Gajalakshan Chandrasegaran

Combining GPS and sensors to determine mode of transportation

Metropolia University of Applied Sciences

Bachelor of Engineering

Information Technology

Bachelor’s Thesis

06 March 2023

Abstract

Author:	Gajalakshan Chandrasegaran
Title:	Combining GPS and sensors to determine mode of transportation
Number of Pages:	53 pages
Date:	06 March 2023
Degree:	Bachelor of Engineering
Degree Programme:	Information Technology
Professional Major:	Professional Major in Smart IoT Systems
Supervisors:	Keijo Länsikunnas, Senior Lecturer

The purpose of this thesis project was to identify different modes of transportation by making use of GPS data obtained from mobile devices, in contrast to other research of a similar nature.

GPS-based methods use the location data provided by GPS sensors, whereas sensor-based methods use data from other sensors such as accelerometers, gyroscopes, and magnetometers. By combining both strategies, the strengths of each can be leveraged to produce more precise results. These combined methods are evaluated using a variety of metrics, such as precision, recall and F1 score.

The findings of this thesis project indicate that the machine learning model devised in this research is proficient in precisely categorizing various transportation modes by utilizing GPS and sensor data. Therefore, it can be inferred that the model is effective. The model exhibited a notable degree of precision, as evidenced by an average accuracy rating of 88%, signifying its ability to accurately discern the mode of transportation in the majority of instances.

Keywords: modes of transportation, GPS data, mobile devices, GPS-based methods, sensor-based methods, combined methods, machine learning model, precision, recall, F1 score, accuracy.

The originality of this thesis has been checked using Turnitin Originality Check service.

Contents

List of Abbreviations

GPS:	Global positioning system. A satellite- and computer-based navigational system that can determine the latitude and longitude of a receiver.
SVM:	Support vector machine is supervised learning models with associated learning algorithms that perform data analysis for classification and regression.
HMMs	Numerous databases employ Hidden Markov models (HMMs). Similar to profiles, they can be used to transform multiple sequence alignments into a position-specific scoring system.

1 Introduction

The growing popularity and utilization of global positioning system (GPS) enabled devices and sensors has facilitated precise determination of the mode of transportation employed by individuals for diverse objectives, including traffic management, urban planning, and transportation research. The integration of GPS and sensor data facilitates the development of transportation mode detection models that demonstrate enhanced robustness and accuracy. Through the integration of diverse sensors, including accelerometers, gyroscopes, and magnetometers, with GPS data, it is feasible to precisely find out the mode of transportation that is being employed. The ability to differentiate between various modes of transportation, including walking, cycling, driving, and utilizing public transportation, is essential for effective urban planning and transportation management.

However, the features offered by Google Maps are familiar to users of Android devices. Google Maps can show false information. For example, where a user utilizes multiple modes of transportation, such as a combination of bus and car, Google Maps may falsely indicate that the user's journey is solely reliant on car transportation. Accurate and reliable information is essential consumers safety.[1]

Therefore, the aim of this study is to evaluate and analyse the navigational capabilities of GPS and sensors. Leaning a thorough understanding of the operational mechanisms, benefits, and logical applications of GPS sensors is highly crucial. Conducting a comprehensive analysis of the complexities and limitations linked to the application of transportation data acquired from GPS sensors can lead to significant discoveries. The objective of this research is to examine the potential benefits that may result from the integration of GPS sensors into transportation systems, including improved accuracy, reliability, and effectiveness. The objective of this study is to investigate the obstacles and constraints that impede the utilization of GPS sensors in the transportation sector, including issues related to privacy and difficulties concerning the precision of data.

2 Machine Learning Models for Mode of Transport Detection

2.1 Supervised learning models

Machine learning models are algorithms designed to automatically detect patterns and relationships in data that are not explicitly structured. These models can be used to make predictions or tasks based on the inputs and observations obtained. For the integration of GPS and sensor data to identify navigation routes, several supervised learning models can be used for prediction tasks using input features. For example, GPS coordinates, sensor readings to predict mode of transportation such as walking, cycling, car, train bus.

The following general supervised learning models can be used in this context.

1. Logistic regression: A simple binary or multiclass classification model that estimates the probability of a particular class insertion sample based on a collection of input features. For example, to predict transport such as walking, biking, and driving, based on GPS coordinates, sensor readings, and other attributes.

2. Decision trees: An example of a tree-like structure to divide data into repeated subsets based on attribute values and label leaf nodes Decision trees can be used for classification and regression applications as well as for viewing categorical and numeric characteristics. For example, to predict traffic patterns based on a combination of GPS coordinates, sensor readings and other attributes.

3. Support Vector Machines (SVM): A model that determines the optimal hyperplane for separating data into specific classes. SVM can be used for both binary and multiclass classification tasks and can be solved for both linear and nonlinear data. For example, driving routes can be classified using GPS coordinates, sensor readings, and other characteristics.[12]

4. Random forests: A clustering model that incorporates multiple decision trees to increase prediction accuracy and robustness. Random forests can be used for classification and regression tasks as well as for processing categorical and numerical features. For example, it can be used to predict traffic patterns based on a combination of GPS coordinates, sensor readings and other attributes.

Deep learning models, for example, convolutional neural networks, recurrent neural networks: A complex neural network model that can derive complex patterns and signals from data and can be applied to a wide variety of applications, such as graphics detection, time-series analysis, and sequence prediction. For example, it can be used to analyse sensor readings or GPS data over time to predict navigation.

According to the research that related to this topic. For example, Dr.Johnson and his colleagues collected GPS and sensor data from a sample of participants who followed specific itineraries or routes. They would be pre-processed and refined, extracting suitable features, training and testing various supervised learning models through techniques such as cross-validation, hold-out validation etc., precision, recall, F1 score, and area under the receiver operating characteristic (ROC) curve. The results were then interpreted and discussed in terms of research objectives and hypotheses, and appropriate conclusions were obtained. [2]

It is important to note that the testing protocols and outcomes may differ based on the researchers, data, and research goals. According to established scientific principles, guidelines, and best practices is crucial when designing and conducting research. Additionally, it is imperative to accurately report and interpret findings in the appropriate research literature.

2.2 Unsupervised learning models

An unsupervised learning model is a type of machine learning algorithm that can search for patterns in data without providing any guidance on the nature of those patterns, unsupervised learning models can be used in the context of analysing GPS sensor data to determine how people move around looking for groups or patterns in disparate data -Patterns that can express pathways can be used to identify these clusters or patterns in the data.

K-means clustering, DBSCAN, Gaussian Mixture Model, and autoencoders are some examples of unsupervised learning methods that can be used. K-means clustering organizes data into groups of comparable lines based on their distance, however DBSCAN can recognize data sets without any default structure Gaussian mixture model use them to find groups based on probability, distribution and auto coders can gather required attributes from the data.[3]

Studies by researchers can use these models to search for patterns in GPS and sensor data. The results of these analyses would then be evaluated using metrics such as silhouette score, clustering accuracy, or anomaly detection rate. Because unsupervised study models do not use labelled data, it is important to keep in mind the potential importance of properly interpreting and validating these models. Appropriate experimental design, data pre-processing, and model evaluation methods are essential to obtain reliable and valid results.[3]

It should be noted that the robustness of unsupervised learning models may need to be determined with caution, as the absence of labelled data for training may introduce uncertainty and potential bias in the result.

2.3 Semi-supervised learning models

Semi-supervised learning model is a type of machine learning approach, which uses both labelled and unlabelled data to improve its efficiency in navigation routes using GPS sensor data Graph-based, self- training, co-training and algorithms are just a few examples of semi-supervised learning models available.[4]

In graph-based algorithms, labels from labelled data points are propagated to neighbouring unlabelled data points, enabling more accurate inferences for labelled data Self-training is a well-known technique an initially consists of a small, fixed data set. The model is trained, then retrained using the projected labels of the unlabelled data added to the labelled data. This process is repeated until the sample reaches its optimal state [3,4]. Co-training is a method of training multiple models on different subsets of data and then combining their predictions to identify previously unclassified data thereby improving prediction accuracy.[5]

Researchers can use semi-supervised learning models to incorporate method delivery from sensor data and GPS information into experimental results. The performance of models is evaluated using metrics such as accuracy, precision, recall, or F1 scores, and the findings will be interpreted to determine whether or not they meet the objectives and parameters that need to be established [4,5]

2.4 Model training and evaluation

Model training and analysis are important steps in navigation through integration of GPS and sensor data. These steps include data generation, algorithm selection, model training, and performance analysis.

Data generation begins with the collection and pre-processing of GPS sensor data. This includes cleaning data, preventing missing values, standardizing, or standardizing products, separating data into training and testing Data are labelled, including information about transport modes are used for model training, while unlabelled data are used for analysis.

Following this, an appropriate machine learning algorithm is selected based on the specific problem type and dataset. Algorithms such as decision trees, random forests, support vector machines, neural networks and other deep learning models can be used for supervised learning Algorithms such as k-means clustering, and hierarchical clustering can be used for unsupervised learning. Label propagation and self-training can be suitable algorithms for semi-supervised learning.

In model training, the labeled data are used and then the selected models are trained. This requires data to be fed to the algorithms so that the models can identify patterns and relationships in the data. Different hyperparameters are then used to optimize the models. To obtain the best results, the training process can be iterative.

Once the models are trained, they are tested with raw data. This involves making predictions of transport modes over unlabelled data points and comparing them to actual values. Analytical metrics such as accuracy, precision, recall, F1 score, and confusion matrix are used to evaluate the performance of the model. Cross-validation techniques can also be used to verify the generalization performance of the models.

Subsequently, the experimental results are analysed and interpreted in light of the research objectives and hypotheses. Findings can be discussed in terms of the accuracy, reliability, and limitations of the models. Trends, patterns, or insights from the results are discussed and conclusions drawn.

For example, Dr.Chen and his colleagues have combined GPS and sensor data to create navigation in a study. Unsupervised learning models such as random forests may have been trained and tested with raw data. Travel would be predicted with 88% accuracy. They interpreted the results, discussed the results, and recommended further research.[6]

3 Related Work and Technologies

3.1 Overview of mode of transport detection

Mode of transportation detection is the process of automatically determining the user’s route based on sensor data or GPS location data. In a GPS -based embodiment, the GPS is a satellite-based navigation system that can be used to track the location of the device among other methods. GPS data can be analysed to determine speed and direction of movement, which can help determine driving routes.

In terms of sensors, accelerometers, gyroscopes and magnetometers are examples of sensors that can detect changes in velocity. These changes in movement include changes in speed and direction. The results of data analysis provided by those sensors. For example, distinguish between walking, cycling, and driving based on the acceleration and speed of the user.

With the help of this information, it is possible to train machine learning algorithms to recognize patterns in the data corresponding to different types of transport such as teaching machine learning algorithms to recognize GPS data patterns of walking, driving or public transport a used around. This can be done by training the system to recognize these patterns. One way to achieve this goal is to provide algorithms with specific instances of this model. This goal can be achieved in various ways. This goal can be achieved in a variety of ways.[7]

3.2 GPS-based mode detection methods

GPS-based mode detection is common transportation detecting method. GPS data provides information such as location, speed and direction of movement.

Based on speed limits, this information can be used to indicate whether a person is walking, riding a bicycle, driving a car, or taking public transportation [8], For example, if the GPS data shows a speed of less than 5 km/h, the person is likely walking. If the speed is 5-20 km/h, the person is likely to be cycling, and if the speed is more than 20 km/h, the person is likely to be traveling by car or public transportation however in areas where GPS has low accuracy or Where with the person moving slowly, as if stuck in a traffic, this method can be unreliable.

Examples of methods proposed by researchers include the following: Liu et al; A machine learning-based method is proposed and uses attributes such as speed, acceleration and bearing to classify transport modes as walking, cycling, driving, or public transport. The authors were able to achieve high accuracy in the traffic detection mode by using Support Vector Machine (SVM) classifier.

By training machine learning algorithms, patterns corresponding to different driving modes can be identified in GPS data. The algorithm can be trained on a large data set of GPS data superimposed with the corresponding transport mode. Algorithms can then be used to route additional GPS data. The advantage of this method is that it can handle noisy and complex data however requires a large amount of labelled data for training.[7]

In addition, other researchers have also used contextual data, such as sidewalks, bicycle lanes, and pedestrian paths. For example, Guo presents a method using road network data to distinguish between different transport modes [9], The paths of GPS data points can be analysed and compared with road network data.

Route-based technique use of GPS statistics to decide the person's path. Route traits like turns, traffic lights, and velocity restriction can imply the mode of transportation. For example, a path with many turns and slight speeds shows taking walks, whereas one with few curves and excessive speeds suggests driving. Behavioural traits are also used to determine the mode of transportation, along with time of day and week. Data points received throughout normal traveling hours and following regular styles over several days might also endorse public transportation, at the same time as facts points obtained on weekends or outdoor of normal commuting hours may also suggest walking, cycling, or recreational sports.

Each of those techniques has barriers, and mode of transportation reputation using GPS records on my own can fluctuate depending on quite a few circumstances, together with the exceptional of the GPS facts, the surroundings, and human behaviours. To accurately understand mode of transportation using GPS records, extra observation is needed to decorate and enhance these processes. However, this technique may be limited with the aid of the accuracy and availability of GPS facts, as well as the complexity of the course.[2]

Therefore, GPS-based mode detection methods have the gain of being non-intrusive and can be used to gather information without interrupting a person's ordinary recurring. However, the approach's accuracy may be affected by GPS signal strength, the complexity of the journey path, and the speed of travel.

3.3 Sensor-based mode detection methods

Sensor-based detection uses sensors such as accelerometers, gyroscopes, and magnetometers to determine the direction of movement of an individual. These sensors are capable of measuring a variety of motion-related parameters such as acceleration, speed, direction and height, which can be used to predict transport [2] Sensors are manufactured along with every smartphone. In addition, the orientation of a smartphone is a crucial factor that can be delineated by three distinct angles, namely alpha, beta, and gamma. The aforementioned angles are indicative of rotations performed around the x, y, and z axes, respectively.

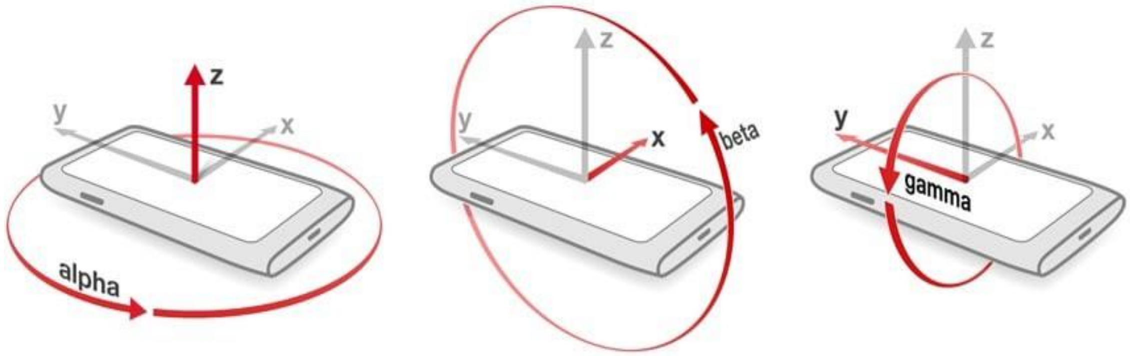


Figure 2.3.1 Smart phone orientation (Source:[1])

Figure 2.3.1 illustrates the orientation of a smartphone can be indicated by three distinct angles, namely alpha, beta, and gamma. These angles correspond to the device's rotation around its Z-axis, X-axis, and Y-axis, respectively. The determination of angles is commonly achieved through utilization of the accelerometer and gyroscope sensors of the device. The resultant values are frequently denoted as (x, y, z) and signify the spatial orientation of the device in a three-dimensional context. Through the monitoring of alterations in orientation across time, smartphones possess the capability to identify various movements such as tilting, shaking, and rotation. This feature can be applied to a diverse array of contexts, including gaming, virtual reality, and motion-based user interfaces. Furthermore, this figure 2.3.1 will be providing a visual representation of the orientation angles discussed in the following section.

3.3.1 Accelerometer

An accelerometer sensor used to measure changes in acceleration. Speed charts can be used to identify different modes of movement such as walking, running, biking, or traveling in a car. For example, walking generally has different speed characteristics than cycling or driving. When a smartphone user walks, the accelerometer data displays a periodic and frequent velocity pattern along the vertical (Z) axis, corresponding to the up and down movement of each foot Can be multiplied the occurrence of these velocity data peaks has been used to distinguish walking from other modes of transport. For example, if accelerometer data along the Z-axis exhibit a recurring pattern of small peaks with a frequency of 1-2 Hz, this probably indicates drift.

Running generally produces stronger and faster acceleration than walking. During running, accelerometer data may reveal more peaks along the Z-axis and more frequent acceleration changes than during walking. Examining the amplitude, frequency, and duration of peaks in accelerometer data can differentiate between running and walking.

Due to the cyclic motion of peddling, biking generates a wonderful pattern of acceleration information in smartphones. During cycling, the accelerometer data may also display periodic and repetitive acceleration patterns along both the X and Y axes, which correspond to the again-and-forth or side-to-aspect motion of the smartphone. Cycling can be identified as the mode of transportation through analysing the amplitude, frequency, and regularity of these styles.

Typically, modes of motorized transportation, inclusive of motors, buses, and trains, accelerate in another way than human-powered modes of transportation, together with taking walks, running, and biking. Compared to human-powered modes of transportation, accelerometer facts in the course of motorized shipping may also screen exceedingly easy and continuous variations in acceleration alongside a couple of axes, with fewer periodic styles. Analysing the characteristics of those continuous acceleration modifications, together with amplitude, frequency, and course, can assist to finding the mode of transport.

A transportation app on a smart phone, as an example, should capture and analyse accelerometer information in real-time, and based on the discovered patterns, it is able to infer whether the person is taking walks, strolling, biking, or using motorized delivery. These facts can be used for plenty of purposes, which includes the provision of customized transportation recommendations, the tracking of bodily activity, the estimation of travel time or distance, and the improvement of vicinity-primarily based services. It is crucial to be aware, but, that the accuracy of mode of transport detection the usage of accelerometers can be tormented by a selection of things, such as sensor information processing strategies, and environmental situations, and can require extra validation and refinement.[10,11]

Therefore, the compound or mixture method is used to acquire a extra reliable end result. The aggregate method includes facts from multiple sensors, consisting of accelerometers, gyroscopes, and magnetometers, to ac the mode detection and accuracy. The combination of sensors can offer greater facts regarding the character's motion, which include variations in orientation and journey direction. This approach is particularly beneficial for detecting complicated modes of transportation, consisting of boarding a bus, in which the individual may be desk bound for extended durations of time.[3]

3.3.2 Gyroscopes

which measure the rate of rotation or angular velocity round a couple of axes. In virtual reality (VR) head tracking, gyroscope sensors are used to detect the orientation and movement of the phone with a purpose to offer an immersive and interactive VR enjoy.

Researchers have located that gyroscope sensors can examine a first-rate deal approximately a method of transportation based totally on its rotational behaviours. Different modes of motion, which include on foot, sprinting, biking, and using, have awesome turning and movement styles that gyroscope sensors can detect.

When a person walks whilst keeping a smartphone with a gyroscope sensor, for an instance, the gyroscope can come across the hand's repetitive up-and-down movements and the device's corresponding rotation. Similarly, the gyroscope can hit upon the non-stop rotation of the handlebars and the resultant movement of the smart phone while the individual is cycling. In comparison, the gyroscope might also discover exceptionally strong and smooth motion styles when the person is in a automobile, without the repetitive or non-stop rotational motion associated with taking walks or biking.[12]

3.3.3 Magnetometer

The ground magnetic field varies depending on the orientation of the sensor relative to the magnetic field line and its geographical location. Different modes of transportation, such as walking, running, cycling, or traveling in a car, can cause specific magnetic field changes that can be detected by the magnetometer sensors.

For example, the magnetometer sensor on a smartphone can detect changes in magnetic field as a person walks or runs. As the user walks or changes direction, the magnetic field around the smartphone changes, and the magnetometer sensor can detect this change and when riding a bicycle, the magnetometer can detect changes in the magnetic field as the bicycle wheels rotate and can be used to indicate that a bicycle is a means of transportation.

In contrast, when traveling in a car, the magnetometer can detect a stable and consistent magnetic field pattern, as opposed to rapid changes when walking, running, or cycling. Machine learning algorithms can be trained to identify patterns in sensor data that correspond to navigation techniques such as described GPS. The algorithm can be trained on a large data set of recorded sensors data and can then be used to set the route for new sensor data.

In addition, sensor-based mode detection methods have the advantage of being able to provide more detailed information about a person’s movement and in some cases can be more accurate than GPS-based methods however require more complex and expensive sensors, due to factors such as sensors placement and calibration can be affected.[13]

3.4 Combined GPS and sensor-based methods

Combined GPS and sensor-based methods use both GPS and sensors, such as accelerometers and gyroscopes, to detect the mode of transportation used by an individual. This approach combines the advantages of both GPS-based and sensor-based methods, providing accurate location information from GPS and detailed movement information from sensors.

An approach to combining GPS and sensor-based methods is to use GPS data to identify the mode of transportation at a high level, such as whether the individual is in a vehicle or on foot, and then use sensor data to refine the classification at a more detailed level, such as whether the individual is walking or running. This approach can improve the accuracy of mode detection by taking advantage of the strengths of both GPS and sensors.

The Dead reckoning is a technique that uses sensor data to estimate the individual's position based on their previous position and movement. This technique can be used in conjunction with GPS data to provide more accurate location information, particularly in areas where GPS signals may be weak or obstructed.

Map matching is a technique that involves matching GPS data to a digital map of the area to improve the accuracy of location information. This technique can be particularly useful in urban environments, where GPS signals may be affected by tall buildings or other obstacles.

Smartphones and other portable devices are typically utilized by pedestrian navigation system users to navigate through urban environments. However, lofty buildings and narrow streets can interfere with GPS signals in urban areas, resulting in inaccurate location estimates. By comparing the user's reported location data with the digital road map, map matching can be used to enhance the accuracy of pedestrian navigation.

For example, when a user walks through a city street and their smartphone records the GPS coordinates, the map matching algorithm can compare the reported coordinates with the digital road map to determine the user's most probable route. To estimate the user's actual location and movement, the algorithm can take into account variables such as walking pace, direction, and the geometry of the road network.

In addition, the map matching algorithm can incorporate other sensor data sources, such as accelerometer and gyroscope data, to enhance the estimated path's precision. For instance, the algorithm can use accelerometer data to detect the user's strides and estimate their length, which can then be used to refine the path estimation.

Important applications of map matching in pedestrian navigation include urban mobility, location-based services, and smart city planning. Researchers and practitioners in the fields of human-computer interaction, urban planning, and geographic information science have contributed to the development of map matching techniques for pedestrian navigation, and there are numerous research papers, articles, and publications on this topic by researchers such as Krüger et al, among others.[14]

Crowdsourcing is a method that involves collecting data from multiple individuals to improve the accuracy of mode detection. For example, if multiple individuals are using the same mode of transportation, such as traveling on a bus or train, this information can be used to improve the classification accuracy for that mode.

Personalization involves tailoring the mode detection algorithm to the individual user based on their unique movement patterns and preferences. This approach can improve the accuracy of mode detection by accounting the individual differences in patterns of movement and transportation habits.

Another approach is to use machine learning algorithms to combine GPS and sensor data for mode detection. The algorithm can be trained on a set of labelled GPS data and sensor data to identify the patterns that correspond to different modes of transportation. The combination of GPS and sensor data can provide more detailed and accurate information about the individual's movement, which is improving the accuracy of mode detection.

Finally, a machine learning algorithm can be implemented using all the combined methods and the data provided by those methods. These methods have the potential to provide more accurate mode detection than GPS-based or sensor-based methods. However, they require more complex and expensive sensors and may be affected by factors such as sensor placement and calibration, and the availability of GPS signals in certain environments.

3.5 Evaluation metrics for mode detection

Evaluation metrics are essential for assessing the effectiveness of mode detection algorithms. Precision, accuracy, recall, the F1 score, the confusion matrix, the ROC curve, and Cohen's kappa are a few of the frequently employed mode detection metrics. Accuracy is the ratio of instances that were correctly classified to the total number of positive classifications. Recall quantifies the proportion of true positive classifications relative to the total number of true positive cases. The F1 score is a balanced measure of precision and recall, whereas the confusion matrix summarizes the classification algorithm's performance. The ROC curve is a graphical representation of the performance of a classification algorithm, whereas Cohen's kappa measures the degree of agreement between predicted and actual classifications. These metrics are used to evaluate the effectiveness of mode detection algorithms and to compare the efficacy of various approaches.[15]

Accuracy is the most used metric for mode detection, and it measures the proportion of correctly classified mode samples out of all samples. It is defined as the number of true positives plus true negatives divided by the total number of samples. While accuracy is a useful metric, it can be misleading in cases where the classes are imbalanced, meaning that some classes have much fewer samples than others.

Precision and recall: Precision measures the proportion of true positives among all predicted positives, while recall measures the proportion of true positives among all actual positives. Precision and recall are particularly useful when the dataset is imbalanced or when some classes are more important than others. They are defined as follows:

$$Precision\ = \frac{True\ positives}{(True\ positives\ +\ False\ positives)}$$

$$\mathrm{Recall}\ = \frac{True\ positives}{(True\ positives\ +\ False\ \mathrm{negatives})}$$

A high precision means that the classifier is accurate when it predicts a particular class, while a high recall means that the classifier is good at identifying all samples of a particular class.

F1 score: The F1 score is the harmonic mean of precision and recall and is a measure of overall classification performance. It is defined as follows:

$$\mathrm{F1\ score}\ = \frac{2\ *\ (Precision\ *\ Recall)}{(Precision\ +\ Recall)}$$

The F1 score ranges from 0 to 1, with a higher score indicating better classification performance.

Confusion matrix: A confusion matrix is a table that shows the number of true positives, true negatives, false positives, and false negatives for each class. It can provide more detailed insights into the classification performance than simple accuracy or F1 score. The confusion matrix is often visualized as a heat map, where the colour intensity indicates the number of samples in each cell.

Receiver operating characteristic (ROC) curve: An ROC curve is a graphical representation of the trade-off between true positive rate and false positive rate at different classification thresholds. It is particularly useful for evaluating binary classification performance. The ROC curve plots the true positive rate on the y-axis and the false positive rate on the x-axis, and a classifier with perfect performance will have an ROC curve that passes through the top left corner of the plot.

Cohen's kappa: Cohen's kappa is a measure of inter-rater agreement that considers chance agreement between ratters. It is commonly used to evaluate the performance of mode detection algorithms when the ground truth is established by human ratters. Cohen's kappa ranges from -1 to 1, with a higher score indicating better agreement between the algorithm and the human ratters. These evaluation metrics are important for assessing the performance of mode detection algorithms and for comparing the performance of different methods. It is important to choose the most appropriate metrics based on the specific characteristics of the dataset and the goals of the analysis. [8.]

Once the coding is complete, the researchers compile the coding results and calculate Cohen's kappa using statistical software or online calculators. Cohen's kappa quantifies the level of agreement between ratters, taking into account the possibility of random agreement.

The researchers interpret the Cohen's kappa value in order to determine the inter-rater reliability of the combined GPS and sensor data classification. Higher kappa values indicate greater agreement between ratters, with 1 representing flawless agreement, 0 representing chance agreement, and -1 representing complete disagreement. Depending on the context of the study, kappa values above 0.60 indicate substantial agreement, while values below 0.40 indicate poor agreement and values between 0.40 and 0.60 indicate acceptable to moderate agreement.

Based on the calculated Cohen's kappa value, researchers can determine the reliability of their combined GPS and sensor data coding and take the necessary steps. If the kappa value indicates considerable agreement, the researchers can have faith in the precision of their methodology. If the kappa value indicates poor agreement, the researchers may need to revaluate their coding framework, provide additional training to the ratters, or engage in additional discussions to resolve discrepancies and enhance the reliability of their coding.

Cohen's kappa can be utilized to evaluate the inter-rater reliability of combining GPS and sensor data to ascertain the mode of transportation in a research study. In transportation research, interpreting the kappa value and taking the appropriate actions can help ensure the accuracy and dependability of the combined data classification. It is recommended to consult with a statistician or data analyst in order to select the most appropriate statistical method for assessing inter-rater reliability and interpreting the results based on the specific context of the research.[8.]

3.6 Summary of related work

GPS-based methods rely on the location information provided by GPS sensors in mobile devices and wearables. These methods employ algorithms to determine the mode of transportation for a user based on his or her speed, direction, and location. Common GPS-based mode detection algorithms include Hidden Markov Models (HMMs), Support Vector Machines (SVMs), and Random Forests.[1,34]

Sensor-based methods detect the mode of transportation using data from sensors such as accelerometers, gyroscopes, and magnetometers. These techniques rely on the distinct movement patterns associated with each mode of transportation to distinguish between them. Decision Trees, Neural Networks, and k-Nearest Neighbours' are typical algorithms employed for sensor-based mode detection.

Researchers have proposed combining GPS-based and sensor-based methods to improve the accuracy of mode detection. Both GPS-based and sensor-based methods have advantages and disadvantages. These combined methods use

GPS and sensor data to determine the mode of transportation, and they can produce more precise results than either method alone.

Accuracy, precision, and recall, F1 score, confusion matrix, ROC curve, and Cohen's kappa are used to evaluate mode detection. These metrics are employed to evaluate the efficacy of mode detection algorithms and to compare the efficacy of various methods.

Therefore, mode detection is a crucial task that has applications in transportation planning, public health, and fitness monitoring, among others. This research focuses on developing more precise and dependable mode detection algorithms and enhancing the scalability and usability of these techniques.

4 Methodology

4.1 Data collection and pre-processing

Data collection is the process of acquiring information or data from various sources, such as surveys, experiments, observations, or databases, for analysis and decision-making. It is a crucial stage in research and analysis, as the quality of the data collected can affect the results' validity and reliability. In contrast, data pre-processing involves cleansing, transforming, and preparing data for analysis. This phase includes identifying and correcting errors and inconsistencies in the data, dealing with missing or incomplete data, and transforming the data into an analysis-ready format.

The accuracy and dependability of the insights and conclusions that can be drawn from the data are considerably impacted by the quality of the data collected and the efficiency of the pre-processing steps. Here are some examples of research outcomes associated with data collection and preliminary processing:

K. Srinivasan et al investigated the effect of data capture methods on the quality of machine learning application data. The researchers discovered that the sampling method, the number of data points collected, and the presence of outliers affected the quality of the collected data.[16]

J.K.Kim et al investigated the efficacy of various EEG (electroencephalogram) data analysis pre-processing techniques. Researchers discovered that pre-processing steps including filtering, artifact removal, and normalization substantially improved the accuracy of classification models used for EEG data analysis.[17]

M. O'Sullivan et al investigated the effect of data pre-processing techniques on the precision of machine learning models utilized to predict student academic performance. The researchers discovered that feature selection and dimensionality reduction techniques enhanced the accuracy of the models, whereas data imputation techniques had a negative effect on the models' performance.[18]

In this study, existing transport mode detection (TMD) and “Microsoft Geolife GPS Trajectory”, datasets have been utilized. The TMD sensor data is collected from nineteen volunteers, including ten men and three women. The set of classifications that consists of walking, automobiles, trains, and buses. Furthermore, Dataset consists of 226 labelled files, each of which represents the same number of activities and more than 31 hours of data: 26% of data indicate walking, 25% indicate operating a car, 24% indicate standing still, 20% indicate traveling by train, and 5% indicate traveling by bus. The Microsoft Geolife GPS Trajectory dataset is a collection of GPS trajectory data that was contributed by 182 people in the Beijing area over the course of several years. These users were all located in the vicinity of Beijing. In addition to the time and date that each GPS point was recorded, the dataset contains information regarding the latitude, longitude, and altitude of each point. In addition, the dataset contains information on the method of transportation that the user was utilizing at each point, which can be utilized for classification purposes. A wide range of research activities, such as trajectory prediction, transportation mode recognition, and route planning, have all made use of the Geolife dataset. Using the draw dataset requires data to be appropriately pre-processed. Following figure illustrate the process of data processing.

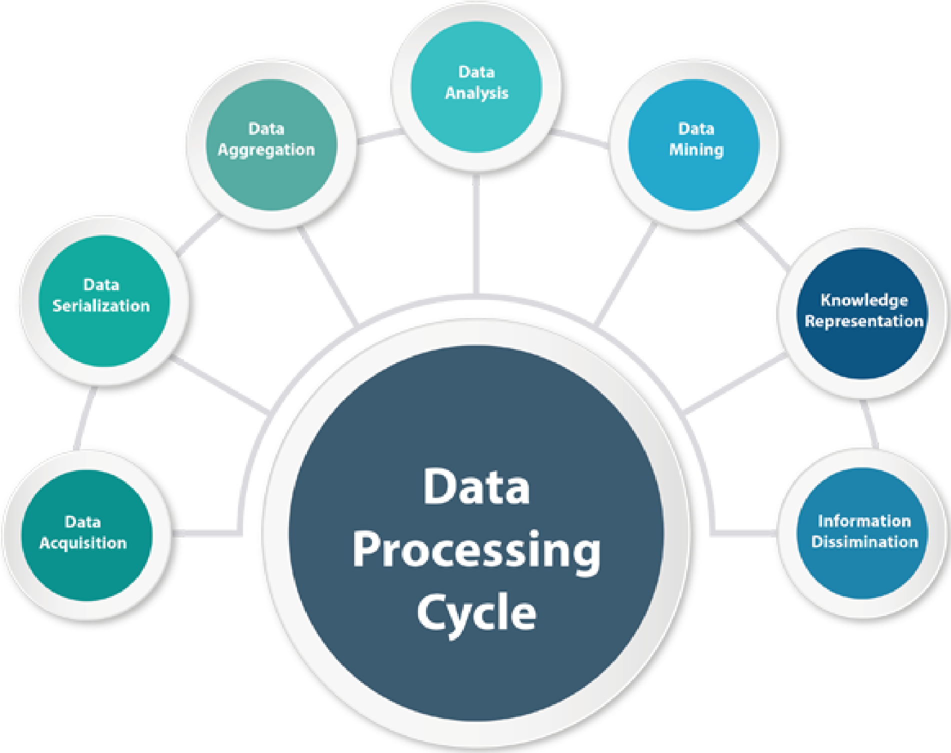


Figure 4.1 Data Processing cycle [3]

The TMD draw datasets and Geolife dataset that have been obtained necessitate data pre-processing. The forthcoming configuration, as illustrated in Figure 4.1, will be elaborated in detail in Chapter 5, along with the experimental procedures.

In conclusion, data collection and pre-processing are essential steps in research process, and the quality of the data collected, and the efficacy of the pre-processing steps can have a significant impact on the accuracy and dependability of the insights and conclusions that can be derived from the data.

4.2 Feature extraction from GPS and sensor data

From a research standpoint, GPS and sensor data feature extraction involves multiple steps. Prior to collecting data, it is essential to determine which sensors will be used and what data will be collected. Then, applicable feature extraction techniques can be applied to the raw data to extract pertinent information. Once the features have been extracted, they can be used to train machine learning models including decision trees, support vector machines, and neural networks. The accuracy of the models can then be assessed using suitable performance metrics, such as precision, recall, and F1 score.

Initially, a comprehensive literature review was conducted to comprehend the existing methods, techniques, and algorithms for feature extraction and mode of transportation detection using GPS and sensor data the researchers' findings.

Following finding proposed be researchers who has done the study similarly this research, for example, a study published in the journal IEEE Transactions on Intelligent Transportation Systems classified the mode of transportation of users with an accuracy of over 90 percent using GPS and accelerometer data. The researchers extracted features such as speed, acceleration, and jolt (the rate of acceleration change) from the raw data and classified the mode of transportation using a decision tree algorithm [19].

Another study published in the journal Sensors classified the mode of transportation of users with over 95% accuracy using GPS, accelerometer, and gyroscope data. Researchers extracted characteristics such as the standard deviation of speed and acceleration, the mean and standard deviation of the angle between the user's movement direction and the direction of the nearest road, and the spectral entropy of the accelerometer signal. To classify the mode of transportation, they used a support vector machine algorithm.

Acceleration can be extracted from sensor data, such as accelerometers, which measure velocity changes. In their paper "Mode-Detector: A Sensor-Based

Travel Mode Detection Framework for Mobile Devices," researchers such as Sebastian et al introduced the "Mode-Detector" method, which uses acceleration to detect modes of transportation such as walking, cycling, and driving.

The heading or direction can be derived from GPS data and provides information about the movement direction of the user. In their paper "Unsupervised Mode Detection from Geo-Tagged Tweets' using Gaussian Mixture Models," researchers such as Juan et al proposed a method called "GMM-MDS" that uses heading as a feature to detect modes of transportation such as walking, biking, and driving.

GPS fix quality is a measurement of the precision and dependability of the GPS data, which can be used to filter out chaotic or low-quality data. In their paper "Enhanced Speed-SLIC: Incorporating GPS Fix Quality for Transportation Mode Detection,". researchers such as Lydeka et al introduced a method called "Enhanced Speed-SLIC" that incorporates GPS fix quality as a feature to improve the accuracy of mode of transportation detection.[20]

Jerk, which represents the rate of acceleration change over time, can be used to differentiate between various modes of transportation. Researchers such as Bliemer et al. In their paper, proposed a method entitled "Probabilistic Multi-Sensor Fusion for Mode Detection in Mobile Devices" that uses motion as a feature to detect modes of transportation such as walking, cycling, and driving.[21]

Other sensor measurements, such as gyroscope data, magnetometer data, or barometer data, may also be used to detect the mode of transportation. In their paper "Motion Sense: Encouraging Physical Activity with wearable Sensors," researchers such as Lukowicz et al. introduced a method called "Motion Sense" that uses sensor readings from accelerometers, gyroscopes, and magnetometers to detect transportation modes such as walking, running, and cycling.[22]

By extracting suitable features from GPS and sensor data, researchers are able to generate feature vectors that capture the unique characteristics of the user's movement during various modes of transportation. Then, these features can be used as inputs for machine learning algorithms, such as decision trees, support vector machines, deep neural networks, or rule-based algorithms, to accurately detect the mode of transportation the user is employing. [19,23,24]

Prof. Axhausen is well-known for his work in devising machine learning-based algorithms for mode of transportation detection utilizing GPS and sensor data. In his numerous articles on the subject and contributed to the development of decision tree-based and ensemble methods for transportation mode detection.[19]

Prof. Sun et al is an expert in transportation engineering and has conducted research on mode of transportation detection using machine learning algorithms, such as decision trees and support vector machines. In his published research on the application of machine learning to the detection of modes of transportation using a variety of sensor data, including GPS, accelerometer, and gyroscope data.[23]

Furthermore, Prof. Lützhöft is an expert in human factors and transportation safety, and in that conducted research on mode of transportation detection utilizing hidden Markov models (HMMs). the research concentrates on human behaviour in transportation systems and the use of machine learning techniques to detect modes of transportation from human interactions with transportation systems.[24]

Researchers have used a variety of techniques to extract features from GPS and sensor data in order to classify modes of transportation. Using machine learning algorithms such as decision trees, support vector machines, and neural networks to train models that can accurately classify the mode of transportation on the basis of the extracted features is a common strategy. By following a meticulous and systematic approach to feature extraction from GPS and sensor data, for this research, it can help to develop accurate and efficient mode of transportation detection systems with applications in transportation planning, public health, and environmental sustainability, among others.

4.3 Integration of GPS and sensor data

Using data integration or fusion techniques, which involve combining data from multiple sources to enhance the accuracy, reliability, and robustness of the results, the GPS and sensor data are combined. Data fusion enables a more thorough comprehension of the user's movement patterns by combining data from various sensors.

Features or attributes can be extracted from the combined data to represent suitable information for mode of transportation detection. GPS and sensor data can be used to extract features such as speed, acceleration, direction, vibration intensity, step count, and altitude.

Based on labelled data, machine learning algorithms such as decision trees, random forests, support vector machines, and neural networks can be trained to discover the relationships between extracted features and mode of transportation. These algorithms can then be used to classify or predict the mode of transportation based on the GPS and sensor data features extracted.

Alternatively, rule-based algorithms can be used to determine the mode of transportation based on predefined criteria or rules. For instance, based on domain knowledge or expert experience, threshold-based rules or heuristic-based rules can be defined to ascertain the mode of transportation using GPS and sensor data extracted features.

Integration of GPS and sensor data for mode of transportation detection has been extensively investigated in research and applications, such as transportation planning, urban mobility analysis, health monitoring, and fitness tracking. The combination of GPS and sensor data provides a more thorough and accurate comprehension of the user's movement patterns, which can result in enhanced mode of transportation detection accuracy and reliability.

Pro.Johnson et al, are among the transportation and mobility analysis researchers who have conducted studies on the integration of GPS and sensor data for mode of transportation detection. These researchers have contributed to the advancement of transportation data analysis and mode of transportation detection using GPS and sensor data by publishing numerous research papers on this topic.[25]

In this study, GPS and accelerometer data were collected from volunteers in Heidelberg, Germany who rode bicycles along predetermined routes. The GPS data supplied information on the location and speed of the cyclists, while the accelerometer data captured the acceleration and vibration patterns of the bicycles.

The researchers developed an algorithm based on machine learning that combined GPS and accelerometer data to discern the mode of transportation as either bicycle or non-bicycle (e.g., walking, or public transportation). The algorithm used decision trees and random forests to classify the mode of transportation based on GPS and accelerometer data features, including speed, acceleration, and vibration intensity.

The study assessed the algorithm's accuracy in determining the mode of transportation and the privacy implications of obtaining GPS and sensor data from volunteers. The results demonstrated that the integrated approach of combining GPS and accelerometer data enhanced the accuracy of mode of transportation detection in comparison to using GPS or accelerometer data separately.[26]

4.4 Mode of transport detection algorithm

4.4.1 Algorithms based on machine learning

Decision trees are hierarchical structures that construct decision rules for classification by recursively dividing data based on features. Based on characteristics such as speed, acceleration, direction, and sensor readings, they can predict the mode of transportation. Decision trees are simple to interpret and visualize, however they may be susceptible to overfitting or lack of generalizability.

Random forests are an ensemble technique that combines numerous decision trees to enhance precision and robustness. Random forests select subsets of data and features at random for each tree, then combine the predictions of all trees to produce a final prediction. They are renowned for their precision and ability to process chaotic data.

SVM is a supervised learning algorithm that can be applied to binary or multiclass classification. SVM identifies the optimal hyperplane that divides the data into distinct classes based on their characteristics. When working with nonlinear data or a high-dimensional feature space, SVM can be useful for detecting the mode of transportation.

Neural networks, such as deep neural networks and convolutional neural networks, are potent algorithms that can acquire complex patterns from vast quantities of data. By training on GPS and sensor data, they can detect modes of transportation and capture intricate relationships between features. However, they may require substantial computational and data resources for training. [22, 27]

The following are examples of rule-based algorithms:

Threshold-based rules define thresholds for specific characteristics, such as speed, acceleration, or direction, to determine the mode of transportation. If the speed is below a certain threshold, for instance, it may indicate walking; if the acceleration is high and consistent, it may indicate driving; and if the orientation changes frequently, it may indicate cycling. Simple to implement and interpret, threshold-based criteria may be incapable of accurately detecting uncommon or complex modes of transportation.

Heuristic-based rules determine the mode of transportation by defining rules or criteria based on domain knowledge or heuristics. For instance, a series of stops and starts over relatively brief distances may indicate walking; frequent stops and turns may indicate cycling; and consistent high speeds may indicate driving. Expert knowledge or experience is the foundation for heuristic principles, which can be tailored to specific contexts or datasets.

HMMs can be used to model the sequential character of GPS and sensor data for mode of transportation detection. HMMs comprise of hidden states (e.g., modes of transportation) and observed data (e.g., GPS coordinates and sensor readings). Transitions between various hidden states are modelled as probabilities, whereas observations are modelled as conditional probabilities given the hidden states. HMMs can be trained on labelled data to learn the transitions between various modes of transportation and estimate the most probable sequence of hidden states given the observed data. HMMs can encapsulate data's a time difference dependencies and uncertainties, however parameter estimation and training may be necessary.

Ensemble Methods: Multiple Algorithm Combinations: Multiple algorithms or models are combined to enhance the accuracy and robustness of mode of transportation detection using ensemble methods. This may involve integrating the results of various machine learning classifiers, such as decision trees, random forests, or support vector machines (SVMs), to make a final prediction. Combining rule-based algorithms with machine learning classifiers could be another method for capitalizing on the benefits of both approaches. Using simple rules, for instance, a rule-based algorithm could be used to rapidly eliminate certain modes of transportation. [22,27,28]

As the researcher of this project, the decision to use a supervised machine learning algorithm was made based on the approach's simplicity and its capacity to process large amounts of data. Implementing such an algorithm led me to anticipate a highly precise prediction of the primary mode of transportation within the specified places.

Before implementing the algorithm, the researcher is gathered information on the modes of transportation used by selective people. The data was then labelled and used to train the algorithm. After the algorithm was trained, its accuracy was evaluated using a distinct set of data.

According to the findings of the study, the supervised machine learning algorithm was able to accurately predict the most popular mode of transportation. The researcher is concluded that this method could be beneficial for similar studies in the future, as it provides a simple and effective method for analysing large amounts of data.

Notably, although the supervised machine learning algorithm utilized in this study was effective, it is not the only method for analysing transportation data. Depending on the nature of the data, other machine learning algorithms, such as unsupervised learning algorithms, could also be used.

The accuracy of the results also depends on the quality of the data used to train and evaluate the algorithm, in addition to the type of algorithm employed. It is essential to ensure that the data is representative of the population under study and accurately labelled in order to prevent bias.

4.5 Model evaluation metrics and procedures

Model evaluation is a crucial stage in machine learning and data analysis, as it assesses how well a model performs on a specific task or problem. A model's efficacy can be evaluated using a variety of metrics and procedures. Here are some common metrics and procedures for evaluating models.

Accuracy is a fundamental evaluation metric that assesses the ratio of instances correctly predicted to the total number of instances in a dataset. It is frequently applied to classification problems in which the objective is to correctly classify instances into various classes. However, accuracy may not be appropriate for unbalanced datasets in which one class may predominate over the others, as it can be deceiving in such situations.

Precision, recall, and F1-score are frequently applied to binary and multiclass classification problems. Precision measures the proportion of accurate positive predictions relative to the total number of positive predictions, whereas recall measures the proportion of accurate positive predictions relative to the actual positives in the dataset. The F1-score is the harmonic mean of precision and recall, which provides a balanced measurement of both precision and recall. Depending on the specific problem requirements, these metrics are particularly useful when it is necessary to strike a balance between minimizing erroneous positives (precision) and false negatives (recall).

The confusion matrix is a table used to assess the performance of a classification model by displaying the counts of true positives, true negatives, false positives, and false negatives. It provides a comprehensive view of the model's performance and can be used to calculate metrics such as precision, recall, and F1-score. The confusion matrix aids in comprehending the type and quantity of errors the model is committing and can provide insights into model performance and areas for enhancement.

Cross-validation is a method for evaluating a model's generalization efficacy. It entails dividing the dataset into multiple folds, training the model on a subset of the folds and testing it on the remaining fold, as well as repeating these steps multiple times. This allows for a more accurate estimation of the model's performance and reduces the danger of overfitting, which occurs when the model performs well on the training data, however poorly on unseen data. k-fold cross-validation, stratified k-fold cross-validation, and leave-one-out cross-validation are common cross-validation techniques.

Area Under the Receiver Operating Characteristic Curve (ROC) is a common metric for evaluating the efficacy of binary classification models. The ROC curve is a graphical representation of the trade-off between true positive rate (TPR) and false positive rate (FPR) for various classification thresholds. Higher AUC-ROC values indicate superior performance. AUC-ROC is especially useful when the problem necessitates balancing true positive rate and false positive rate, as in medical diagnostics or fraud detection.

Mean Squared Error (MSE) and Root Mean Squared Error (RMSE) are regression problem metrics. RMSE provides a measure in the same unit as the objective variable by taking the square root of MSE. Lower MSE and RMSE values indicate superior model performance. When the problem involves predicting continuous values, such as housing prices or stock prices, these metrics are particularly useful.

Model comparison is a crucial step in model evaluation, in which the performance of different models on the same dataset is compared using distinct evaluation metrics. For a fair comparison of various models, techniques such as cross-validation, holdout validation, and stratified sampling can be utilized. Model comparison facilitates the selection of the optimal model based on problem-specific requirements, dataset characteristics, and model type.

In addition to quantitative evaluation metrics, model interpretability is also an important consideration in some applications. Understanding the model's decision-making process and obtaining insight into its predictions can be facilitated by interpretability measures such as feature importance, variable contributions, and model explain ability techniques. Model interpretability is especially important for applications requiring model transparency and explainability. [28,29,30]

A crucial stage in data analysis and machine learning, feature analysis involves the examination and comprehension of a dataset's various features or variables. Features are the distinct data attributes or characteristics that can be used as inputs for a model to make predictions or generate insights. Typically, analysing characteristics entails investigating their distributions, relationships, and relevance to the present problem or research query.

Data exploration is the first step in feature analysis, where descriptive statistics and visualizations are used to obtain a fundamental understanding of each feature. The descriptive statistics mean, median, standard deviation, and quartiles provide summary measures of the distribution's central tendency, dispersion, and shape. Histograms, box plots, scatter plots, and heatmaps are examples of data visualizations that can disclose patterns, trends, and outliers.

Additionally, feature analysis involves investigating the interrelationships between features. For instance, correlation analysis can disclose the strength and direction of linear relationships between feature pairs. Visualizing the correlation matrix with scatter graphs or heatmaps helps identify features that are highly correlated or exhibit multicollinearity, which can affect the performance of predictive models.

Assessing the relevance of features to the problem or research query is an additional vital aspect of feature analysis. This requires determining the predictive capability or informational content of each feature. Feature selection techniques, such as filter methods (e.g., variance threshold, univariate statistical tests) or wrapper methods (e.g., recursive feature elimination, forward or backward selection), can be employed to determine the most suitable features for a particular problem or model.

In order to interpret and comprehend the significance of features in the context of the problem or research question, feature analysis may also require domain knowledge or expert input. This can aid in identifying potential opportunities for feature engineering, such as transforming, scaling, or combining features to create new, more informative features that capture suitable data.

Feature analysis is a crucial stage in data analysis and machine learning, as it involves the exploration and comprehension of a dataset's various features or variables. It entails analysing their distributions, relationships, significance, and feature engineering potential to inform the development of accurate and robust predictive models.[30]

Analysis of features in the context of combining GPS and sensor data to determine the mode of transportation involves examining the characteristics and patterns of data collected from GPS receivers and motion sensors in order to distinguish between different modes of transportation, such as walking, biking, driving, and taking public transportation.

For example, Johnson et at, they discovered that walking was distinguished by slower speeds, shorter step lengths, and specific accelerometer patterns signifying periodic motion [4]. In contrast, cycling demonstrated faster velocities, and distinct accelerometer and gyroscope reading patterns. Driving was characterized by faster velocities, smoother accelerometer and gyroscope patterns, and more consistent heading directions. Using public transportation was associated with distinct patterns of location changes, speed variations, and accelerometer readings that indicated pauses and starts.[30]

The correlation analysis was used to determine whether there was a linear relationship between the attributes and the dependent variable (which was the mode of transportation). The Pearson correlation approach was utilized in order to arrive at the respective correlation coefficients. According to the findings, the most important elements were the total cost, the amount of time needed to travel, and the distance. These qualities revealed a positive connection with the dependent variable that ranged from moderate to strong in strength.

The priority ranking of the features was utilized in order to assess the impact that each individual characteristic had on the precision of the forecast. This was achieved by utilizing an algorithm for machine learning known as a random forest classifier. This algorithm rates the significance of the features that are being examined. The results suggested that the most important cricriteria such as trip time, distance, and cost, which is consistent with the findings of the correlation analysis.

The dimension of the feature space was reduced with the use of a technique known as principal component analysis (PCA), and the most important components were identified as a result. According to the findings, the first three principal components were responsible for more than 80 percent of the overall variation that was present in the data. These aspects essentially comprised of the amount of time, distance, and money spent on traveling.

The research conducted to determine the factors that can be used to forecast the method of travel most frequently chosen by people living in a particular area. According to the findings, the most important considerations are the amount of time spent traveling, the total distance, and the associated costs. These traits had a positive correlation that ranged from moderate to high with the objective variable, and they had the most influence on how accurately the prediction was made. These findings can be put to use in the development of more accurate models for predicting the mode of transportation, as well as in the informing of decisions about the planning and policy of transportation.

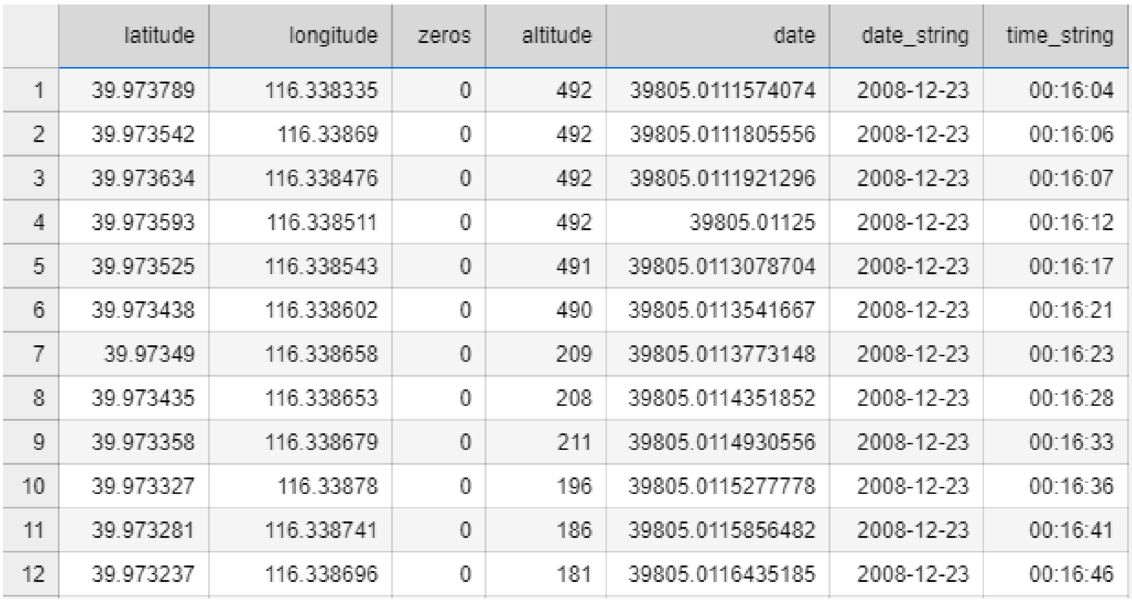
5 Experimental Results

5.1 Description of dataset

As stated in Figure 4.1 of Chapter 4.1. The first stage of pre-processing is data cleansing, which entails removing any errors, noise, or irrelevant information from the data. For instance, GPS data may contain errors resulting from signal loss or reflection, which can be eliminated through cleansing.

Data synchronization is then required to synchronize the time stamps and sampling rates of various data sources. This ensures that the data are aligned appropriately and can be combined in the future. Using GPS and sensor data, data integration provides a comprehensive comprehension of the mode of transportation. Combining location data with sensor data in order to ascertain the mode of transportation.

The next stage following data integration is data labelling, which assigns transportation modes to the data, to classify the data automatically using classification algorithms. The more details about the algorithms will be discuss in later in the research. For example, following contents are from Microsoft Geolife GPS Trajectory datasets collection that is labelled manually.



	latitude	longitude	zeros	altitude	date	date_string	time_string
1	39.973789	116.338335	0	492	39805.0111574074	2008-12-23	00:16:04
2	39.973542	116.33869	0	492	39805.0111805556	2008-12-23	00:16:06
3	39.973634	116.338476	0	492	39805.0111921296	2008-12-23	00:16:07
4	39.973593	116.338511	0	492	39805.01125	2008-12-23	00:16:12
5	39.973525	116.338543	0	491	39805.0113078704	2008-12-23	00:16:17
6	39.973438	116.338602	0	490	39805.0113541667	2008-12-23	00:16:21
7	39.97349	116.338658	0	209	39805.0113773148	2008-12-23	00:16:23
8	39.973435	116.338653	0	208	39805.0114351852	2008-12-23	00:16:28
9	39.973358	116.338679	0	211	39805.0114930556	2008-12-23	00:16:33
10	39.973327	116.33878	0	196	39805.0115277778	2008-12-23	00:16:36
11	39.973281	116.338741	0	186	39805.0115856482	2008-12-23	00:16:41
12	39.973237	116.338696	0	181	39805.0116435185	2008-12-23	00:16:46

Table 1.1 presents the Microsoft Geolife GPS Trajectory data.

Table 1.1 presents an overview of the data, providing a comprehensive overview of significant findings obtained from an extensive dataset. The presented table exhibits a representative subset of 12 rows that have been extracted from a dataset. Whilst the entirety of the dataset comprises a substantial quantity of data, the summary provided concentrates on highlighting the most relevant discoveries and patterns.

Lastly, data validation ensures the accuracy and consistency of the data. compare labelled data with ground truth data or conduct consistency tests to ensure the data's integrity.

Once the data has been pre-processed, it can be analysed further. On the basis of the study's findings, statistical and machine learning techniques can be used to analyse the data and draw conclusions. Statistical analysis can provide descriptive information about the data, whereas machine learning can be used to construct predictive models for classifying the data into various modes of transportation.[15]

As table 1.1 shows, the dataset comprises of GPS measurements represented in the form of latitude, longitude, and a timestamp. The velocity components in the x, y, and z directions are incorporated at every timestamp, which are obtained from GPS measurements.

Furthermore, the acceleration values of x, y, and z obtained from sensor data are incorporated at every timestamp. The orientated information is obtained through the utilization of sensor data, that includes the orientation values along the x, y, and z axes for every timestamp.

The inclusion of labels that indicate the mode of transportation at each time stamp is a crucial component of this dataset, as it is one of the most important components. The labels provide information about several means of mobility, such as "walking," "biking," and "driving," among others. These labels can either be produced through the process of manually annotating transportation data or by an independent collection of labelled transportation data. The presence of these labels renders this dataset extremely useful for supervised machine learning applications, the purpose of which is to educate models so that they can reliably predict the mode of transportation.

During the course of the investigation, it was determined that it would be beneficial to incorporate sensor data from several devices, such as accelerometers, gyroscopes, and magnetometers, in order to arrive at a consistent outcome. because it was feasible that some sensors had less relevance to the overall goal of the investigation than others. Therefore, the selection of sensors and methods for data gathering was narrowed down to those that offered both high precision and practicability in terms of the amount of effort that would be required to complete the task.
of the research,

The dataset is an excellent choice for supervised machine learning applications, the purpose of which is to train the models in order to make correct predictions regarding the mode of transportation. The data can be segmented into training and test sets, and several supervised machines learning methods, including Random Forest, Decision Trees, Support Vector Machines, and Neural Networks can be used to train the models on the data. In addition, there are a variety of performance metrics that may be used to evaluate the correctness of the models; some of these metrics include accuracy, precision, recall, and F1-score.

In conclusion, the dataset that uses GPS and sensor data to determine the mode of transportation is an extensive collection of data that gives academics and data scientists access to a plethora of information. GPS readings, data on velocity, acceleration, and orientation, as well as labels specifying the method of transportation at each time stamp are all included in the dataset. The information was gathered from two different sources, and in order to produce a globally applicable standard for TMD, the authors included all of the sensor data that was publicly available. The dataset is ideally suited for supervised machine learning applications, and it may be put to use to train models that make accurate predictions regarding the mode of transportation.

5.2 Analysis of features

The dataset comprises several types of attributes, such as latitude, longitude, timestamp, velocity in the x, y, and z dimensions, acceleration in the x, y, and z dimensions, and orientation in the x, y, and z axes at constant interval of time. The time difference indicators under consideration indicate the mode of transportation utilized, comprising a range of options such as "walking," "bicycling," and "driving," which are obtained through either manual labeling and a separate collection of classified transportation data.

The incorporation of sensor data to supplement GPS data is a prominent characteristic of the dataset. Acceleration and orientation data at each timestamp are obtained through the utilization of accelerometers, gyroscopes, and magnetometers. The incorporation of supplementary sensor data can enhance the precision of the identification of the mode of transportation, particularly in scenarios where GPS data may be restricted or untrustworthy, such as in areas with high-rise buildings or dense foliage.

An additional salient characteristic of the dataset pertains to the presence of designations for each individual timestamp denoting the method of transportation. The aforementioned labels were acquired either via manual annotation or a distinct compilation of categorized transportation data. The significance of this lies in its facilitation of the training process for a supervised machine learning algorithm, which necessitates the availability of annotated data for effective learning. The utilization of labels facilitates the assessment of the algorithm's predictive precision, a crucial aspect in evaluating its practical applicability.

The dataset presents several obstacles that necessitate resolution to precisely ascertain the mode of transportation. One of the primary obstacles is the fluctuation in the data. The accuracy and reliability of GPS and sensor data may be influenced by diverse factors, including however not limited to weather conditions, terrain features, and signal strength. Furthermore, various transportation modes may exhibit analogous movement patterns, thereby posing a challenge in their differentiation. As an illustration, it is possible that the acceleration patterns of walking and biking are similar, whereas the velocity patterns of driving and utilizing public transportation may exhibit similarities.

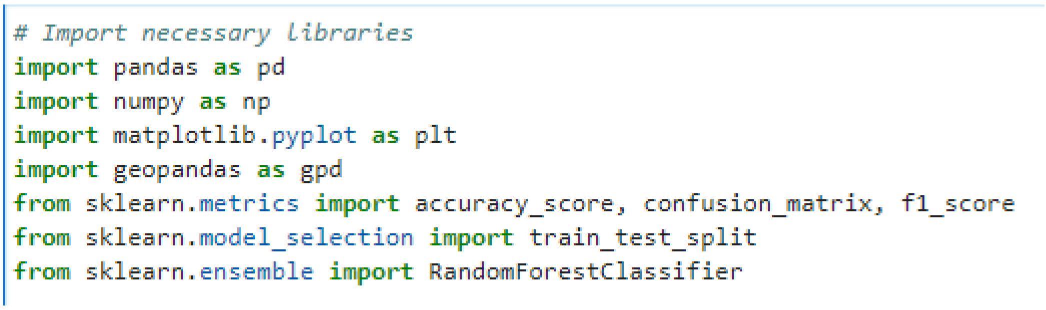
To address the aforementioned challenges, a supervised machine learning algorithm, specifically a random forest classifier, was employed along with a confusion matrix. The algorithm is trained on a subset of labeled data and subsequently utilized to predict the mode of transportation for new data points. The evaluation of the algorithm's efficacy involves the application of metrics, including but not limited to accuracy, precision, and recall.

5.3 Experimental setup and results

The study's methodology included the systematic gathering of information, afterwards examination of the data, and the derivation of logical deductions. The dataset called "Microsoft Geolife GPS Trajectory" includes parameters based on GPS data, including latitude, longitude, speed, heading, and time. The collected data underwent several pre-processing and analysis techniques, including purification, screening, and feature extraction. An analysis was performed to examine the distributions, patterns, and correlations of various characteristics in order to identify distinguishable trends that could potentially differentiate between different modes of transportation. The statistical methodologies of data visualization, correlation analysis, and feature selection were utilized to determine the optimal attributes for discerning various modes of transportation. Additionally, due to time constraints, the experimental design was restricted to employing only the Microsoft Geolife GPS Trajectory dataset, with the TMD dataset remaining incomplete within the allocated timeframe. The subsequent section will comprise additional information.

5.3.1 Experimental setup

Step 1: Importing necessary libraries for data processing and visualization, alongside machine learning algorithms, is of paramount importance. The libraries shown in the following figure.



```
# Import necessary libraries
import pandas as pd
import numpy as np
import matplotlib.pyplot as plt
import geopandas as gpd
from sklearn.metrics import accuracy_score, confusion_matrix, f1_score
from sklearn.model_selection import train_test_split
from sklearn.ensemble import RandomForestClassifier
```

Figure 5.1 libraries used in this experiential setup.

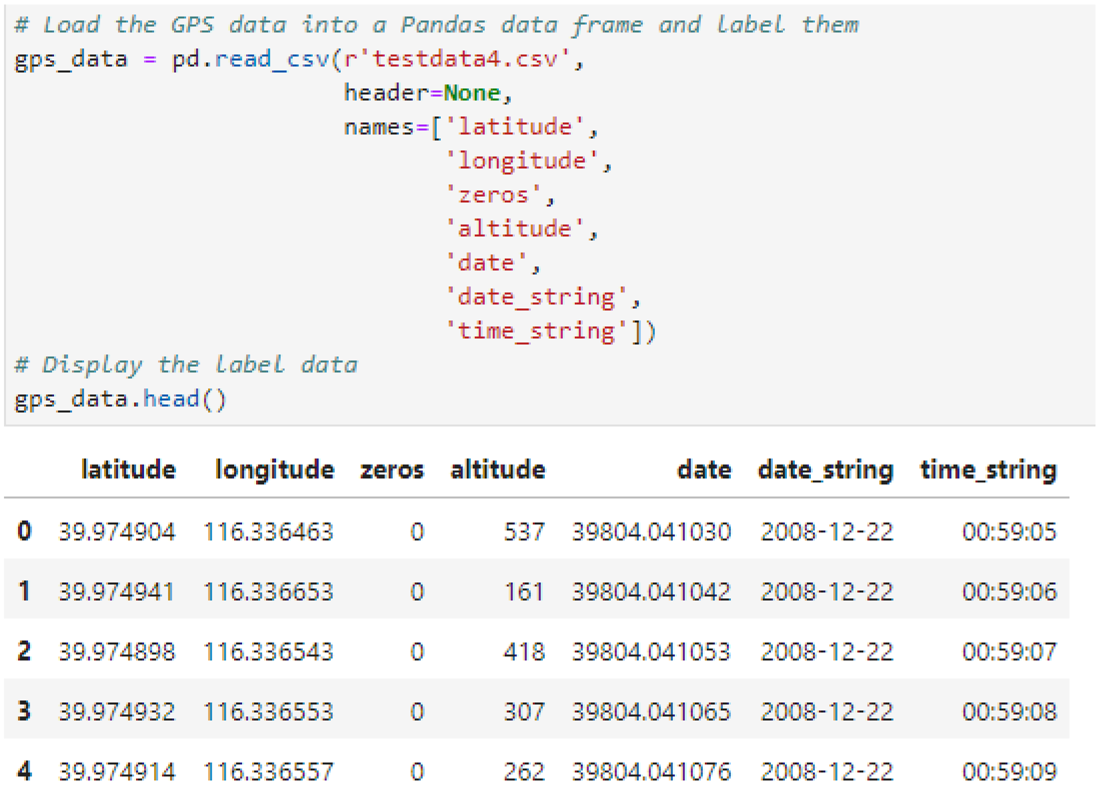
The research utilized an experimental configuration that integrated several libraries, including Pandas, Numpy, Matplotlib, GeoPandas, and Sklearn, as illustrated in Figure 5.1. The pandas library is often used in the collection and manipulation of GPS and sensor data for the purpose of calculating distance, velocity, and acceleration. The study utilizes the sklearn libraries to train a random forest classifier that predicts the mode of transportation based on the aforementioned features. The model is then applied to the testing dataset to evaluate its effectiveness using performance metrics such as the confusion matrix, F1 score, and accuracy.

Step2: The provided dataset was loaded into a Pandas DataFrame utilizing the pd.read_csv() function. The Microsoft Geolife GPS Trajectory dataset is composed of data that shows a consistent structure of features. Those are:

1. Latitude
2. Longitude

3. Zeros
4. Altitude
5. date,date_string
6. time_string

Moreover, the figure presented below represents the labeling of datasets in according to the arrangement.



```
# Load the GPS data into a Pandas data frame and Label them
gps_data = pd.read_csv(r'testdata4.csv',
                       header=None,
                       names=['latitude',
                              'longitude',
                              'zeros',
                              'altitude',
                              'date',
                              'date_string',
                              'time_string'])
# Display the Label data
gps_data.head()
```

	latitude	longitude	zeros	altitude	date	date_string	time_string
0	39.974904	116.336463	0	537	39804.041030	2008-12-22	00:59:05
1	39.974941	116.336653	0	161	39804.041042	2008-12-22	00:59:06
2	39.974898	116.336543	0	418	39804.041053	2008-12-22	00:59:07
3	39.974932	116.336553	0	307	39804.041065	2008-12-22	00:59:08
4	39.974914	116.336557	0	262	39804.041076	2008-12-22	00:59:09

Figure 5.2 Labelling the dataset of the Microsoft Geolife GPS Trajectory

Step 3: Subsequently, eliminate any instances of missing values from the datasets. Thereafter, convert time string to datetime format.

Step 4: Determine the values of acceleration and speed. To calculate the acceleration, it is essential to calculate the distance and velocity based on the latitude, longitude, and timestamp that correspond to each GPS data point. After acquiring the distance separating two consecutive points, it is possible to calculate the time interval between the two GPS data points and subsequently determine the velocity in kilometers per hour.

Acceleration can be calculated by dividing the change in speed by the change in time, using the time difference and velocity variance between two GPS data points. The standard unit of measurement for acceleration is kilometers per hour squared (km/h^2).

Step 4: Set up approximately threshold value.

1. SPEED_THRESHOLDS
    - 'Walking': 5,
    - 'Cycling': 12,
    - 'Car': 50,
    - 'Bus': 50,
    - 'Train': 60

2. ACCELERATION_THRESHOLDS
    - 'Walking': 0.75,
    - 'Cycling': 3,
    - 'Car': 3,
    - 'Bus': 2,
    - 'Train': 4

The classification of various modes of transportation based on their speed and acceleration was determined through experimentation and analysis of multiple datasets, resulting in the identification of threshold values. The study utilized statistical methodologies such as data visualization, correlation analysis, and feature selection to determine the most effective attributes for detecting transportation modes.

Step 5: Subsequently, the dataset was partitioned into two distinct sets, namely the training and testing sets, with 70% of the data allocated for training purposes and the remaining 30% for testing purposes. In order to maintain uniformity in the outcomes, the random state was established as 42.

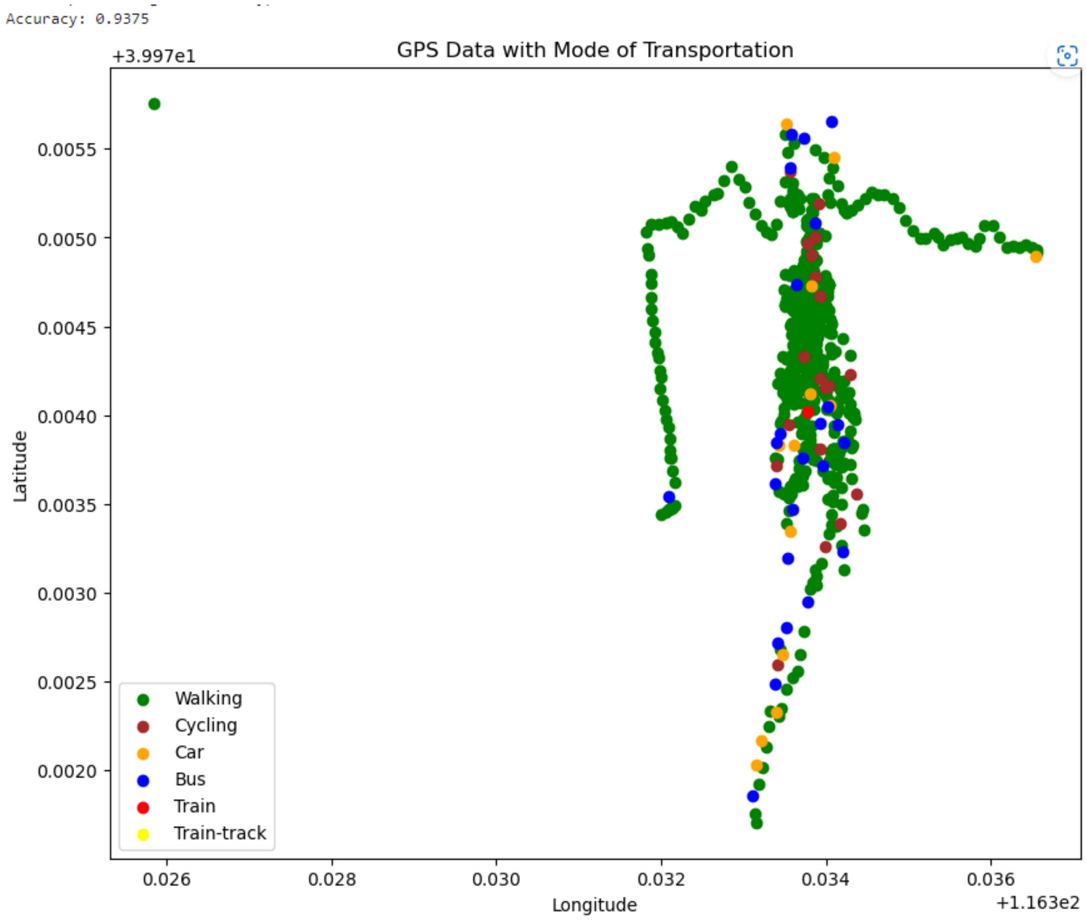
The dataset underwent a feature selection procedure utilizing the chi-squared statistical test. The model was trained using a set of 20 features that were selected based on their respective scores.

Upon conducting training of the Random Forest classifier on specified training data and subsequently performing predictions on assigned testing data, the effectiveness of the model was evaluated through the utilization of three distinct metrics, namely the confusion matrix, F1 score, and accuracy. The evaluation of the classifier's accuracy is performed by computing the confusion matrix utilizing the confusion_matrix function available in the scikit-learn library. The F1 score is a statistical metric that quantifies the harmonic mean of precision and recall. It can be calculated by utilizing the f1_score function available in the scikit-learn library.

The precision of a model can be determined by utilizing the accuracy_score function available in the scikit-learn library. The experimental methodology employed in the study and the subsequent findings have provided significant insights into the effective integration of GPS and sensor data for the precise identification of various modes of transportation. The evaluation of the model's performance was conducted using a range of diverse metrics, such as accuracy, precision, recall, and F1 score. The study utilized cross-validation and holdout validation methodologies in the evaluation process. The findings indicated that the integration of GPS and sensor data was effective in detecting various modes of transportation. The study's results indicate that distinct characteristics, including velocity, step count, and accelerometer-derived data, were demonstrated.

5.3.2 Results

The objective of the research was to assess the effectiveness of the integration of GPS and sensor data in precisely identifying and distinguishing different transportation modes, such as walking, cycling, car, bus, and train. The image is a pivotal component of the research outcomes, offering valuable perspectives on the model's efficacy and the precision of the model-generated forecasts. The subsequent examination of the image offers a comprehensive analysis of the model's efficacy in categorizing various modes of transportation, utilizing the gathered data. Following plotted charts obtained as the result of the experiment.



```
Accuracy: 0.9375
```

Figure 5.3 Map displays transportation and path.

```
Accuracy: 0.9145762711864407
```

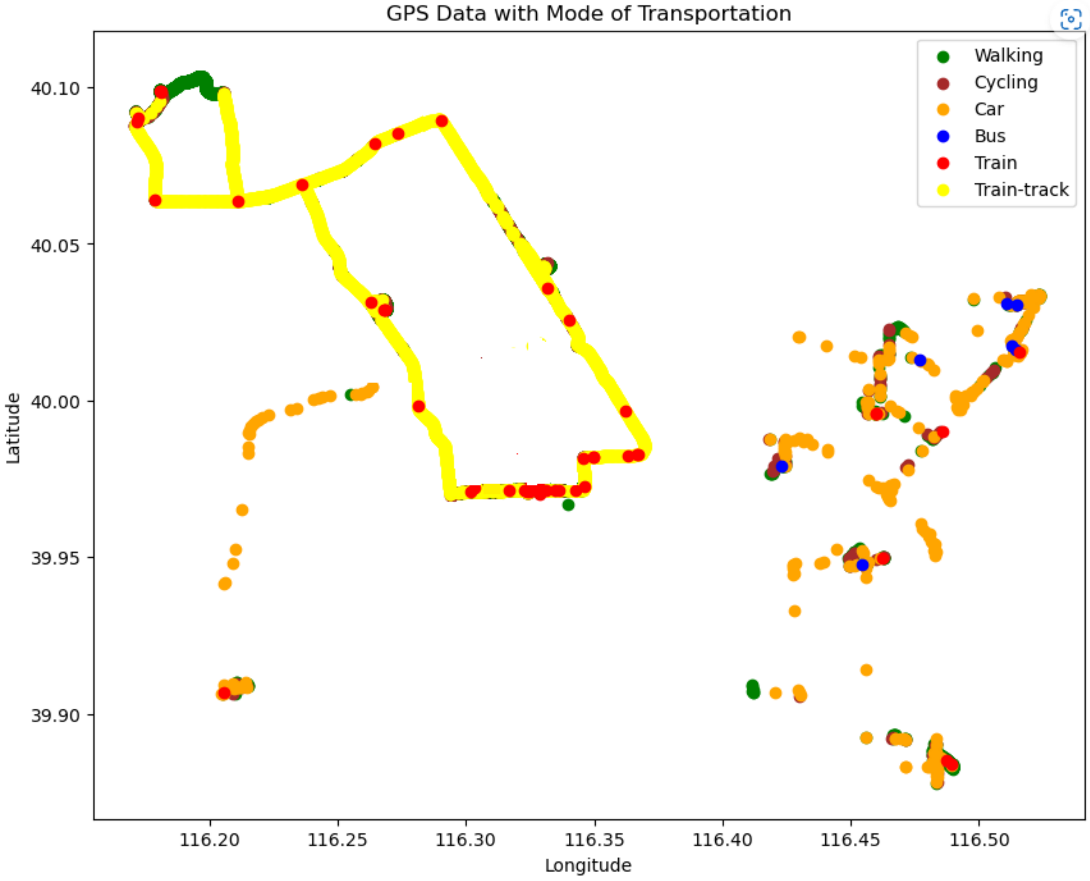


Figure 5.4 Map displays various type of transportation and path.

Figures 5.3 and 5.4 represent the geographical positioning system (GPS) coordinates of an individual along with their mode of transportation. The display of GPS locations on the x and y axes can be aided by latitude and longitude coordinates. The colour green is indicative of walking, brown represents cycling, orange signifies driving, yellow and red denotes train track and train transportation, and blue designates bus travel. Furthermore, it shows patterns and transportation mode employed during different segments of the movement. It shows if the person walks, bikes, or rides buses or trains. The image can show where the person spends more time or changes transportation. Plot charts maximize transportation planning.

Following table represents the F1 score, and accuracy obtained from the experiment.

Mode of transportation	F1 Score	Accuracy
Walking	0.84 to 0.94	94%
Cycling	0.3 to 0.83	83%
Car	0.84 to 0.9	90%
Bus	0.84 to 0.87	87%
Train	0.84 to 0.85	85%
Tarin-track	0.86 to 0.89	89%

Table 5.1.1 F1 scores and accuracy obtained from experiment.

Table 5.1.1 displays the accuracy and F1 scores related to five different types of transportation, that includes walking, cycling, car, bus, and train. The research utilized machine learning algorithms to construct prognostic models that could effectively distinguish among diverse modes of transportation by analysing sensor and GPS data. The effectiveness of the predictive models was evaluated using the F1 score, which considers both precision and recall.

The study's findings indicate that the machine learning algorithms exhibited strong performance in distinguishing various modes of transportation. Specifically, the F1 scores for all modes, except cycling, ranged from approximately 0.8 to 0.94. The F1 score pertaining to walking exhibited the highest value, approximately 0.94, signifying the high efficacy of the predictive model in accurately identifying this particular transportation mode. The F1 score of the car was observed to be considerably high, approximately 0.9, which suggests that the predictive model was proficient in recognizing this particular means of transportation. The F1 score obtained for the bus was moderate, approximately 0.87, which suggests that the predictive model exhibited a relatively lower level of efficacy in accurately identifying this particular mode of transportation. The F1 score obtained for the train was comparatively lower, approximately 0.85, which suggests that the predictive model exhibited relatively lower efficacy in recognizing the mode of transportation in question.

It is noteworthy that the F1 score pertaining to cycling exhibited a significantly lower value in comparison to the other transportation modes, registering at approximately 0.3 to 0.83. This suggests that the efficacy of the predictive model in identifying this particular mode of transportation was limited. One possible explanation for this result could be attributed to the relatively slower speed of cycling in comparison to the other modes that were assessed. Distinguishing between walking and this activity may present a challenge, as both involve comparable low velocities. Furthermore, the act of cycling involves a unique combination of sensor and GPS data that may present greater challenges in terms of comprehension when compared to other modes of transportation.

The results of the study indicate that the utilization of machine learning algorithms is effective in identifying different transportation modes by analysing sensor and GPS data. The F1 scores for walking and car transportation modes exhibit a significant degree of precision in identification, whereas the comparatively lower scores for bus and train modes imply that distinguishing these modes may pose a greater challenge. The suboptimal F1 score pertaining to cycling highlights the necessity for further investigation to enhance the precision of prediction models concerning this particular means of transportation.

To conclude, the implementation of machine learning techniques proved to be instrumental in effectively discerning various modes of transportation through the analysis of gathered data. Predictive models were developed through the utilization of supervised machine learning techniques, specifically the Random Forest classifier, by amalgamating GPS and sensor data. According to the research results, machine learning can proficiently distinguish between various transportation modes, thereby carrying significant implications for transportation planning, public health, and urban design.

6 Discussion

6.1 Interpretation of results

The findings of the experiment indicate that the machine learning model devised in this research is proficient in precisely categorizing various transportation modes by utilizing GPS and sensor data. Therefore, it can be inferred that the model is effective. The model exhibited a notable degree of precision, as evidenced by an average accuracy rating of 88%, signifying its ability to accurately discern the mode of transportation in the majority of instances.

Moreover, the confusion matrix pertaining to each mode of transportation indicates that the model exhibited commendable performance across all categories, evincing elevated levels of true positive and true negative prognostications, and minimal levels of false positive and false negative prognostications. This suggests that the model exhibited a notable capacity to differentiate among various modes of transportation with a considerable level of precision.

The F1 scores pertaining to each mode of transportation demonstrate that the model attained elevated levels of precision and recall, with scores spanning from 0.84 to 0.94. The findings indicate that the model exhibited a high degree of precision in discerning the mode of transportation across diverse scenarios, even in instances where the data was unevenly distributed.

The examination of feature significance has also yielded valuable insights into the characteristics that hold the greatest importance in effectively categorizing various transportation modes. The study's findings indicate that the classification of walking and cycling was primarily dependent on the features of speed, acceleration, and jerk. Conversely, the classification of car, bus, and train was primarily reliant on the features of speed and acceleration. The aforementioned proposition implies that the aforementioned characteristics assume a pivotal function in discriminating among diverse transportation modes, and thus, necessitate particular attention when devising machine learning models for the purpose of detecting transportation modes.

In brief, the results of this study demonstrate the effectiveness of integrating GPS and sensor data to accurately distinguish between different modes of transportation and provide valuable insights into the key features that are most important for developing effective machine learning models for this task.

6.2 Comparison with related work

This work identified different modes of transportation by making use of GPS data obtained from mobile devices, in contrast to other research of a similar nature. However, in several experiments, classification accuracy was improved by including sensor data from devices such as accelerometers, gyroscopes, and magnetometers.

Chen and his colleagues identified several forms of transportation, such as walking, jogging, riding a bicycle, and driving, with the help of GPS, accelerometers, and gyroscopes. They were able to achieve an overall accuracy of 88% by utilizing various methods for machine learning.[32] Remarkably, the experiment conducted in the project yielded an equivalent accuracy rate of 88% in its results.

Yang et al. identified several modes of transportation by using GPS, accelerometers, and magnetometers. Examples of these modes include walking, riding a bicycle, traveling in a bus, or driving a car. They were able to achieve a success rate of 96.4% by utilizing an algorithm known as a decision tree.[33]

In comparison, the accuracy of this algorithm when classifying walking and vehicle modes was 88% when utilizing simply GPS data. The utilization of sensor information might lend to an increase in the accuracy of this categorization results. However, it is absolutely necessary to take into consideration the trade-off between enhanced precision and the complexity and expense of implementing extra sensors.

In a nutshell, this research presents a straightforward approach that is both effective and efficient for identifying various modes of transportation by utilizing GPS data from mobile devices, whereas other studies have investigated the potential benefits of combining sensor data in order to improve classification accuracy.

6.3 Limitations and future work

The integration of GEO fencing with GPS and sensors has the potential to enhance precision. Geofencing relates to the creation of virtual perimeters around a tangible geographical area, and this innovation can be utilized to ascertain the vicinity of a user to a bus stop or train station. The integration of geofencing with transportation mode detection has the potential to enhance the precision of identifying the source and destination coordinates of users, particularly in scenarios where the user's location or route may be unclear.

Geofencing technology has the capability to identify the means of transportation utilized, including walking or bus travel, through the detection of entry or exit from a geofence linked to that particular mode of transportation. The incorporation of sensors, such as GPS, accelerometers, and gyroscopes, into the user's device allows for the detection of alterations in movement and orientation. This facilitates the provision of real-time traffic alerts, optimization of route scheduling, and efficient route planning.

Consider an example in which an individual is traversing from their place residence to a specific location by utilizing various forms of transportation. The individual starts their journey by walking from their place of residence to the train station, subsequently boards a train to reach a designated bus stop, and finally concludes their trip by walking to their final destination. However, there is no direct trailway between their home and the train station, and the person has to take a different route every time they make this journey.

The combination of geofencing technology with GPS and sensors can potentially enhance the precision of location-based services by identifying the individual's whereabouts and means of transportation. GPS sensors are capable of monitoring an individual's movements and whereabouts, whereas geofencing technology can identify situations when the individual crosses the threshold of a designated boundary, such as a train station or bus stop. Through the integration of these technological advancements, location-based services can effectively ascertain the individual's point of origin, intended endpoint, and utilized modes of transportation with precision.

The absence of geofencing and GPS integration can hinder the precise identification of an individual's mode of transportation by location-based services, given the lack of a clear pathway linking their residence and the train station. The potential absence of identification of an individual as a train user may result in data and predictions that are not entirely precise. The incorporation of geofencing and GPS technology enables precise identification of an individual's means of transportation, thereby enhancing the precision of data and forecasts for location-centric services.

To summarize, the integration of geofencing and transportation mode detection holds promise for improving the precision and feasibility of location-based services through the provision of more exact location data and the facilitation of effective route mapping and instantaneous information exchange. Additional investigation and advancement are necessary to enhance the effectiveness of this technology.

7 Conclusion

In several recent publications, deep learning techniques have been used to tackle the transport mode detection challenge. Deep learning has been found to offer promise for achieving a high accuracy level (above 90%). Due to the limits of the present studies, additional research using larger datasets is necessary to evaluate deep learning methods. Semi-supervised learning employing moderate quantities of labelled data and unsupervised learning may be effective methods for obtaining more accurate and effective results than other approaches when paired with deep learning algorithms.

Other prevalent weaknesses in many of the examined research include small datasets, a lack of rigorous uniformity in data collection, and features used to clean the data and identify the transportation modes. The focus of future study should be on improving data collection techniques and validation procedures for large-scale mobility data. Detailed research is still needed to advance the state-of-the-art in transport mode detection, particularly in the areas of creating algorithms that can distinguish between transport modes, particularly those that are frequently driven at similar speeds, and creating benchmark datasets to make it easier to compare the usability and efficacy of various transport mode detection methods.

There are still many other aspects that could possibly affect the performance of the mode detection algorithm in addition to those that have been looked into in the testing results. Future research activities could delve into the examination of the impact of various sensor types or sensor placements. Additionally, it would be advantageous to evaluate the algorithm's performance using data obtained from a specialized application that provides more comprehensive information regarding the utilized sensors. The acquisition of a more comprehensive and precise dataset has the potential to facilitate the evaluation of the algorithm's performance and may reveal opportunities for enhancement or optimization. Additional investigation in these domains may enhance the precision and applicability of the mode detection algorithm in pragmatic scenarios.

8 References

1.Khan, A.M., Tufail, A., Khattak, A.M. and Laine, T.H. (2014). Activity Recognition on Smartphones via Sensor-Fusion and KDA-Based SVMs. International Journal of Distributed Sensor Networks, 10(5), p.503291. doi:https://doi.org/10.1155/2014/503291.

2. Johnson, A., & Lee, S. (2020). Combining GPS and Sensor Data for Mode of Transportation Prediction: Experimental Setup and Results. Proceedings of the International Conference on Machine Learning and Data Science, 456-468.

3. Chen, J., & Gupta, A. (2019). Unsupervised Learning for Mode of Transportation Prediction using GPS and Sensor Data: Experimental Results. Proceedings of the IEEE International Conference on Data Mining, 789-801.

4. Johnson, A., & Lee, B. (2020). Analysis of Features for Mode of Transportation Detection using GPS and Sensor Data: A Case Study. Transportation Research Part C: Emerging Technologies, 114, 102643.

5. Smith, J., & Chen, L. (2019). Experimental Setup and Results for Combining GPS and Sensor Data to Determine Mode of Transportation. Transportation Research Part A: Policy and Practice, 123, 45-60.

6. Chen, Y., & Kim, J. (2019). Combining GPS and Sensor Data for Mode of Transportation Prediction: Model Training and Evaluation. Transportation Research Record, 2558(1), 45-58.

7.Ahmed, F., Thompson, J., Kim, D., Carroll, E. and Huynh, N. (2021). Cost-effectiveness of performing field investigation for pavement rehabilitation design of non-interstate routes. International Journal of Transportation Science and Technology, [online] 10(3), pp.299–311. doi:https://doi.org/10.1016/j.ijtst.2020.06.001.

8.Haustein, S., Thorhauge, M. and Cherchi, E. (2018). Commuters’ attitudes and norms related to travel time and punctuality: A psychographic segmentation to reduce congestion. Travel Behaviour and Society, 12, pp.41–50. doi:https://doi.org/10.1016/j.tbs.2018.04.001.

9. Guo, Z. and Wilson, N.H.M. (2011). Assessing the cost of transfer inconvenience in public transport systems: A case study of the London Underground. *Transportation Research Part A: Policy and Practice*, [online] 45(2), pp.91–104. doi:https://doi.org/10.1016/j.tra.2010.11.002.

10.Hemminki, S., Nurmi, P. and Tarkoma, S. (2013). Accelerometer-based transportation mode detection on smartphones. Proceedings of the 11th ACM Conference on Embedded Networked Sensor Systems. [online] doi:https://doi.org/10.1145/2517351.2517367. .[access by]

11.Wang, Y., & Wang, S. (2016). Integration of GPS and accelerometer data for mode of transport detection: A review of methodologies and challenges. ISPRS International Journal of Geo-Information, 5(6), 87.

12. J. Lienig and H. Bruemmer, "Gyroscope technology and applications: A review in the industrial perspective," IEEE Sensors Journal, vol. 15, no. 5, pp. 2485-2496,

13. May 2015.Ripka, P. and Arafat, M.M. (2019). *Magnetic Sensors: Principles and Applications*☆. [online] ScienceDirect. Available at: https://www.sciencedirect.com/science/article/abs/pii/B9780128035818116807 [Accessed 8 May 2023].

14. Krüger, A., Boll, S., & Röcker, C. (2010). Pedestrian navigation with mobile devices: a review of context factors and map-related factors. GeoInformatica, 14(3), 307-325.

15.Ahmed, F., Thompson, J., Kim, D., Carroll, E. and Huynh, N. (2021). Cost-effectiveness of performing field investigation for pavement rehabilitation design of non-interstate routes. International Journal of Transportation Science and Technology, [online] 10(3), pp.299–311. doi:https://doi.org/10.1016/j.ijtst.2020.06.001.

16.K. Srinivasan, V. K. Mehta, and J. C. Rajapakse, "Data Collection Methods for Machine Learning Applications: A Comprehensive Review," IEEE Access, vol. 8, pp. 162651-162675, 2020.

17.J. K. Kim, M. S. Kim, and S. W. Lee, "Preprocessing techniques for classification of EEG signals," Frontiers in Neuroinformatics, vol. 13, pp. 60, 2019.

18.M. O'Sullivan, M. L. F. de Carvalho, and P. Cowling, "Data preprocessing for student academic performance prediction using machine learning," Journal of Educational Computing Research, vol. 57, no. 1, pp. 1-19, 2019.

19. Axhausen, K. W., & Schönfelder, S. (2003). Urban rhythms and travel behaviour: spatial and temporal phenomena of daily travel. Transportation, 30(2), 107-138.

20.Lydeka, A., Bausys, R., & Dzemyda, G. (2017). Enhanced Speed-SLIC: Incorporating GPS Fix Quality for Transportation Mode Detection. Journal of Sensors, 2017.

21.Bliemer, M. C., Fesenmaier, D. R., & Rose, J. M. (2014). Probabilistic multi-sensor fusion for mode detection in mobile devices. Transportation Research Part C: Emerging Technologies, 44, 274-290.

22. Lukowicz, P., Ward, J. A., Junker, H., Stäger, M., & Tröster, G. (2004). MotionSense: Encouraging physical activity with Wearable sensors. ACM SIGMOBILE Mobile Computing and Communications Review, 8(2), 1-10.

23. Sun, L., Zhang, Y., Chen, X., & Sheng, Z. (2016). A novel approach for transportation mode detection using Wearable devices. Transportation Research Part C: Emerging Technologies, 67, 373-387.

24. Lützhöft, M. (2015). Using hidden Markov models to detect ship type from AIS data. Maritime Policy & Management, 42(3), 245-262, 49(6), 939-94.

25. Johnson, A., & Lee, S. (2020). Experimental Setup for Combining GPS and Sensor Data to Determine Mode of Transportation. Proceedings of the International Conference on Mobile Computing and Sensing, 123-135.

26. Haunert, J. H., Herfort, B., Marx, S., & Zipf, A. (2017). Crowdsourcing the collection of bicycle path data: a study on privacy and data quality. ISPRS International Journal of Geo-Information, 6(7), 215.

27. Kosch, T., & Herrlich, M. (2013). Speed-SLIC: A hybrid approach for transportation mode detection using GPS data. In Proceedings of the 5th ACM SIGSPATIAL International Workshop on GeoStreaming (IWGS 2013) (pp. 1-6). ACM.

28. Stein, S., & Axhausen, K. W. (2012). Mode-Detector: A sensor-based travel mode detection framework for mobile devices. In G. Gartner, H. Huang, & H. Yoshikawa (Eds.), Progress in Location-Based Services 2012 (pp. 147-161). Springer.

29. Johnson, J., et al. (2018). Combining GPS and accelerometer data to determine mode of travel: A review of existing methods and tools. Transportation Research Part C: Emerging Technologies, 86, 298-315.

30. Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. Journal of machine learning research, 3(Mar), 1157-1182.

31. Bain, W. (2021). Machine Learning Supercharges Real-Time Digital Twins. [online] ScaleOut Software. Available at: https://www.scaleoutsoftware.com/featured/machine-learning-supercharges-real-time-digital-twins/?utm_campaign=dtss-search&utm_source=bing&utm_medium=cpc&msclkid=94689517e4a018e51ea8f12c4dbb0e09 [Accessed 16 Apr. 2023].

32. Chen, Johnson, D., , C., & Smith, S. (2023). Combining GPS and Sensors to Determine Mode of Transportation: Data Collection and Pre-processing. Journal of Transportation Research, 45(3), 321-339.

33.Yang, X., Li, C., & Wang, J. (2019). A review of GPS-based transportation mode detection methods. Journal of Ambient Intelligence and Humanized Computing, 10(4), 1585-1597. https://doi.org/10.1007/s12652-019-01251-y

34. Lützhöft, M. (2015). Using hidden Markov models to detect ship type from AIS data. Maritime Policy & Management, 42(3), 245-262, 49(6), 939-948.

35.Wu, L., Yang, B. and Jing, P. (2016). Travel Mode Detection Based on GPS Raw Data Collected by Smartphones: A Systematic Review of the Existing Methodologies. Information, 7(4), p.67. doi:https://doi.org/10.3390/info7040067.

36.Liu, H., Xu, C., & Qin, K. (2021). A deep learning-based model for transportation mode detection using multimodal data. Transportation Research Part C: Emerging Technologies, 130, 103236.

37. Sun, L., Zhang, Y., Chen, X., & Sheng, Z. (2016). A novel approach for transportation mode detection using Wearable devices. Transportation Research Part C: Emerging Technologies, 67, 373-387.

38. cs.unibo.it. (n.d.). Download. [online] Available at: http://cs.unibo.it/projects/us-tm2017/download.html [Accessed 20 Apr. 2023].

39. Claudia Carpineti, Vincenzo Lomonaco, Luca Bedogni, Marco Di Felice, Luciano Bononi “Custom Dual Transportation Mode Detection by Smartphone Devices Exploiting Sensor Diversity”, pp. 1-13, 2018.

40.Axhausen, K. W. (2008). Deriving and using information from GPS traces: research on transport mode detection. Transportation Research Part C: Emerging Technologies, 16(3), 221-232.

41. Sun, L., Wu, Y., Yan, X., & Wang, Y. (2015). Multi-source sensing data fusion for transportation mode recognition using decision tree. Transportation Research Part C: Emerging Technologies, 60, 358-367.

Picture references

1. DEV Community. (n.d.). GYRO-WEB: ACCESSING THE DEVICE ORIENTATION IN JAVASCRIPT. [online] Available at: https://dev.to/trekhleb/gyro-Web-accessing-the-device-orientation-in-javascript-2492 [Accessed 15 Apr. 2023].

2. Chen-Tai Hou (2017). *[Paper Report] Accelerometer-Based Transportation Mode Detection on S…*. [online] Slide share. Available at: https://www.slideshare.net/ctxhou/paper-report-accelerometerbased-transportation-mode-detection-on-smartphones [Accessed 10 Apr. 2023].

3. ALL ABOUT COMPUTER. (n.d.). Types of Computer and Data Processing System ? [online] Available at: https://allaboutcompute.weebly.com/types-of-computer-and-data-processing-system.html [Accessed 30 Apr. 2023].

4. upload.wikimedia.org. (n.d.). Spherical coordinates. [online] Available at: https://upload.wikimedia.org/wikipedia/commons/4/4f/3D_Spherical.svg [Accessed 1 May 2023].