COMPARISON OF THE PERFORMANCE OF GRADIENT BOOSTING, LOGISTIC REGRESSION, AND LINEAR SUPPORT VECTOR CLASSIFIER ALGORITHMS IN CLASSIFYING TRAVEL MODES BASED ON GNSS DATA

O. Shamohammadi [1], P. Pahlavani[1 *], M.A. Sharifi [1]

[1] School of Surveying and Geospatial Engineering, College of Engineering, University of Tehran, Tehran, Iran
(omid.shahmohammadi, pahlavani, sharifi)@ut.ac.ir

Commission IV, WG IV/3

KEY WORDS: Gradient boosting, Linear support vector classifier, Logistic regression, Streaming GNSS data, Transportation mode.

ABSTRACT:

Public transportation system capacity must be compatible with the frequency of daily trips. Smart mobile phones can collect positioning data at different times, which can detect transportation modes people use for their daily commutes. This information helps the government predict how many vehicles are needed to satisfy public transportation system demands. This article investigates the performance of three different machine learning models, including Gradient Boosting (GB), Logistic Regression (LR), and linear Support Vector Classifier (SVC) in classifying the trip types. Thirty-nine features, including statistical parameters of velocity, acceleration, and jerk, and also parameters representing the time of each trip, are given to the models as input. To increase the performance of the models, with the help of thresholding, points corresponding to noise are detected and removed from the dataset. Moreover, to fill the possible gaps and smooth the trajectories, spline interpolation and Savitzky-Golay filter are also investigated in feature calculation. The results show that the linear models are incapable of distinguishing between different classes well and they are over-fitted to classed with more samples. Hence, the GB by 0.93 recall, precision, and F-score was the best model in determining the vehicle used compared to LR and linear SVC.

1. INTRODUCTION

For many years, regional planning organizations have been investigating daily trips with the aim of obtaining and analysing of information facilitating the management of transportation system such as estimating the amount of demand for public vehicles (Mumford et al., 2002). Knowing the transportation mode helps transportation agencies to utilize appropriate strategies, leading to reductions in durations of trips, traffic, and air pollution. For example, by accurately identifying each user's transportation mode, it becomes possible to provide a more realistic understanding of how many vehicles are needed to move people from a specific place to another each day, which is a great help in reduction of traffic over the roads and motivate people to take public transportation more. Also, High-occupancy vehicle (HOV) lanes could also be introduced.(Grennfelt et al., 2020).

Collecting information related to travel mode was traditionally obtained through written surveys or telephone interviews, which were time-consuming and expensive and usually led to low response rates and incomplete information. Due to the popularity of smartphones and other electronic devices which can measure their positions via different sensors such as Global Positioning System (GPS), Galileo, and GLObal NAvigation Satellite System (GLONASS), these tools have become replacements for traditional methods. Collecting data with these devices is more accurate, cost-effective, and demands fewer human resources. Nowadays, several large datasets are provided by smart devices, such as the dataset of GPS lst2016(Erener & Sarp, 2018) and the Geolife dataset. These datasets contain different trajectories, each consisting of many points with known latitude, longitude, and measurement time. Although many of these devices could collect appropriate datasets, smartphones are more welcomed in this field because a majority of the population in almost all countries use and carry their mobile phones everywhere they go(Busch & McCarthy, 2021).

Although smartphones give no explicit information about the modes of transportation used, by processing their data, some parameters can be extracted to distinguish among the type of vehicles each user takes. For example, the average velocity for a car driver might be much more than the same parameter for a man who walks(Huang et al., 2018).

This paper compares three different machine learning models in terms of classifying vehicles taken by passengers based on navigation data provided by smartphones. The aims of this paper are as follows:

- Developing an integrated method for removing noises from the dataset and extracting features simultaneously,
- Filling the possible gaps in the dataset using spline interpolation,
- Utilizing simple machine learning models which can be processed in normal computers.

The rest of this paper is organized into five sections. While related works are reviewed in the next section, the third one describes the methodology in detail. The fourth section introduces the dataset used, and in the fifth section, the technical details of the implementation scenario and results are expressed. Finally, the last section is dedicated to the conclusion.

2. RELATED WORKS

In (Erdelić et al., 2022), a real-time method of segmenting trajectory based on transport mode change was developed. The

[*] *Corresponding author*

This contribution has been peer-reviewed.
https://doi.org/10.5194/isprs-archives-XLVIII-4-W2-2022-103-2023

transport mode changing points were automatically detected using Transition State Matrices (TSM).

(Nawaz et al., 2020) fused Microsoft Geolife dataset with weather data to analyze how human mobility is affected by geospatial region. For this purpose, they proposed a deep learning-based convolutional long short-term memory (LSTM) model for transportation mode learning. In that research, LSTM was responsible for learning the sequential patterns in the data, while the convolutional neural network duty was extracting high-level features. Their model could improve the accuracy by 3% compared to the benchmark models when it only uses GPS features. Also, the respective improvements of 4% and 7% in accuracy were seen if the impacts of geospatial region and weather attributes were considered.

(Erdelić et al., 2019) used static locations of bus stations, rail lanes, and real-time locations of buses, as well as the GPS data to classify motor and non-motor movement.

Recently, researches are using deep learning methods to solve a variety of problems including those related to computer vision, biomedical, and speech recognition. Deep learning methods are also used to deal with transportation issues. However, by now, the applications developed this domain are too limited(Nguyen et al., 2018). Deep learning methods use non-linear functions in order to transform data from a space to another. Sequential arrangement of these space-to-space functions which forms a deep neural network, can learn complex structures and functions (LeCun et al., 2015).

3. METHODOLOGY

This paper proposes a two-stepped procedure to determine the type of the vehicle a commuter uses based on smartphones' GNSS data. At first, for any trip in the dataset, 39 features were extracted to analyze and recognize the pattern of the trips. In addition to extracting the features, this step also detects and removes noises from the raw data in an iterative process. In the second step, the remaining features are given to three machine learning models that classify trips into five classes. These two steps are described in detail in sections 3.1 and 3.2.

3.1 Step 1: Feature extraction and noise removal

The dataset used has sampled the trajectory at different times. As a result, the data consists of many points represented by coordinates and the measuring time for each point.

Based on these data, 39 features are extracted. These features are mostly statistical parameters about the trajectory's velocity, acceleration, and jerk. Also, as traffic is a function of time as well, weekdays and the average of starting and ending hours are recorded as features. Moreover, the calculated features contain three parameters: bearing angle, Heading Change Rate (HCR), and Stop Rate (SR). These parameters that were used in previous literature and showed their ability to differentiate different travel modes are introduced in the following.

- **Bearing Angle:** This parameter determines the angle between the heading direction of two consecutive points. People who ride bicycles or walk could change their direction more sharply than those sitting in a motorized vehicle. Eqs. (1-4) show how this parameter can be estimated.(Wang et al., 2020)

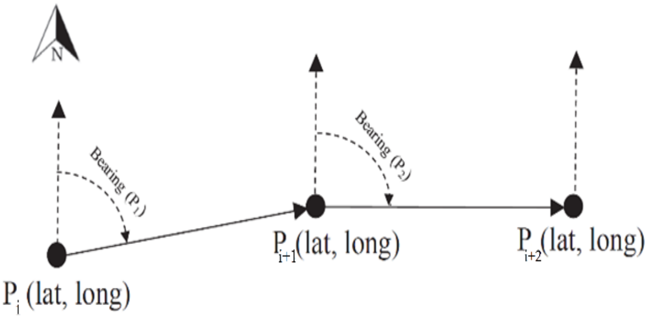


Figure 1. Bearing angles

$$y_i = \sin(lat_{i+1} - lat_i) \times \cos lat_{i+1} \quad (1)$$

$$x_i = \cos\ lat_i\ \times\ \sin lat_{i+1} - \sin lat_i \times \cos lat_{i+1} \times \cos(lat_{i+1} - lat_i) \quad (2)$$

$$B_i = tan^{-1}({}^{y_i}/_{x_i}) \quad (3)$$

$$BR_i = |B_{i+1} - B_i| \quad (4)$$

where P$_i$, P$_{i+1}$, P$_{i+2}$ = Three consecutive points in trajectory
x$_i$, y$_i$ = Different in coordinate of P$_i$, P$_{i+1}$
B$_i$ = Bearing angle between P$_i$, P$_{i+1}$
BR$_i$ = Difference of B$_i$ and B$_{i+1}$

- **HCR:** This parameter counts how many GPS points the bearing angle changed over a certain threshold. This parameter that is introduced by (Zheng et al., 2008) can mostly distinguish between motorized and non-motorized transportation modes. Eq. (5) Shows how this parameter can be determined.

$$HCR = {}^{P_{HC}}/_{D} \quad (5)$$

where P$_{HC}$= Number of GPS points having high change in bearing angle
D = Overall distance of the trajectory

- **SR:** This parameter shows the number of points in which a user's velocity was less than a certain threshold. This parameter can also be beneficial as buses and taxis have to stop much more than private cars. Eq. (6) represents how this parameter can be calculated.

$$SR = {}^{P_{SR}}/_{D} \quad (6)$$

where P$_{SR}$ = Number of GPS points having a low velocity
D = Overall distance of the trajectory

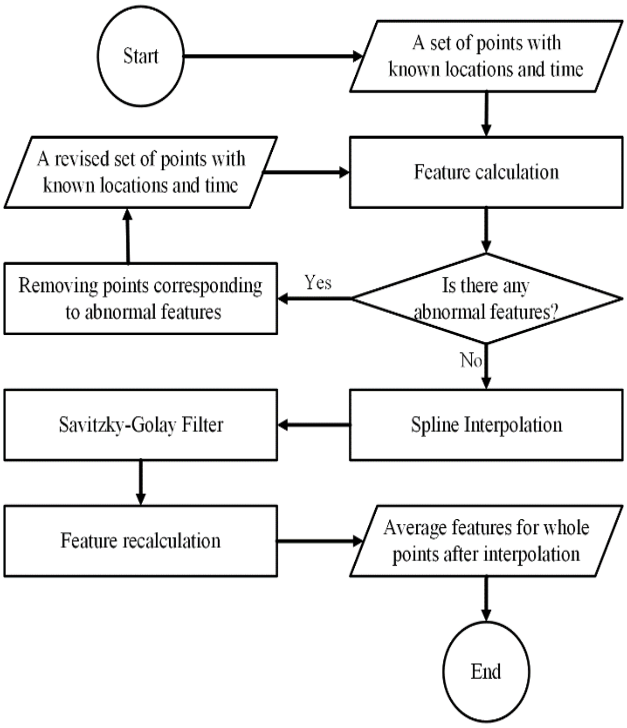

Start
A set of points with known locations and time
Feature calculation
A revised set of points with known locations and time
Is there any abnormal features?
Yes
Removing points corresponding to abnormal features
No
Spline Interpolation
Savitzky-Golay Filter
Feature recalculation
Average features for whole points after interpolation
End

Figure 2. Flowchart of feature extraction and noise removal

As the dataset contains noises affecting the final results, these noises must be removed. Since the velocity and acceleration for each kind of vehicle cannot be more than certain values, abnormal features can be detected. Therefore, the points corresponding to these features are considered noises and are removed. Then, the features are recalculated using the new set of points. This process would be repeated until no other noises are detected. Afterward, if the temporal distance between a pair of consecutive points of a trip is more than a certain value, the gap would be filled by spline interpolation method. Then, the Savitzky-Golay filter is investigated to smooth the final trajectory and then, the parameters are recalculated again. Finally, since the numbers of points in different trips were not the same, for any specific trip, the average values of features calculated among all points were stored as features of that trip. Figure 2 shows the flowchart of this step.

3.2 Step 2: Machine learning classification

In this step, machine learning models are utilized to classify the trips into five classes, namely walk, bicycle, bus, train, and car. Gradient Boosting (GB), Logistic Regression (LR), and linear Support Vector Classifier (SVC) were the classification methods used. Here, each of these methods are introduced briefly.

The first machine learning algorithm used is called Gradient Boosting which is useful for both regression and classification problems. In this method, an ensemble of weak prediction models (base learner) is trained through an iterative process that uses least squares to sequentially fit a base learner to current “pseudo” residuals. These residuals are the gradient of the loss function w.r.t the model values for each training sample during the current step. [friedman2002] This method usually outperforms Random Forest.(Madeh Piryonesi & El-Diraby, 2021), (Piryonesi & El-Diraby, 2020),(Hastie et al., 2009)

Figure 3 depicts a simple diagram of using Gradient Boosting for classification problems.

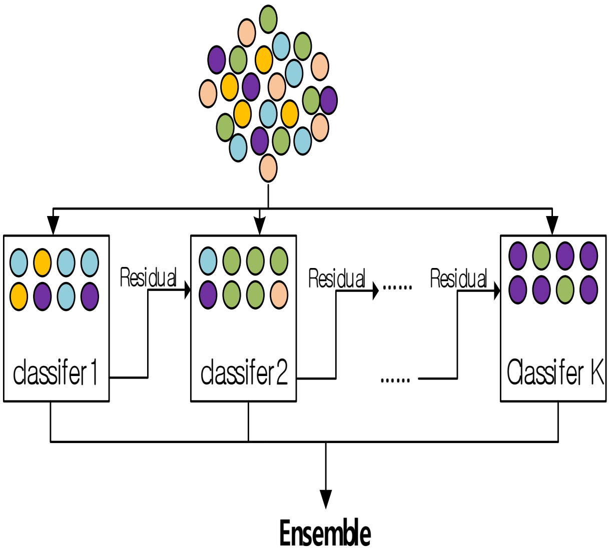


Figure 3. Classification using Gradient Boosting

The second algorithm is Logistic Regression in which a linear function with multiple parameters is usually used. Each input (sample) can have multiple features and each one of these features is multiplied by a corresponding variable (weight) in the linear function. These multiplication terms are then summed together to build the function. These steps are illustrated in Figure 4. In each step, the objective is to fit (train) the weights in a way that a defined cost function becomes minimum. The cost function can usually be the mean square of the distance between the actual predicted outputs for each sample. The predicted part is calculated using the function for each input in the current iteration. (Huang, 2022), (Hosmer Jr et al., 2013), (Sperandei, 2014).

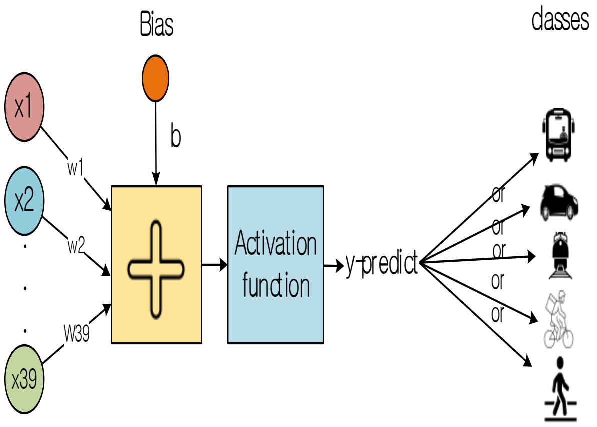


Figure 4. A simple diagram of using Logistic Regression for classifying transportation modes

Looking to the last classifier used, Support Vector Machines (SVM) methods have always been a robust method for regression and classification. When it was used for linear classification with straight lines as kernels, the method is called Linear SVC. In other words, in this method, the hyperplane that is used for classification is a linear condition. Basically, the objective of this method is to distinguish different classes with the help of a margin whose distance from classes is maximum. So, this margin minimizes the classification error. Figure 5 shows a simple example of classification objectives in Linear SVC. Here, for simplification, we have only drawn two dimensions (instead of 39 features). (Suh et al., 2021), (Ajimakin & Devi, 2021), (Ma et al., 2020)

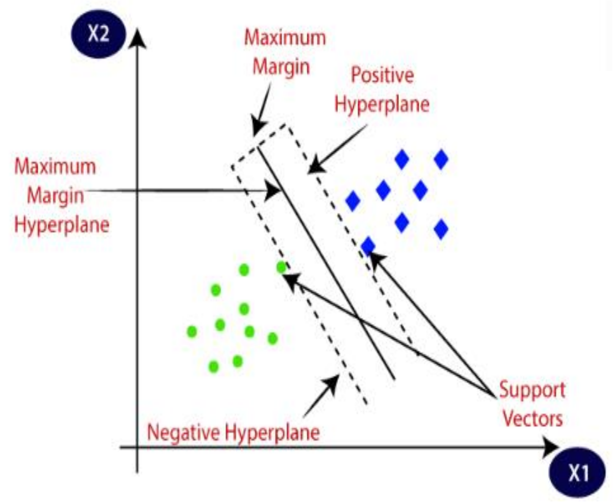


Figure 5. Classification Margins and Hyperplanes in Linear SVC

4. DATASET

In this research, "Geolife" project data was used to evaluate the proposed method.(Zheng et al., 2011), This dataset reflects a wide range of users' commutes, with different destinations including workplaces, stadiums, shopping and entertainment centers. The dataset was collected in more than 30 cities in different countries such as China, USA, and some European countries. Figure 7 shows the distribution of the dataset in Beijing, the city where most of data are created. Also, table 1 provides more details about the dataset. Also, we have observed that our data is unbalanced. In other words, we have more samples for classes "Walk" and "bus", while there are much fewer samples labeled as "Train", "Car", and "Bicycle". Figure 6 shows the distribution of the number of samples for each class within the dataset.

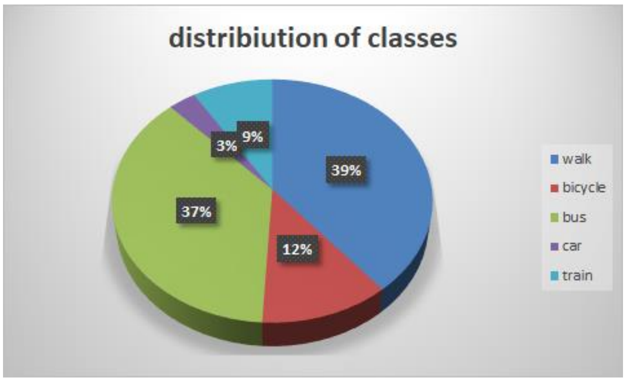


Figure 6. A pie diagram of class distribution in the Geolife dataset after removing noise and extracting features

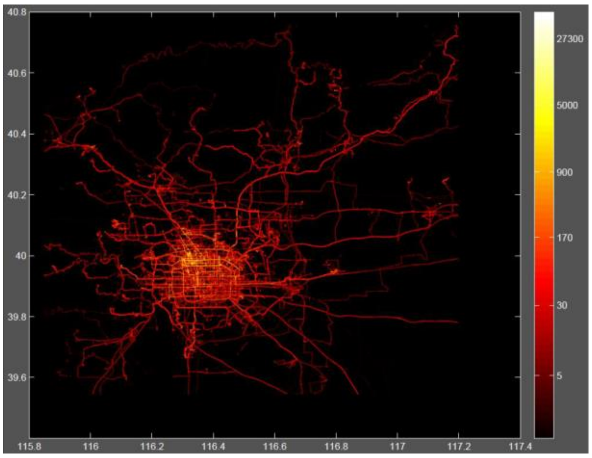


Figure 7. Database distribution in Beijing, China

Number of users	Number of trajectories	Total distance (km)	Total duration (h)
182	17621	1292951	50176

Table 1. Specification about "Geolife" dataset

5. EXPERIMENTAL RESULTS

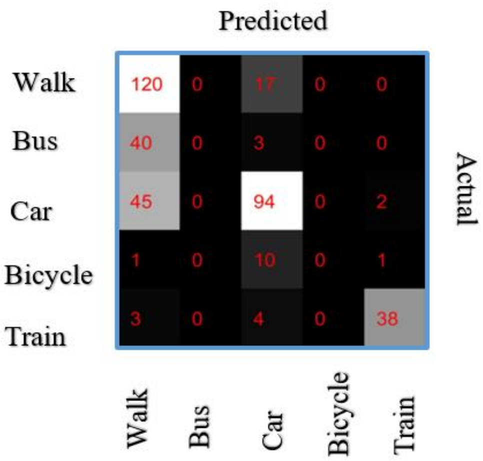
As mentioned in section 3, noises in the dataset are detected with the help of velocity and acceleration thresholding. The chosen thresholds are expressed in Table 2. Also, the threshold for HCR and SR parameters are set to be 5 degree and 1 meter/second respectively.

Parameter	**Velocity (m/s)**	**Acceleration (m/s^2)**
Walk	7	3
Bicycle	12	3
Bus	34	2
Train	50	10
Car	75	3

Table 2. Thresholds controlling the noise removal process

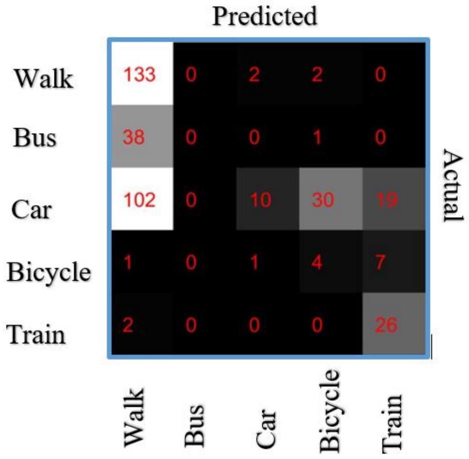
Also, for any trip, if the temporal distance between two consecutive points was more than the median value among all points of the trip, some other points were created and added to the trip to fill the gap. Finally, the coordinates of these points are calculated through spline interpolation.

In this paper, the train-test cross-validation split methodology was used. So, in the train-test split step, the data was separated into two groups of test and train that contain 70% and 30% of the data respectively. Then, the train group was used to produce ten different folds of training and validation split in a validation split operation. In each fold, 70% is considered as the training data and 30% as the validation data. Figures 8, 9, 10 show the confusion matrices for LR, linear SVC and GB respectively.



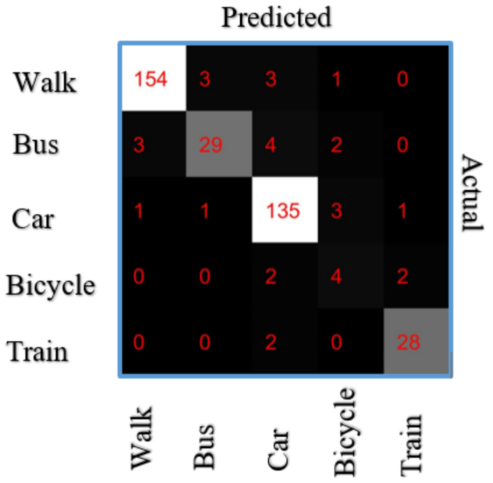
	Walk	Bus	Car	Bicycle	Train
Walk	120	0	17	0	0
Bus	40	0	3	0	0
Car	45	0	94	0	2
Bicycle	1	0	10	0	1
Train	3	0	4	0	38

Figure 8. Confusion matrix for LR classifier



	Walk	Bus	Car	Bicycle	Train
Walk	133	0	2	2	0
Bus	38	0	0	1	0
Car	102	0	10	30	19
Bicycle	1	0	1	4	7
Train	2	0	0	0	26

Figure 9. Confusion matrix for linear SVC classifier



	Walk	Bus	Car	Bicycle	Train
Walk	154	3	3	1	0
Bus	3	29	4	2	0
Car	1	1	135	3	1
Bicycle	0	0	2	4	2
Train	0	0	2	0	28

Figure 10. Confusion matrix for GB classifier

In addition to the confusion matrices, precision, recall, and F1-score are the metrics used to evaluate the performance of models, as shown in Table 3.

Model	Precision	Recall	F1-score
LR	0.60	0.67	0.62
Linear SVC	0.54	0.46	0.34
GB	0.93	0.93	0.93

Table 3. Results for each machine learning algorithm described by precision, recall and F1-score

As shown in Table 2, for F1-score measurement, the GB algorithm by 93% achieved the highest score, while the LR and the linear SVC reached 62% and 34% respectively. In terms of precision, the sequence of the algorithms performance was also similar and respective values of 93%, 60% and 54% are recorded for the GB, LR and linear SVC. For recall, the GB algorithm shows its effectiveness with a score of 93%, while the LR reached 67%, and linear SVC achieved 46%. Based on these results, in all calculated performance metrics, the GB algorithm outperforms the other machine learning algorithms.

6. CONCLUSION

This paper proposed a method to classify the vehicles used for trips based on the data acquired by smartphones. The core of the classification scheme was three different machine learning models, including GB, LR, and linear SVC. In the feature extraction process, noises in the datasets are removed based on thresholding, and possible gaps are filled using spline interpolation. Also, a filter has been investigated to make the trajectories smoother. The results showed that linear classifiers are inappropriate for the classification of travel types as both linear classifiers could not reach high accuracy. The most probable cause of this underperformance is that these models are over-fitted to the Walk and Car classes. This is, in turn, because of both the fewer samples of the Bicycle and Train classes, and the complexity of the problem in general. On the other hand, the Gradient Boosting algorithm was able to better distinguish between different classes and achieved average accuracy of 93%, which is the highest among the evaluated algorithms.

REFERENCES

Ajimakin, A. D., & Devi, V. S. (2021). Estimation of von mises-fisher distribution algorithm, with application to support vector classification. Proceedings of the Genetic and Evolutionary Computation Conference Companion,

Busch, P. A., & McCarthy, S. (2021). Antecedents and consequences of problematic smartphone use: A systematic literature review of an emerging research area. *Computers in human behavior*, *114*, 106414.

Erdelić, M., Carić, T., Erdelić, T., & Tišljarić, L. (2022). Transition State Matrices Approach for Trajectory Segmentation Based on Transport Mode Change Criteria. *Sustainability*, *14*(5), 2756.

Erdelić, M., Carić, T., Ivanjko, E., & Jelušić, N. (2019). Classification of travel modes using streaming GNSS data. *Transportation Research Procedia*, *40*, 209-216.

Erener, A., & Sarp, G. (2018). Spatiotemporal distribution of Industrial Regions and Impact on LST in the case of Kocaeli, Turkey.

Grennfelt, P., Engleryd, A., Forsius, M., Hov, Ø., Rodhe, H., & Cowling, E. (2020). Acid rain and air pollution: 50 years of progress in environmental science and policy. *Ambio*, *49*(4), 849-864.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning , chapter 10. Boosting and Additive Trees. In: New York: Springer.

Hosmer Jr, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). *Applied logistic regression* (Vol. 398). John Wiley & Sons.

Huang, F. L. (2022). Alternatives to logistic regression models in experimental studies. *The Journal of Experimental Education*, *90*(1), 213-228.

Huang, J., Hu, P., Wu, K., & Zeng, M. (2018). Optimal time-jerk trajectory planning for industrial robots. *Mechanism and Machine Theory*, *121*, 530-544.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

Ma, T. M., Yamamori, K., & Thida, A. (2020). A comparative approach to Naïve Bayes classifier and support vector machine for email spam classification. 2020 IEEE 9th Global Conference on Consumer Electronics (GCCE),

Madeh Piryonesi, S., & El-Diraby, T. E. (2021). Using machine learning to examine impact of type of performance indicator on flexible pavement deterioration modeling. *Journal of Infrastructure Systems*, *27*(2), 04021005.

Mumford, M. D., Schultz, R. A., & Osburn, H. K. (2002). Planning in organizations: Performance as a multi-level phenomenon.

Nawaz, A., Zhiqiu, H., Senzhang, W., Hussain, Y., Khan, I., & Khan, Z. (2020). Convolutional LSTM based transportation mode learning from raw GPS trajectories. *IET Intelligent Transport Systems*, *14*(6), 570-577.

Nguyen, H., Kieu, L. M., Wen, T., & Cai, C. (2018). Deep learning methods in transportation domain: a review. *IET Intelligent Transport Systems*, *12*(9), 998-1004.

Piryonesi, S. M., & El-Diraby, T. E. (2020). Data analytics in asset management: Cost-effective prediction of the pavement condition index. *Journal of Infrastructure Systems*, *26*(1), 04019036.

Sperandei, S. (2014). Understanding logistic regression analysis. *Biochemia medica*, *24*(1), 12-18.

Suh, Y. S., Shin, S. K., Baang, D., Seo, S. M., & Lee, J. B. (2021). A Brief Review of Linear Support Vector Machine for Machine Learning Programming. Transactions of the Korean Nuclear Society Virtual Spring Meeting,

Wang, X., Lu, S., & Zhang, S. (2020). Rotating angle estimation for hybrid stepper motors with application to bearing fault diagnosis. *IEEE Transactions on Instrumentation and Measurement*, *69*(8), 5556-5568.

Zheng, Y., Fu, H., Xie, X., Ma, W.-Y., & Li, Q. (2011). Geolife GPS trajectory dataset-user guide. *Geolife GPS trajectories*, *1*, 2011.

Zheng, Y., Wang, L., Zhang, R., Xie, X., & Ma, W.-Y. (2008). GeoLife: Managing and understanding your past life over maps. The Ninth International Conference on Mobile Data Management (mdm 2008),