Journal of the Society of Automotive Engineers Malaysia
Volume 8, Issue 3, pp 153-160, September 2024
e-ISSN 2550-2239

Optimization of Driver Behavior Profiling through K-means Unsupervised Clustering Algorithm Using Real-world Data

M. H. Danial*, Z. Z. Abidin and N. A. Asyqin

Centre for Unmanned Technologies (CUTe), Kulliyah of Eng., International Islamic Uni. Malaysia, Selangor

*Corresponding author: mhariz.hasbullah@live.iium.edu.my

ORIGINAL ARTICLE *Open Access*

Article History:

Received
1 Jul 2024

Accepted
17 Aug 2024

Available online
1 Sep 2024

ABSTRACT – *The use of telematics systems has brought benefits as conventional modern cars are now equipped with the technology, allowing data gathering to a centralized system, allowing vehicle monitoring and management. The data gathered from the sensors can provide insight into driver behavior and driving patterns. Unfortunately, the data is not fully utilized for in-depth analysis, as the pattern can be too sophisticated to understand. This hinders its potential to improve the safety of the road environment, as providing information from the data pattern can be utilized by drivers, law enforcement, policy makers, or anyone related to road safety management. This paper provides an in-depth analysis of real-world driving data behavior using unsupervised algorithms (K-means). The paper aims to assess the data pattern of the driver. The study uses the K-means algorithm to cluster the data of drivers and separate it by pattern. Further analysis is needed to classify driving based on characteristics of the clusters, such as bad and good drivers. The findings reveal that K-means was able to identify patterns of the driving behavior, and analysis was done for categorization. 6 clusters were identified in the algorithm, where clusters 2 and 4 exhibit good driving patterns while clusters 1,3,4, and 6 exhibit bad driving patterns. This research provides crucial information to the driver's awareness, gives insight into policymakers and law enforcement, thus improving the safety of the road.*

KEYWORDS: K-means, telematics systems, driving behavior, road safety

1. INTRODUCTION

Telematics has emerged as a pivotal technology in the automotive industry, transforming how vehicles operate, communicate, and integrate into broader transportation networks. A Telematic Control Unit (TCU) in a vehicle allows data from a variety of sources within a connected vehicle to be collected (Gekker & Hind, 2020). This technology allows important data to be collected and analyzed, understanding information behind the data collection, potentially improving the safety of the road. The majority of driving accidents are due to human errors; these errors are from reckless and undisciplined driving behaviors and have always been the leading contributors to all sorts of incidents across the globe (Ghaffarpasand et al., 2022). More than 90% of traffic accidents are due to drivers' behavior and their interests. Thus, studying how drivers' behavior is important to reduce traffic accidents (Liu et al., 2020).

A telematic system with an intelligent vehicular telematic platform was proposed, which allows real-time monitoring of vehicular information such as vehicle engine speed, oxygen levels, speed per hour, and water temperature. The concept works by connecting the Controller Area Network (CAN) bus in the vehicle with an On-Board Diagnostic (OBD) bridge to receive information. The CAN bus can be used as internal LANs communication in a vehicle; hence, outside communication is unable to connect unless OBD is used to directly connect to the Local Area Networks (LANs) communication. OBD provides significant information because of the diagnostic mode to analyze malfunctions in a vehicle (Chen et al., 2016). The majority of modern cars come with OBD technology, where information about the car can be obtained, such as speed, engine speed, acceleration, and deceleration, to be analyzed (Pereira et al., 2016).

A study by Liu et al. (2020) obtains drivers' data from the OBD terminal and combines it with x-axis and y-axis acceleration changes and behavior duration of the vehicle's three-axis acceleration sensor to identify abnormal driving behavior, establishes a hierarchical driving behavior indicator system, and a judgment matrix. It uses threshold standards as references to detect abnormal driving behavior such as rapid acceleration, rapid deceleration, and sharp turns. By analyzing the data from the drivers, bad driving behavior can be detected and send awareness to the driver. From past studies, different approaches were made to determine driving behavior, such as conducting driving data simulation or collecting data through GPS to obtain speed and acceleration to evaluate driving behavior (Wu, 2004). However, existing methods for detecting aggressive driving often rely on subjective definitions and manually set thresholds based on specific driving parameters (Júnior et al., 2017). For example, threshold used for driving behavior detection is for safe acceleration and deceleration lie approximately within a range of ±0.3 g (3m/s2), sudden acceleration and deceleration lie within a range of ±0.5g (5m/s2), and gradual lane changes produce an average g-force less than ±0.1g. (approximately 1m/s2), and unsafe lane changes have a g-force over ±0.5 g (Fazeen et al., 2012). A comparison between rule-based and pattern-matching algorithms was made. The results reveal that pattern-matching algorithms give better performance than rule-based algorithms (Saiprasert, 2013).

Driving behavior data can be analyzed using algorithms, statistical analysis, deep learning, or machine learning. By analyzing driving behavior, it can be applied to various situations such as road conditions improvement, safety warning systems, and many more (Anil & Anudev, 2022). Furthermore, driving analytics adaptations improve other aspects of transportation systems, such as increasing overall security and reducing the usage of vehicle energy and gas emissions. Therefore, by exploring the potential of driver analysis, it could further improve safer and energy-efficient driving style (Syed Ahmad et al., 2022). Unsupervised clustering algorithms can be used to label unlabeled raw data. In this research, K-means is used as our algorithm to detect driving behavior categories. K-Means clustering is an unsupervised learning algorithm used to group an unlabeled dataset into clusters.

Identification and categorization of driving behavior are necessary because they improve traffic safety, and this application can be used in intelligent transportation systems (Xiao et al., 2023). K-means algorithms have shown satisfactory results in recognizing driving patterns and dividing them into clusters without the need for labelling raw data (Chhabra et al., 2017). Limitations of the algorithm also need to be considered, as high-dimensional data with low variance may pose a challenge for clustering algorithms such as K-means, leading to lower accuracy, less meaningful clustering, and high computational complexity.

Aggressive driving behavior studies based on Principal Component Analysis (PCA) and K-means clustering show the effectiveness in recognizing different driving patterns without the use of thresholds. The ability to separate clusters in the K-means shows the relevance of the clustering algorithm to recognize driving behavior patterns. However, the driving pattern is affected by external factors such as traffic environment, vehicle types, and weather conditions; therefore, it requires further analysis on its dataset (Xiao et al., 2023). Therefore, the enhancement in using the K-means models across different types of driving pattern datasets needs to be studied.

In this study, an exploration was conducted to study the driving patterns of the drivers. The aim is to use the K-means algorithm for cluster separation, result analysis, and driving profiling categorization with the use of a dataset that has higher dimensions, such as speed and engine load data.

2. METHODOLOGY

2.1 Scope of Study

This study will be conducted on a normal one-lane road, and the road conditions are safe for driving. Two parameters will be maintained throughout the data collection, which is a fixed route during the collection; this ensures enough data points are collected. The second parameter is to use the same type 2 passenger vehicle during the collection. These parameters are crucial to maintain data reliability and reduce variability. After a certain amounts of data are collected, it will be inserted into the K-means algorithm to allow clustering and analysis of the data will be done for driving behavior categorization.

2.2 Experimental Setup

The project aims to collect driving data from drivers, where each driver displays their own driving behavior. A hardware telematic system was developed and implemented in a vehicle for data analysis. The proposed systems will help implement the features and technologies required to achieve our objectives by improving the safety of traffic and vehicles. Various data will be collected by the sensors and will be sent to the microcontroller of the telematic system to be processed. Then the data will be transferred to the K-means algorithm for driving behavior analysis. A fixed route is determined shown in Figure 1 where the distance of the route is 3.3 km. Figure 1 shows the route that will be used during the data collection.

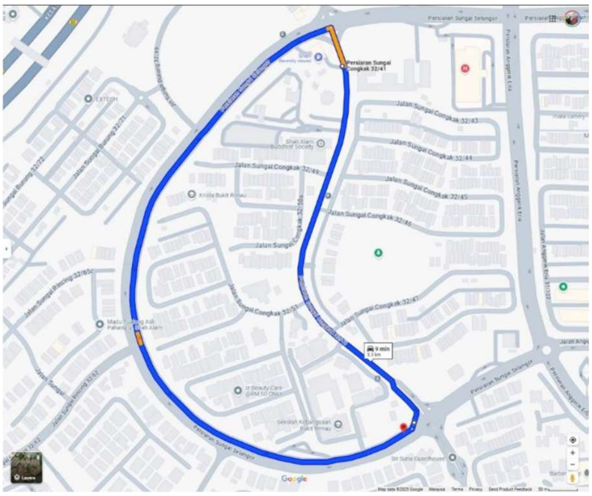
Persiaran Sungai Selangor
Persiaran Sungai Congkak 32/41
Jalan Sungai Congkak 32/43
Jalan Sungai Congkak 32/44
Jalan Sungai Congkak 32/45
Jalan Sungai Congkak 32/46
Jalan Sungai Congkak 32/47
Jalan Sungai Congkak 32/49
Jalan Sungai Congkak 32/50s
Jalan Sungai Congkak 32/51
Jalan Sungai Buring 32/71
Jalan Sungai Buring 32/72
Jalan Sungai Rincing 32/65c
Jalan Sungai Rincing 32/62
Persiaran Anggerik Eria
Shah Alam Buddhist Society
Sekolah Kebangsaan Bukit Rimau
9 min
3.3 km
Google

FIGURE 1: Single-lane road route

2.3 Data Sources

The data for this study was collected using the OBD-II UART board by SparkFun and the MPU-6050 sensor, both of which are essential tools for capturing the necessary data from the vehicle for driving behavior data. A microcontroller, ESP8266, is used to process the data from the sensors and save it to a local machine.

The OBD-II UART Board by SparkFun is a device designed to interface with a vehicle's On-Board Diagnostics (OBD-II) system. It retrieves real-time vehicle parameters such as speed and engine load. Speed, measured in kilometers per hour (km/h), provides critical information on the velocity of the vehicle, while engine load indicates the percentage of engine capacity being utilized. These parameters are vital for profiling driving behavior, as they directly reflect how the vehicle is being operated in different conditions. The board sends requests for data by sending the Parameter identification (PID) to the vehicle computer to request the information. However, PID standards may differ depending on the car manufacturer, therefore, further research is needed to identify what PID of the vehicle. Figure 2 shows the SparkFun OBD-II UART.

The MPU-6050 Sensor captures both acceleration and gyroscopic data along the X, Y, and Z axes. The acceleration data format is in "g" which is the gravitational unit where 1 g refers to 9.8 m/s^2; thus, acceleration data provides insights into linear motion. Meanwhile, gyroscopic data captures rotational movements of the vehicle by giving the data in degrees. Derived metrics such as acceleration magnitude, jerk, and angular velocity are calculated to better represent driving behavior. These metrics are crucial for detecting sudden changes, such as harsh braking, rapid acceleration, sharp turns, or erratic movements, which are often indicative of aggressive or unsafe driving styles. The sensor's high sensitivity and versatility make it a critical component of the data collection system. Figure 3 shows the MPU-6050 accelerometer sensor.

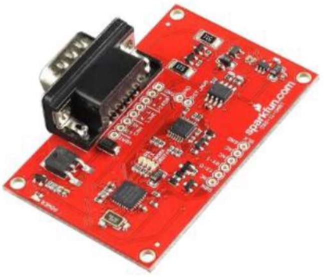

FIGURE 2: SparkFun OBD-II UART

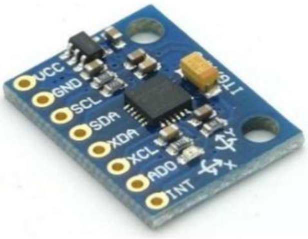

FIGURE 3: MPU-6050 accelerometer sensor

2.4 Hardware Telematic System

The telematic box used for this research is an integrated device that is housed in a compact electrical box containing components necessary to collect data. The components consist of the main microcontroller, which is an ESP8266 microcontroller, to process data from the OBD UART board and MPU-6050 sensors. The microcontroller translates serial communication from these external modules to readable data and sends data requests to the vehicle’s ECU, hence supporting real-time data streaming during the vehicle's operation. The OBD-II UART 2 board by SparkFun is used to retrieve vehicle parameters such as speed and engine load through the OBD-II port of the vehicle. This board is necessary as a gateway between the ESP8266 microcontroller and the vehicle’s Electronic Control Unit (ECU), as certain requirements are needed before requesting the data. Figure 4 shows the telematic box hardware setup.

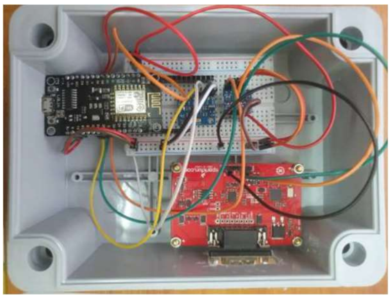

FIGURE 4: Telematic box hardware

2.5 Integration with Vehicle

The telematic box is connected to the vehicle via the OBD-II port and positioned securely within the vehicle cabin. The setup includes a few processes for this research, where, firstly, vehicle compatibility is needed for the telematic box to retrieve data. The telematic box is compatible with any vehicle equipped with an OBD-II port, a standard feature in most modern cars. For this study, the test vehicle used was a 2022 Proton Saga MC2; this car model has an OBD-II port and provided reliable OBD-II data access. The figure below shows the car model used for the setup and the OBD port location.

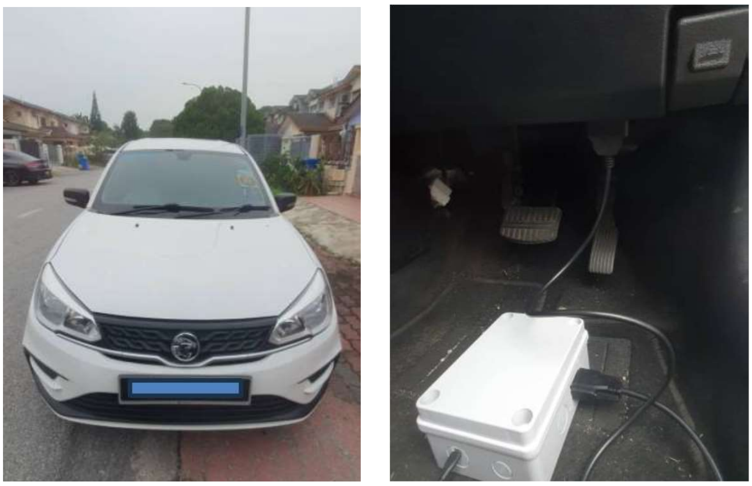

FIGURE 5: Vehicle (left); OBD port location (right)

3. RESULTS AND DISCUSSION

Figure 6 shows the Elbow method used to determine the result in determining optimal number of clusters (K), It plots the sum of squared distances (inertia) between data points and their assigned cluster center for different values of k. The figure shown shows the number of clusters used in clustering on the x-axis, and the y-axis represents the compactness of clusters (inertia). The lower the inertia difference value, the higher the chance of the elbow point occurring. As seen from the graph, the clusters K was tested to 10 clusters. The slope of the data plot decreased gradually at different rates as clusters increased, eventually forming an elbow-like shape. From Clusters 1 to 2, the inertia difference is high, indicating more compact clusters. As the clusters increase, the decrease in inertia becomes less, showing that adding more clusters gives diminishing returns. Six clusters were chosen for clustering because beyond cluster 6 does not significantly reduce the inertia and complexity of the model will also increase.

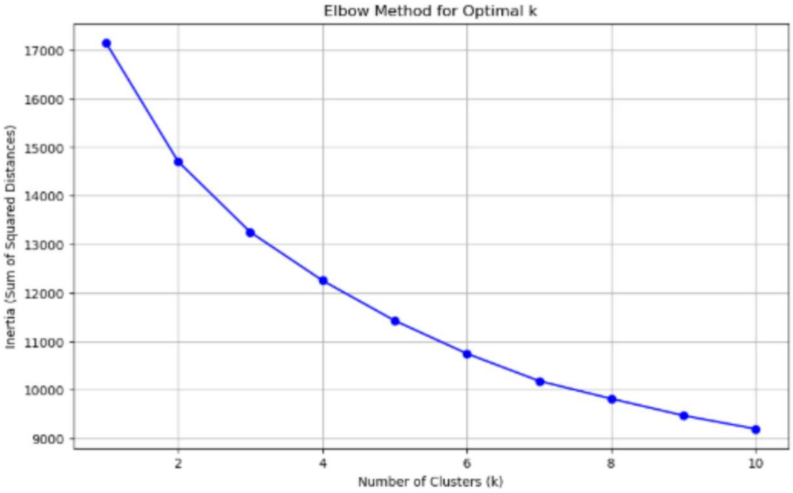

FIGURE 6: Elbow method analysis

Figure 7 shows the K-means clustering visualization with K = 6. The figure shows the clustering result with six clusters, whereby each cluster is color-coded to distinguish the separation. Each cluster represents different driving patterns where K-means effectively separation between the different driving patterns for analysis. The clustering result is represented by Principal Component Analysis (PCA) pattern to reduce the high-dimensional feature space into two principal components for readable analysis. Cluster 0 is represented by purple, Cluster 1 is dark blue, Cluster 2 is light blue, Cluster 3 is light green, Cluster 4 is green, and Cluster 5 is yellow. The PCA-based clustering visualization effectively illustrates six different driving patterns. Figure 8 illustrates the visualization in the 3D component.

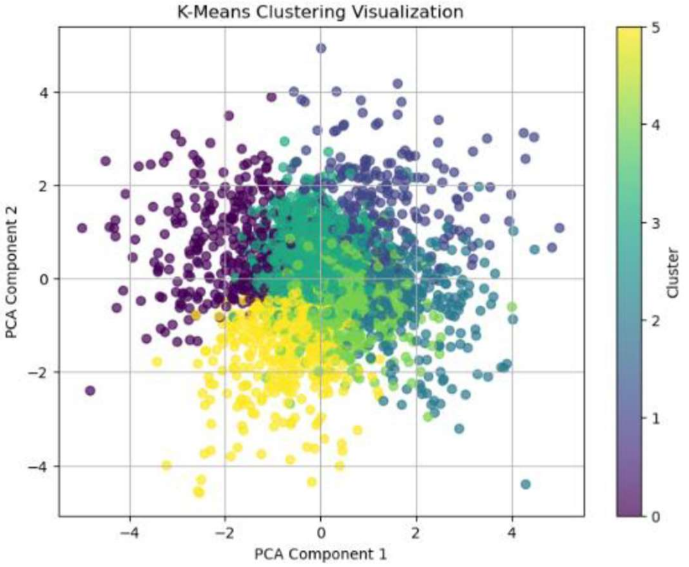

FIGURE 7: 2D PCA visualization clusters

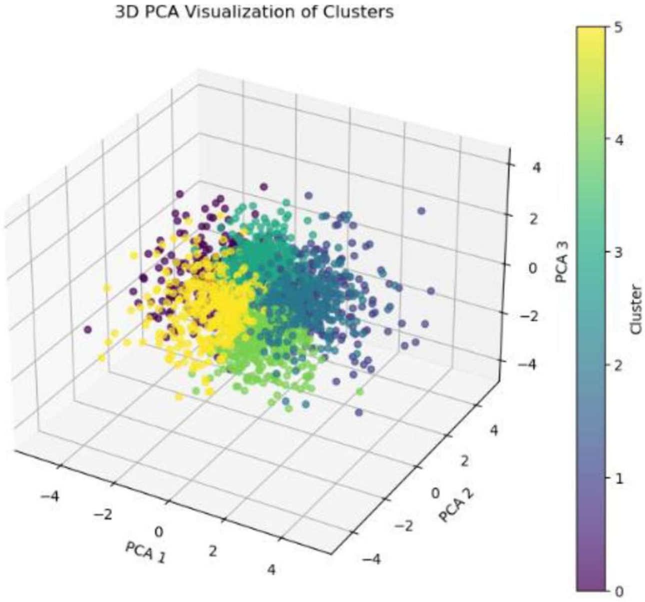

FIGURE 8: 3D PCA visualization clusters

Table 1 shows the summary of the clustering pattern and its explanation. The acceleration data reveals distinct driving behavior patterns categorized into good and bad driving behavior, along with a description of driving patterns. The range of AccX (longitudinal acceleration) and AccY (lateral acceleration) values for each cluster further supports these classifications, as clusters associated with bad driving exhibit wider acceleration variations, whereas good drivers maintain a more stable and controlled range. Cluster 1 shows the AccX values range from -1.385 to 2.980 m/s2, showing the highest magnitude of acceleration but smooth braking (deceleration), for AccY lane changes range from -2.735 to 2.458 m/s2, showing stable lane changes. In cluster 2, it shows AccX value of -2.378 to 2.383 m/s2, indicating a moderate magnitude value of acceleration and braking, as the value is lesser than Cluster 1 for acceleration and Cluster 3 for braking, for AccY lane changes range from -2.337 to 2.403 m/s2, showing stable lane changes. In Cluster 3, the AccX value ranges from -2.772 to 2.383 m/s2, showing the highest magnitude of braking but smooth acceleration, for the AccY range from -2.475 to 2.393

m/s2, indicating stable lane changes. In Cluster 4, the AccX value ranges from -1.692 to 1.918, showing smooth acceleration and braking, for the AccY value range of -0.499 to 2.521 m/s2indicating stable lane changes. For cluster 5, AccX shows a value range -2.592 to 2.667 m/s2showing moderate acceleration and deceleration, for AccY value range -2.824 to 0.324 m/s2, indicating sharp lane changes. Finally, for cluster 6, AccX ranges from -2.627 to 2.018 m/s2, showing moderate acceleration and braking, and for AccY, the range is -2.682 to 2.271 m/s2, showing sharp lane changes.

TABLE 1: Clustering pattern

K=6	AccX (m/s2)	AccY (m/s2)	Description	Category
Cluster 1	(-1.385) to 2.980	(-2.735) to 2.458	**Harsh** acceleration and smooth braking, **stable lane changes**	Bad driver
Cluster 2	(-2.378) to 2.383	(-2.337) to 2.403	**Moderate** acceleration and braking**, stable lane changes**	Good driver
Cluster 3	(-2.772) to 1.857	(-2.475) to 2.393	**Harsh** braking and smooth acceleration, **stable lane changes**	Bad driver
Cluster 4	(-1.692) to 1.918	(-0.499) to 2.521	**Smooth** acceleration and braking, **stable lane changes**	Good driver
Cluster 5	(-2.592) to 2.667	(-2.824) to 0.324	**Moderate** acceleration and braking, **sharp lane changes**	Bad driver
Cluster 6	(-2.627) to 2.018	(-2.682) to 2.271	**Harsh** braking and smooth acceleration, **sharp lane changes**	Bad driver

By comparing the range value in Table 1. The highest and the lowest magnitudes of acceleration data, AccX and AccY, indicate a threshold for driving pattern evaluation. Cluster 1, 3, 5, 6 is categorized as a driver because they exhibit the driving pattern of harsh acceleration, harsh braking, and sharp lane changes. Meanwhile, clusters 2 and 5 are classified as good drivers, showing more stable changes and below the threshold value magnitude.

4. CONCLUSION

This study successfully applied K-means clustering to analyze and classify driving behaviors based on driving data of AccX (longitudinal acceleration) and AccY (Lateral acceleration). The data was captured on a Type 1 vehicle and a fixed route. This is to ensure data has stable variability. The clustering results revealed six distinct driver profiles, where Clusters 2 and 4 were identified as good drivers with smooth acceleration, braking, and stable lane changes, while Clusters 1, 3, 5, and 6 exhibited aggressive tendencies such as harsh acceleration, harsh braking, and sharp lane changes, categorizing them as bad drivers. The acceleration variations in bad driver clusters showed significantly wider ranges compared to good driver clusters, supporting the effectiveness of clustering in differentiating between safe and risky driving behaviors. These findings highlight the effectiveness of K-means clustering in differentiating between aggressive and cautious driving behaviors, providing valuable insights for applications in road safety analysis, driver monitoring systems, and insurance risk assessments. These findings emphasize the potential of unsupervised machine learning in driver profiling, which can be leveraged for applications such as driver monitoring systems, insurance risk assessments, and intelligent transportation safety programs. However, this study is limited to acceleration-based clustering and does not consider external environmental factors such as weather conditions or traffic influences. Moreover, the data quality can be further improved with higher variations and more data points, which allows the clustering process to be more separable, hence more meaningful. Future research can explore real-time clustering applications, multi-sensor data integration, external factor dataset and adaptive learning models to further enhance driver behavior classification accuracy and robustness.

ACKNOWLEDGMENT

The authors would like to thank the reviewers for their detailed comments, which have improved the quality of this paper. The main author is supported by the project of International Islamic University Malaysia (IIUM) with Collaborative Research in Engineering, Science & Technology Center (CREST) and Delloyd R&D (M) Sdn. Bhd. This work is under grant funding (grant ID CREST24-001-0002).

REFERENCES

Anil, A. R., & Anudev, J. (2022). Driver behavior analysis using K-means algorithm. International Conference on Intelligent Computing, Instrumentation and Control Technologies, ICICICT, 1555–1559.

Chhabra, R., Verma, S., & Krishna, C. R. (2017). A survey on driver behavior detection techniques for intelligent transportation systems. In 2017 7th International Conference on Cloud Computing, Data Science & Engineering-Confluence (pp. 36-41). IEEE.

Xiao, Y., Liu, Y., & Liang, Z. J. (2023). Road rage recognition model for truck drivers based on K-means clustering and random forest algorithm. *Advances in Transportation Studies*, *60*.

Chen, L. B., Li, H. Y., Chang, W. J., Tang, J. J., & Li, K. S. M. (2016). An intelligent vehicular telematics platform for vehicle driving safety supporting system. 2015 International Conference on Connected Vehicles and Expo, ICCVE 2015 - Proceedings, 210–211.

Fazeen, M., Gozick, B., Dantu, R., Bhukhiya, M., & González, M. C. (2012). Safe driving using mobile phones. IEEE Transactions on Intelligent Transportation Systems, 13(3), 1462–1468.

Gekker, A., & Hind, S. (2020). Infrastructural surveillance. New Media and Society, 22(8), 1414–1436.

Ghaffarpasand, O., Burke, M., Osei, L. K., Ursell, H., Chapman, S., & Pope, F. D. (2022). Vehicle telematics for safer, cleaner and more sustainable urban transport: A review. Sustainability, 14 (24), 16386.

Júnior, J. F., Carvalho, E., Ferreira, B. V., De Souza, C., Suhara, Y., Pentland, A., & Pessin, G. (2017). Driver behavior profiling: An investigation with different smartphone sensors and machine learning. PLoS ONE, 12(4).

Liu, T., Yang, G., & Shi, D. (2020). Construction of driving behavior scoring model based on OBD terminal data analysis. Proceedings - 2020 5th International Conference on Information Science, Computer Technology and Transportation, ISCTT 2020, 24-27.

Pereira, A., Alves, M., & Macedo, H. (2016, April). Vehicle driving analysis in regards to fuel consumption using Fuzzy Logic and OBD-II devices. In 2016 8th Euro American conference on telematics and information systems (EATIS) (pp. 1-4). IEEE.

Saiprasert, C. (2013). Detecting driving events using smartphone. Retrieved from https://www.researchgate.net/publication/279949922

Syed Ahmad, S. S., Muhammad, M., & Jawi, Z. M. (2022). Driving analytics - Data science approach based on smartphone vehicle telematic data. Journal of the Society of Automotive Engineers Malaysia, 6(2), 77–84.

Wu, J. (2004). Analysis of taxi drivers' driving behavior based on a driving simulator experiment simulator experiment. University of Central Florida.