, Y., & Kamijo, S. (2017). *Multiple Faulty GNSS Measurement Exclusion Based on Consistency Check in Urban Canyons*. *IEEE Sensors Journal*, 17(6), 1909–1917. <https://doi.org/10.1109/JSEN.2017.2654359>
- Ichino, H., Kaji, K., Sakurada, K., Hiroi, K., & Kawaguchi, N. (2016). *HASC-PAC2016: Large scale human pedestrian activity corpus and its baseline recognition*. *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, 705–714. <https://doi.org/10.1145/2968219.2968277>
- I-Hao, L., & Dong-Soo, H. (2024). *Optimizing GNSS Indoor Outdoor Detection: Balancing Observation Window and Sampling for Accuracy and Responsiveness*. *2024 14th International Conference on Indoor Positioning and Indoor Navigation (IPIN)*, 1–6. <https://doi.org/10.1109/IPIN62893.2024.10786149>
- Kahana, E., Han, C., Kahana, B., Bhatta, T., Yu, J., & Iqbal, A. (n.d.). *MAINTAINING LIFE SPACE MOBILITY AMONG ELDERLY MIGRANTS TO THE SUNBELT: DEMOGRAPHIC, HEALTH AND SOCIAL ANTECEDENTS*.
- Kumar, A., Kumar, S., Lal, P., Saikia, P., Srivastava, P. K., & Petropoulos, G. P. (2021). *Introduction to GPS/GNSS technology*. In *GPS and GNSS Technology in Geosciences* (pp. 3–20). Elsevier. <https://doi.org/10.1016/B978-0-12-818617-6.00001-9>
- Lachapelle, G. (2004). *GNSS Indoor Location Technologies*. *Positioning*, 01(08), Article 08. <http://www.scirp.org/journal/PaperInformation.aspx?PaperID=237&#abstract>
- Li, D., Lei, Y., Li, X., & Zhang, H. (2020). *Deep Learning for Fingerprint Localization in Indoor and Outdoor Environments*. *ISPRS International Journal of Geo-Information*, 9(4), Article 4. <https://doi.org/10.3390/ijgi9040267>
- Liao, M., Guo, Y., Qin, Y., & Wang, Y. (2015). *The application of EMD in activity recognition based on a single triaxial accelerometer*. *Bio-Medical Materials and Engineering*, 26(s1), S1533–S1539. <https://doi.org/10.3233/BME-151452>

- Mallik, M., Panja, A. K., & Chowdhury, C. (2023). *Paving the way with machine learning for seamless indoor–outdoor positioning: A survey*. *Information Fusion*, 94, 126–151.
<https://doi.org/10.1016/j.inffus.2023.01.023>
- Martin, P., Kelly, N., Kahana, B., Kahana, E., Willcox, B. J., Willcox, D. C., & Poon, L. W. (2015a). *Defining successful aging: A tangible or elusive concept?* *The Gerontologist*, 55(1), 14–25. <https://doi.org/10.1093/geront/gnu044>
- Martin, P., Kelly, N., Kahana, B., Kahana, E., Willcox, B. J., Willcox, D. C., & Poon, L. W. (2015b). *Defining Successful Aging: A Tangible or Elusive Concept?* *The Gerontologist*, 55(1), 14–25. <https://doi.org/10.1093/geront/gnu044>
- Michel, J.-P., & Sadana, R. (2017). “Healthy Aging” Concepts and Measures. *Journal of the American Medical Directors Association*, 18(6), 460–464.
<https://doi.org/10.1016/j.jamda.2017.03.008>
- Montanha, A., Polidorio, A., & Romero-Ternero, M. (2021). *New signal location method based on signal-range data for proximity tracing tools*. *Journal of Network and Computer Applications*, 180, 103006. <https://doi.org/10.1016/j.jnca.2021.103006>
- Niculescu-Mizil, A., Perlich, C., Swirszcz, G., Sindhvani, V., Liu, Y., Melville, P., Wang, D., Xiao, J., Hu, J., Singh, M., Shang, W. X., & Zhu, Y. F. (2009). *Winning the KDD Cup Orange Challenge with Ensemble Selection*. *Proceedings of KDD-Cup 2009 Competition*, 23–34.
<https://proceedings.mlr.press/v7/niculescu09.html>
- Okamoto, M., & Chen, C. (2015). *Improving GPS-based indoor-outdoor detection with moving direction information from smartphone*. *Adjunct Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing and Proceedings of the 2015 ACM International Symposium on Wearable Computers*, 257–260.
<https://doi.org/10.1145/2800835.2800939>
- Özdemir, A. T., & Barshan, B. (2014). *Detecting Falls with Wearable Sensors Using Machine Learning Techniques*. *Sensors*, 14(6), Article 6. <https://doi.org/10.3390/s140610691>
- Passaro, V. M. N., Cuccovillo, A., Vaiani, L., De Carlo, M., & Campanella, C. E. (2017). *Gyroscope Technology and Applications: A Review in the Industrial Perspective*. *Sensors*, 17(10), 2284. <https://doi.org/10.3390/s17102284>

- Paving the way with machine learning for seamless indoor–outdoor positioning: : A survey: Information Fusion: Vol 94, No C. (n.d.). Retrieved March 19, 2024, from <https://dl.acm.org/doi/abs/10.1016/j.inffus.2023.01.023>*
- Puricer, P., & Kovar, P. (2007). Technical Limitations of GNSS Receivers in Indoor Positioning. 2007 17th International Conference Radioelektronika, 1–5. <https://doi.org/10.1109/RADIOELEK.2007.371487>*
- Puthussery, A. R., Haradi, K. P., Erol, B. A., Benavidez, P., Rad, P., & Jamshidi, M. (2017). A deep vision landmark framework for robot navigation. 2017 12th System of Systems Engineering Conference (SoSE), 1–6. <https://doi.org/10.1109/SYSOSE.2017.7994976>*
- Radu, V., Katsikouli, P., Sarkar, R., & Marina, M. K. (2014). A semi-supervised learning approach for robust indoor-outdoor detection with smartphones. Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems, 280–294. <https://doi.org/10.1145/2668332.2668347>*
- Rajak, S., Panja, A., Chowdhury, C., & Neogy, S. (2021). A Ubiquitous Indoor–Outdoor Detection and Localization Framework for Smartphone Users (pp. 693–701). https://doi.org/10.1007/978-981-15-9927-9_67*
- Röcke, C., Luo, M., Bereuter, P., Katana, M., Fillekes, M., Gehriger, V., Sofios, A., Martin, M., & Weibel, R. (2023). Charting everyday activities in later life: Study protocol of the mobility, activity, and social interactions study (MOASIS). *Frontiers in Psychology*, 13, 1011177. <https://doi.org/10.3389/fpsyg.2022.1011177>*
- Roecke, C., Katana, M., Fillekes, M., Martin, M., & Weibel, R. (2018). Mobility, physical activity and social interactions in the daily lives of healthy older adults: The MOASIS project. *Innovation in Aging*, 2(suppl_1), 274–274. <https://doi.org/10.1093/geroni/igy023.1014>*
- Rokach, L. (2010). Ensemble-based classifiers. *Artificial Intelligence Review*, 33(1), 1–39. <https://doi.org/10.1007/s10462-009-9124-7>*
- Radu, V., Katsikouli, P., Sarkar, R., & Marina, M. K. (2014, November). A semi-supervised learning approach for robust indoor-outdoor detection with smartphones. In *Proceedings of the 12th ACM Conference on Embedded Network Sensor Systems* (pp. 280-294).*

- Tabatabaei, A., & Mosavi, M.-R. (2019). *Performance analysis of GLONASS integration with GPS vectorised receiver in urban canyon positioning*. *Survey Review*, 51(368), 460–471. <https://doi.org/10.1080/00396265.2018.1481181>
- Tinder, R. F. (2007). *Relativistic flight mechanics and space travel: A primer for students, engineers and scientists* (p. 33). Morgan & Claypool Publishers. ISBN 978-1-59829-130-8.
- Van Diggelen, F. (2002). *Indoor GPS theory & implementation*. 2002 IEEE Position Location and Navigation Symposium (IEEE Cat. No.02CH37284), 240–247. <https://doi.org/10.1109/PLANS.2002.998914>
- Wikipedia contributors. (2025, January 3). *Inertial measurement unit*. Wikipedia. Retrieved January 3, 2024, from https://en.wikipedia.org/wiki/Inertial_measurement_unit
- Rindler, W. (2013). *Essential relativity: Special, general, and cosmological* (illustrated ed., p. 61). Springer. ISBN 978-1-4757-1135-6.
- Wikipedia contributors. (2025, January 3). *Gyroscope*. Wikipedia. Retrieved February 3, 2025, from <https://en.wikipedia.org/wiki/Gyroscope>
- Wikipedia contributors. (2025, January 3). *Gyroscope*. Wikipedia. Retrieved February 3, 2025, from <https://en.wikipedia.org/wiki/Accelerometer>
- Wikipedia contributors. (2025, January 3). *Gyroscope*. Wikipedia. Retrieved February 3, 2025, from <https://en.wikipedia.org/wiki/Magnetometer>
- Xu, R., Chen, W., Xu, Y., & Ji, S. (2015). *A New Indoor Positioning System Architecture Using GPS Signals*. *Sensors*, 15(5), 10074–10087. <https://doi.org/10.3390/s150510074>
- Yan, J., Diakit , A. A., & Zlatanova, S. (2019). *A generic space definition framework to support seamless indoor/outdoor navigation systems*. *Transactions in GIS*, 23(6), 1273–1295. <https://doi.org/10.1111/tgis.12574>
- Zeng, Q., Wang, J., Meng, Q., Zhang, X., & Zeng, S. (2018). *Seamless Pedestrian Navigation Methodology Optimized for Indoor/Outdoor Detection*. *IEEE Sensors Journal*, 18(1), 363–374. *IEEE Sensors Journal*. <https://doi.org/10.1109/JSEN.2017.2764509>
- Zhou, P., Zheng, Y., Li, Z., Li, M., & Shen, G. (n.d.). *IODetector: A generic service for indoor outdoor detection*.
- Zhou, T., Lian, B., Yang, S., Zhang, Y., & Liu, Y. (2018a). *Improved GNSS Cooperation Positioning Algorithm for Indoor Localization*.

Zhou, T., Lian, B., Yang, S., Zhang, Y., & Liu, Y. (2018b, December 15). *Improved GNSS Cooperation Positioning Algorithm for Indoor Localization*. | *Computers, Materials & Continua* | EBSCOhost. <https://doi.org/10.3970/cmc.2018.02671>

Zhu, Y., Luo, H., Wang, Q., Zhao, F., Ning, B., Ke, Q., & Zhang, C. (2019). *A Fast Indoor/Outdoor Transition Detection Algorithm Based on Machine Learning*. *Sensors (Basel, Switzerland)*, 19(4), 786. <https://doi.org/10.3390/s19040786>

Appendix

Appendix A

Irchel Campus Scenario 1

1. Start outdoors in the courtyard (outdoor)
2. Enter the teaching building and walk into a room staying for >8 minutes (indoor)
3. Sit by the window for 5-8 minutes (semi-indoor)
4. Walk to Medizinerhof between buildings and stay for 5-8 minutes (semi-indoor)
5. Walk to the library staying for 5-8 minutes (indoor)
6. Walk in the Irchel Park for 5-8 minutes (outdoor)

Train Station Scenario

1. Depart from the central train station (outdoor)
2. Walk through Bahnhofstrasse to Globus (outdoor)
3. Browse at least four stores in the mall staying 5-8 minutes in each (indoor)
4. Stay near the windows for 5-8 minutes before leaving. (semi-indoor)
5. Leave Globus and go to Coop's supermarket floor staying 5-8 minutes

Main Campus Scenario

1. Start at main campus
2. Stay at the Lichthof for > 8 minutes (semi-indoor)
3. Enter a classroom and stay for > 8minutes (indoor)
4. Leave the building and walk to take Polyban (outdoor)
5. Take a walk in the alley (semi-indoor)
6. Take a walk along the Limmatquai (outdoor)

Garden Scenario

1. Start outside in the garden and stay there for 5-8 minutes (outdoor)
2. Stay inside the house > 8 minutes (indoor)

3. Stay under the eaves for 5-8 minutes (semi-indoor)
4. Walk to the community center and stay there for 5-8 minutes (outdoor, indoor)
- (optional) 5. Walk to a nearby gym and stay for 5-8 minutes (indoor)

Irchel Campus Scenario 2

1. Start at Irchel campus (outdoor)
2. Walk to Irchel bar and stay under the roof for 5-8 minutes (semi-indoor)
3. Walk into (outdoor) the cafeteria and stay for 5-8 minutes (indoor)
4. Walk to 'Lichthof' and sit there for 5-8 minutes (semi-indoor)
5. Get into your office room and stay for 5-8 minutes (indoor)

Central Scenario

1. Start at an open area on or near "Central" square (outdoor)
2. Stay under the station's glass roof for 5 minutes (semi-indoor)
3. Take tram to Bellevue (semi-indoor)
4. Get off the tram and walk to the train station Stadelhofen (outdoor)
5. Hop on a train to Zurich HB (indoor)
6. Get off the train and stay on the ground floor of Zurich HB (semi-indoor)

Tram 7 Scenario

1. Start from Milchbuck
2. Take tram 7 to Schwamendingerplatz (indoor)
3. Get off the tram and enter Coop (indoor)
4. Stay inside for 5-8 minutes
5. Enter Migros and stay inside for 5-8 minutes (indoor)
6. Leave Schwamendingerplatz by tram (semi-indoor)

City Center Shopping Scenario

1. Start at Zurich HB
2. Walk along Bahnhofstrasse towards Rennweg (outdoor)
3. Enter 2-3 shops on your way (indoor) (e.g., Lindt chocolate, clothing stores), stay 5-8 minutes in each shop
4. Board the tram at Rennweg till Bürkliplatz (semi-indoor)
5. Take a 5-minute walk along the lake (outdoor)

Appendix B

"Label Logger " Tutorial

This is a detailed introduction and tutorial of the "Label Logger" application.

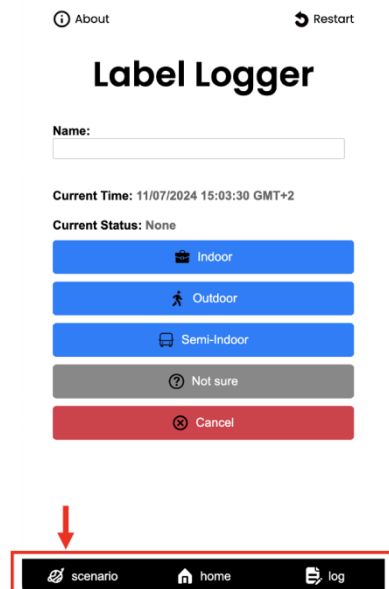
0. Start

To start, please fill in your name on the Home page, and look at the Scenario page to confirm your collection scenario. When you are ready, you could start with your first label choice.

1. Page Overview

There are three pages. You can switch between pages at the bottom.

- Home page: to record the label.
- Scenario page: to see the scenario and corresponding plans.
- Log page: to check your label records and edit them.



2. Home Page

2.1 Name

As stated in the "consent form," we anonymize the data. Therefore, you can use a code name when entering your name. However, please do not change the name during a single data collection process.

2.2 Label Log (end with "" stop!!)

The most important function of the homepage is to record labels.

There are three main labels:



- **Indoor:** Indicates being in a fully enclosed environment.
- **Outdoor:** Indicates being in an open area.
- **Semi-Indoor:** Indicates being in a semi-enclosed environment, such as a glasshouse, tram, bus, or sheltered outdoor garden.

When you are unsure which environment you are in, you could press "**Not sure**" and give a short description of your current environment.

If you change the environment, **it is important to switch the label in time**. To switch the label, **simply click on the new label**.

If you want to **stop** the current label recording, please click the current label again to stop it.

If you think you clicked the current label by mistake, please click the "**Cancel**" button to cancel it. Please note that once canceled, the previous records will be erased. Ensure to switch to the correct label promptly or add the record on the "Log" page.

2.3 Restart

The restart button is located in the upper right corner. Clicking it will **erase all records**. **Please click it when you need to begin a new data collection session.**

 About  Restart 

Label Logger

Name:

2.4 About

  About  Restart

Label Logger

Name:

An information page will pop up after clicking the "About" button. This page includes a brief introduction to the research and a concise tutorial for quick reference.

3. Scenario Page

The scenario page displays the data collection scenarios. You can follow the scenario plan during data collection.

Embark on Your Data Journey!

[Change Scenario](#)

[Create a Scenario](#)

Central Scenario

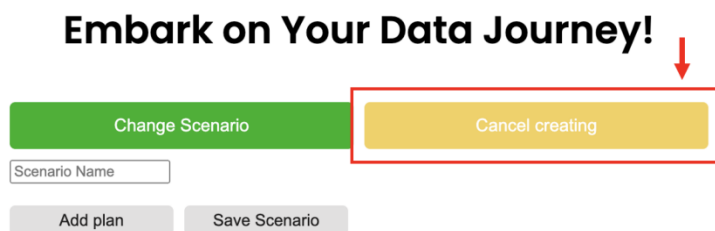
- ☐ 1.Start at a nearby open area
- ☐ 2.Stay under the station's glass roof for 5 minutes
- ☐ 3.Take tram 9 to Bellevue
- ☐ 4. Get off the tram and walk to train station Stadelhofen
- ☐ 5.Hop on a train to Zurich HB (sit on the lower floor)
- ☐ 6.Get off the train and stay on the ground floor of Zurich HB

Each data collection session can randomly switch to a new scenario when you click **“Change Scenario”**, this action can only be taken **once**.

You can create your scenario plan by clicking **“Create a Scenario”**. When creating your scenarios, please follow the design protocol below.

1. Your scenario should include indoors, outdoors, and semi-indoors environments.
2. Please stay for at least 8 minutes before moving to the next place.

Additionally, a valid scenario should have a scenario title and specific plans. You can also click **“Cancel creating”** button to quit.



The purpose is to ensure the diversity of data. The scenario plan page functions like a to-do list where you can click to mark tasks as completed or not completed. Excluding which scenario you follow, other actions will not be recorded.

4. Log page

4.1 Log Editing

The log page shows all the label records. The record has three background colors. The first record is always yellow, taken as an example record. The blue record is added by the “Add record” button. The white record is added by selecting the label on the “Home” page.

 Export

Log Table

Add record

Delete record

Name	Id	Label	Date	Time	TimeZone	Note
John Doe	1	indoor	15/04/2024	8:31:57	GMT+2	Note 1
Chenxi	1	indoor	12/07/2024	20:19:55	GMT+2	
Chenxi	3	semi-indoor	12/07/2024	20:19:59	GMT+2	
Chenxi	1	indoor	12/07/2024	20:20	GMT+2	
Chenxi	2	outdoor	12/07/2024	20:20:02	GMT+2	
Chenxi	4	not sure	12/07/2024	20:20:34	GMT+2	On the balcony
Chenxi	4	not sure	12/07/2024	20:20:39	GMT+2	Stop

You can add a record by clicking “Add record” button and after filling in the necessary information, click “Save”, then it will be saved to your label log. To add a valid record, the environment id, the date, and the time must be filled. For the “Not sure” label, you also need to describe the environment briefly in the note.

 Export

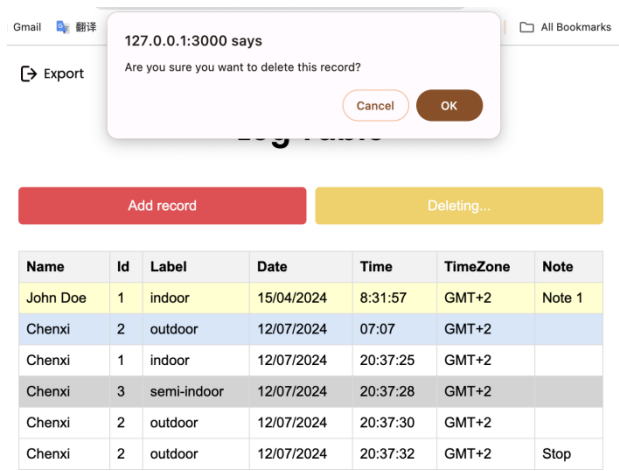
Log Table

Save

Delete record

Name	Id	Label	Date	Time	TimeZone	Note
John Doe	1	indoor	15/04/2024	8:31:57	GMT+2	Note 1
Chenxi	Select an option		2024/07/12	--:--	GMT+2	Note

If you find there is a wrong record, you can delete it by clicking the “Delete record” button. After clicking the record you want to delete, the background color of it will turn grey. If you confirm to delete this record, please choose “OK” in the pop-up window. To quit deleting mode, please click the “Deleting...” button.



4.2 Log Export

When you finish **each scenario's** data collection, please export your log by clicking the “**Export**” icon in the upper left corner. The csv file will automatically pop up, and please download it and **send it to me by email**.



Log Table



Appendix C

Participant	Scenario	Fold Average Accuracy	Test set accuracy	Test set precision	recall	f1
1	Free	0.59666667	0.645429363	0.667935336	0.59666667	0.561718778
1	Tram_7_Scenario	0.643086817	0.326923077	0.797808209	0.643086817	0.606968581
1	Tram_7_Scenario	0.916320475	0.994071146	0.95515614	0.916320475	0.93117993
1	Tram_7_Scenario	0.862207358	1	0.855085883	0.862207358	0.835032902
2	Central_Campus_Scenario	0.567656766	0.620614035	0.743365178	0.567656766	0.592279782
2	City_Center_Scenario_1	0.849122807	0.924124514	0.89126533	0.849122807	0.865852371
2	City_Center_Scenario_2	0.847863248	0.951704545	0.81017298	0.847863248	0.817726343
2	Train_Station_Scenario	0.851219512	0.770889488	0.828274983	0.851219512	0.811783005
2	Tram_7_Scenario	0.945045045	0.28742515	0.952819196	0.945045045	0.945541378
3	Free	0.662056738	0.419811321	0.654544439	0.662056738	0.635562305
3	Irchel_Campus_Scenario_1	0.699734748	0.996466431	0.767326572	0.699734748	0.693717265
3	Irchel_Campus_Scenario_1	0.776208178	0.997524752	0.78032323	0.776208178	0.754640096
4	City_Center_Scenario_2	0.836434109	0.765463918	0.807080538	0.836434109	0.812496633
4	Free	0.954545455	0.693467337	0.98951049	0.954545455	0.965593434
4	Free	0.739338235	0.988977342	0.875927729	0.739338235	0.739782102
4	Free	1	0.627659574	1	1	1
4	Free	0.911214953	0.699066874	0.871852328	0.911214953	0.879976002
4	Irchel_Campus_Scenario_1	0.703053435	0.139593909	0.792032053	0.703053435	0.724233869
4	Irchel_Campus_Scenario_2	0.534951456	0.980707395	0.715598046	0.534951456	0.485472297
5	Free	0.743119266	0.097560976	0.748687073	0.743119266	0.727580246
5	Irchel_Campus_Scenario_1	0.581651376	0.841463415	0.709639106	0.581651376	0.558238867
5	Irchel_Campus_Scenario_2	0.483950617	0.18579235	0.525745737	0.483950617	0.456985827
5	Tram_7_Scenario	0.912987013	0.289416847	0.927196876	0.912987013	0.913194146
6	Free	0.817821782	0.868247694	0.79788197	0.817821782	0.776221532
6	Free	0.704014273	0.742127154	0.679585225	0.704014273	0.667514361
6	Free	0.803331872	0.465225015	0.791356655	0.803331872	0.78579031
6	Irchel_Campus_Scenario_1	0.752380952	0.923076923	0.837553584	0.752380952	0.760586228
7	City_Center_Scenario_1	1	0.970543739	1	1	1
7	Irchel_Campus_Scenario_1	0.593623188	0.938342967	0.720668427	0.593623188	0.616513593
7	Irchel_Campus_Scenario_2	0.666115702	1	0.604964507	0.666115702	0.615190209
8	Free	0.750180505	1	0.79787709	0.750180505	0.763231979
8	Single_Family_Scenario	0.92	0.102222222	0.872298947	0.92	0.890298324
9	Central_Campus_Scenario	0.833566434	0.227906977	0.87147497	0.833566434	0.830962836
9	City_Center_Scenario_2	0.836956522	0.963791968	0.905795372	0.836956522	0.850031263
9	Free	0.691666667	0.842835131	0.825376493	0.691666667	0.690034759
9	Free	0.776252723	0.9173314	0.769133229	0.776252723	0.742297379
9	Irchel_Campus_Scenario_1	0.78877551	0.731748727	0.808829923	0.78877551	0.793997366
9	Irchel_Campus_Scenario_2	0.691774892	0.235632184	0.713461554	0.691774892	0.663073142
10	Free	0.695061728	0.090349076	0.710204472	0.695061728	0.676140218
10	Irchel_Campus_Scenario_1	0.507507508	0.998003992	0.667270874	0.507507508	0.504054871
10	Irchel_Campus_Scenario_2	0.433463035	0.15245478	0.630923863	0.433463035	0.407898541
10	Tram_7_Scenario	0.874267101	0.350649351	0.928484218	0.874267101	0.885195096
11	City_Center_Scenario_1	0.786187845	0.981651376	0.6889231	0.786187845	0.719787973
11	City_Center_Scenario_1	0.833076923	0.138107417	0.809736456	0.833076923	0.804393516
11	Single_Family_Scenario	0.819587629	0.902229846	0.818071112	0.819587629	0.803560738
11	Tram_7_Scenario	0.731271478	0.96347032	0.783465325	0.731271478	0.708395844
12	City_Center_Scenario_1	0.825559105	1	0.802588969	0.825559105	0.804089291
12	Free	0.786567164	0.104218362	0.815360332	0.786567164	0.776217022
12	Irchel_Campus_Scenario_1	0.662903226	0.992831541	0.57059971	0.662903226	0.594927179
12	Irchel_Campus_Scenario_2	0.577304965	1	0.524553146	0.577304965	0.538052619
12	Tram_7_Scenario	0.818412698	0.980972516	0.85254533	0.818412698	0.826972516

Personal declaration

I hereby declare that the submitted thesis is the result of my own, independent work. The AI tool Grammarly and ChatGPT used for grammar checking and language refinement. All external sources are explicitly acknowledged in the thesis.

Chenxi Jiang

Zurich, 31 January 2025