SEMARAK ILMU PUBLISHING

Journal of Advanced Research in Applied Sciences and Engineering Technology

Journal homepage:
https://semarakilmu.com.my/journals/index.php/applied_sciences_eng_tech/index
ISSN: 2462-1943

Driver Behaviour Classification: A Research using OBD-II Data and Machine Learning

Nur Farisya Aqilah Muhamad Fadzil[1], Hilda Mohd Fadzir[1], Hafizah Mansor[1,*], Untung Rahardja[2]

[1] Department of Computer Science, Kulliyyah of ICT, International Islamic University Malaysia, 53100 Gombak, Selangor, Malaysia
[2] Faculty of Science and Technology, University of Raharja, Kota Tangerang, Banten 15117, Indonesia

ABSTRACT

Classification of driver behaviour has gained much attention due to its potential in a variety of applications, and On-Board Diagnostic (OBD) real-time data is often under-utilised. Hence, using On-board Diagnostic-II (OBD-II) data by categorising drivers based on their driving behaviour can be an efficient method. The objective of this study is to identify groups of drivers based on their driving styles using the collected OBD-II data. This study uses a Kaggle-obtained online dataset of OBD-II. The suggested model in this study analyses driving behaviour using both supervised and unsupervised methods. The relationship between all features and engine speed is analysed to select the optimal features, which include engine speed, vehicle speed, throttle position, and calculated engine load. Then, the proposed model makes use of the K-Means algorithm to create driving behaviour labels whether belong to safe or aggressive - validated by the safety score criteria. Different machine learning models including Random Forest (RF), Decision Tree (DT), Support Vector Machine (SVM), Multi-Layer Perceptron (MLP), AdaBoost (AB), Linear Combination (LC) and Weighted Linear Combination (WLC) are used, customised, and compared to get the most accurate prediction of driver behaviour. Experimental results indicate that the suggested driving behaviour analysis can reach an average rate of 98.72% accuracy using DT. However, implementing the ensemble method AB has improved the accuracy to 99.48%.

Keywords:

Driving behaviour analysis; Unsupervised and supervised; On-Board Diagnostic-II (OBD-II)

1. Introduction

In the automotive industry, vehicles are becoming increasingly software-intensive, complex systems in which software and electronics are the primary sources of innovation. More than 100 Electronic Control Units (ECUs) are installed in modern automobiles, and these ECUs are mostly compact computers that continuously execute gigabytes of software [1]. Not only that, but driver behaviour analysis is also a new trend that meets the demands of a variety of markets. Therefore, grouping the drivers based on their driving styles can efficiently utilise the data. Methods of data analysis are essential for analysing the ever-increasing volume of high-dimensional data, and driving behaviour analysis is one of the necessary studies. Driver behaviour affects traffic safety, fuel or

* *Corresponding author.*
E-mail address: hafizahmansor@iium.edu.my

https://doi.org/10.37934/araset.56.2.5161

energy consumption and gas emissions. According to current research on road traffic accidents in Malaysia, the total fatality caused by road accidents is growing year by year. Accident incidents in Malaysia are often due primarily to the driver's factor, which includes aggressive, careless, and reckless driver behaviour. In other terms, it is more attributable to human error [2].

One of the approaches to solving these concerns is to study how car drivers behave when driving by utilising the On-Board Diagnostic II (OBD-II) dataset. The OBD is a built-in self-diagnostic system that can access, monitor, communicate, and report vehicle operational statuses such as engine speed, engine load, throttle position and many more parameters through multiple vehicle subsystems.

The proposed model measures driving behaviour by utilising four main attributes from the original dataset: engine speed (also known as Revolutions per Minute (RPM)), vehicle speed, engine throttle position and calculated engine load. These values are chosen after determining the correlation between RPM and all other parameters. This proposed model uses both supervised and unsupervised to analyse driving behaviour as the data has no class label but only attributes describing the features of each type of sensor. In this work, K-Means algorithms are applied for the clustering method in the first experiment. Then, the second step is to apply supervised techniques to classify the driver's behaviour. Besides that, the ensemble algorithms are also used to see the improvement in the accuracy of machine learning. This study compares selected algorithms (Decision Tree, Support Vector Machine and Multi-Layer Perceptron) and hybrid machine learning algorithms (AdaBoost, Random Forest, Linear Combination and Weighted Linear Combination) to discover the algorithm with the highest performance for the driving behaviour grouping model. Thus, this paper aims to evaluate the performance of multiple combinations of machine learning algorithms and explore new ways to improve the accuracy.

2. Literature Review

Researchers have recently focused a significant interest on studies regarding driver behaviour and the prediction of automotive systems' problems [3,4]. This growing trend is foreseen, given the rapid growth of computer technology and machine learning approaches [5]. We have compiled a series of driver behaviour studies to improve automobile-related problems including the underutilisation of real-time OBD-II data and the inability to directly measure driving style information. These studies are also analysed to optimise the method of identifying driver behaviour.

The study by da Silveira Barreto *et al.*, [6] is the closest to our work since both utilise the same methods to categorise driver behaviour, which is a combination of unsupervised and supervised algorithms. The study groups the driving pattern into 'high', 'mid', and 'low' categories. They divided the procedure into unsupervised and supervised steps. The first experiment utilised three clustering approaches, the second experiment used six machine learning algorithms. For profiling, K-Means produced the best results, with Silhouette Index (SI) (0.349) and Davies-Bouldin (DB) (0.995), while the maximum accuracy was obtained from partitions with three groups, using Multi-Layer Perceptron (MLP) (99.253%). Then, platforms are created as a distributed tool to collect data from car sensors and categorise the driver's profile. The author provides a distributed system for car engine sensing that clusters and classifies driver usage. This platform will help with fleet management, insurance, fuel efficiency, and carbon dioxide emissions. The author plans to interface with other platforms to collect traffic data, speed restrictions, Global Positioning System (GPS), and other information to bring new services to the recommended platforms.

K-Means and Hierarchical Agglomerative Clustering (HAC) were implemented on a Principal Component Analysis (PCA) and T-distributed Stochastic Neighbourhood Embedding (t-SNE)-reduced

dataset [7]. The goal is to identify groups of drivers based on driving style and assess K-Means and HAC algorithms in reduced data (post t-SNE and post PCA) to choose the optimal clustering method. Driving style can impact several automotive systems and be utilised to understand quality concerns, such as maintenance, trouble codes and many more according to the author. The t-SNE dimensionality reduction approach with the K-Means clustering algorithm delivers the best result with the greatest SI (0.39), Calinski-Harabasz Index (CHI) (485.37), and lowest DB (0.571). Future work will involve building a classification model utilising the label and a machine-learning model to predict monitoring failures. K-Means clustering algorithm is also applied to conduct clustering which helps to improve the performance of a recommendation system in another domain [8].

Classification algorithms were used on the feature-engineered data [9]. The authors employed Support Vector Machine (SVM), Decision Tree (DT), Random Forest (RF), K-Nearest Neighbour (KNN), Extreme Learning Machine (ELM), and a customised feed-forward Deep Neural Network (DNN) to evaluate the driver's risk profile. The author compared chosen and customised machine learning algorithms for risk prediction. RF classifiers surpass all others with 97% accuracy and the highest Area Under Curve (AUC) score (0.89). The authors proposed a cloud-based profiling system that collects real-time vehicle network data, radar range, and smartphone inertia measurements to identify driving behaviours using sequence modelling as their future work.

The study by Júnior *et al.*, [10] applies supervised machine learning algorithms to construct a classification that characterises the driver aggressiveness profile. This work attempts to analyse the performance of the many combinations of machine learning algorithms and Android smartphone sensors while conducting different approaches. The performance of Artificial Neural Networks (ANN), SVM, RF and Bayesian Networks are examined and compared. The author determined that RF is by far the best-performing algorithm, followed by MLP. In future studies, the authors intend to gather a bigger number of driving events samples using different cars, Android smartphone modes, road conditions, weather, and temperature. They also intended to integrate additional machine-learning techniques into their evaluation.

Supervised machine learning methods were used to describe a platform that incorporates well-known machine and deep learning approaches [11]. This research seeks to classify drivers' behaviours as either eco-friendly or not. K-Means has been used as the clustering algorithm. The conclusion from this experiment demonstrates that k=2 has the optimal number of clusters. After the labelling operation, the new labelled dataset was delivered as the input to Machine Learning and Deep Learning for the classifying process. The classification experiment employed these algorithms, Logistic Regression, SVM, MLP, RF, and Recurrent Neural Network (RNN). The accuracy score of three classical algorithms is over 95% in LR at 98.2%, SVM and RF at 100% however the accuracy rate for MLP is 99.8% and RNN (Long-Short Term Memory (LSTM)) at 100%. For future works, more labels should be supplied during the execution of the real unsupervised analysis. The evaluation of each driver might be returned in real time using the proposed cloud-based platform.

The same dataset was used in a study published by Kumar *et al.*, [12]. In the model, both unsupervised and supervised approaches were used. The objective of the study is to develop a driving prediction model that would allow insurance firms to charge their customers for their driving behaviour rather than being taxed with a predetermined amount. Expectation Maximisation, Agglomerative Hierarchical and K-Means are employed using Weka as the dataset is unlabelled. The experiment then continues with measuring the quality of the cluster using Silhouette and Davies Bouldin indices. Seven distinct algorithms were used for the classification which include AB, DT, KNN, MLP, Naive Bayes (NB), RF, and SVM. The 10 cross-fold is then applied in which the dataset is split into 10-fold, 90% for training and 10% for testing. The results obtained after training the model reveal that the accuracy for all techniques is over 94.3% and based on mean and standard deviation, MLP

(99.2%) has the best accuracy score. MLP then is being utilised for cloud application development. Lastly, in future works, the authors suggested that the right approach is needed to restrain the ML method in acquiring new data from new drivers.

A unique approach for driver behaviour profiling by using time frames and data segmentation was proposed by Al-Hussein *et al.*, [13]. The rows and segments are labelled and classified to compare modulated models and choose the best for the recognition system. Deep Neural Networks (DNN), Recurrent Neural Networks (RNN), and Convolutional Neural Networks (CNN) are employed. According to the confusion matrix, DNN and RNN have identical performance and no overfitting. CNN has 96.1% accuracy and 95.2% f-measure. DNN's accuracy is 82.8% and the f-measure is 81.5%. Therefore, CNN has been chosen. The authors said future studies should compare LSTM's performance. A recognition mechanism should be created online.

Data-gathering studies were focused on driving behaviour by Ameen *et al.*, [2]. Based on real-time data acquired from cars and reference data supplied by earlier researchers, the authors created a method to identify driving patterns divided into four groups: risky, aggressive, safe, and typical conduct to minimise the likelihood of collisions. Then, comparison and statistical approaches were utilised to establish the ideal strategy for collecting driving data, utilising independent-sample t-tests and Statistical Package for the Social Sciences (SPSS) statistics to compare the means between groups on the same continuous dependent variable.

Experiments employing GPS and OBD data from a hybrid car during real-world driving cycles were outlined by Puchalski *et al.*, [14]. The work focuses on the search for vehicle speed and acceleration signals and metrics generated from them that best represent a safe driving style. The author implements K-Means clustering, the most popular unsupervised machine learning algorithm for partitioning data into k groups. Drivers were categorised based on a statistical analysis of input data and Modulation Transfer Function (MTF) by clustering. The author analysed driving cycles based on two criteria: (1) Driver preferences and driving characteristics (aggressive and not aggressive); (2) Ecological safety (eco-driving, eco-neutral, and not eco-driving).

The study by Fan *et al.*, [15] presents eco-driving assistants that analyse driving performance and give feedback to minimise fuel consumption. In addition to using the same dataset as this paper, this study employs the same relative scoring methodology for their model. This study also implements a white-box analysis using explicit driving categorization metrics and a black-box analysis using K-Means clustering to categorise driving styles based on driving data. Regression techniques were evaluated using Root Mean Square Error (RMSE) and Mean Absolute Percentage Error (MAPE). MLP, Linear Support Vector Regression (SVR), Extra Tree, AB, RF, DT, KNN, Stochastic Gradient Descent (SGD), Automatic Relevance Determination (ARD), SVR, and Gaussian Process are regression methods. Models with RMSE 2 or MAPE 0.05 are used to extract significant features and discard unnecessary ones to reduce noise [15].

3. Methodology

3.1 Data Collection

Figure 1 shows the experimental process flowchart in our work that starts with data collection. The original dataset utilised for experiments in the project is obtained from a Kaggle public database [16]. The owner of the dataset acquired fifteen different OBD-II connectors, and the values were sent to a smartphone through OBD-II for recording [6]. These values are then transmitted to a decentralised online platform capable of storing, fusing, and analysing records for all registered users. From the literature review, we have discovered a few reliable datasets that may be utilised for the study. However, this dataset was chosen since it is the most recent (2018) OBD-II dataset out of all.

Since the dataset is unlabelled, it meets our learning objective of employing both unsupervised and supervised approaches. Following the collection of data from all 19 drivers, there are 28 distinctive features and about 8261 instances, representing an average of 434 rows for each drive.

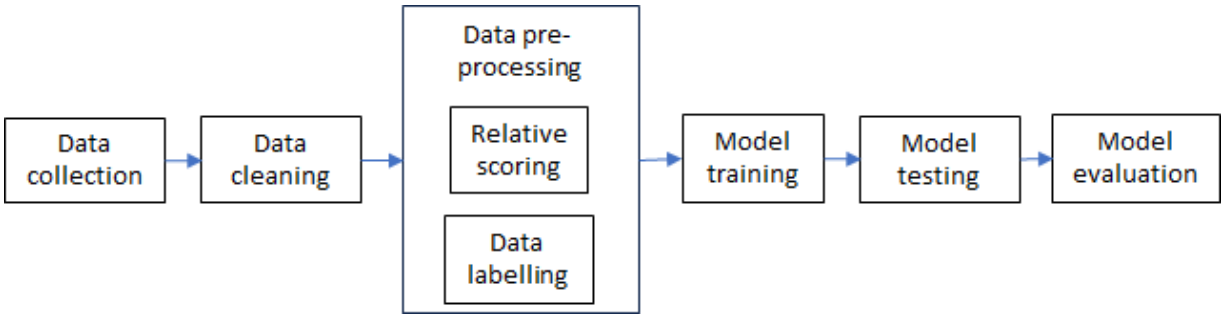

Fig. 1. Experimental process flowchart

3.2 Data Cleaning

After considering the correlations between all attributes and engine speed, we selected only a few attributes from the original dataset. This includes the engine speed (RPM), vehicle speed, throttle position and engine load. Engine speed (RPM) is one of the most essential engine output data variables. The optimal driving operation is not only characterised by a high and low engine speed but also by the driver's ability to move the vehicle through the low-efficiency area of the engine in the shortest time possible while maintaining a constant vehicle speed and engine speed. Next, engine load is representative of the operating environment of a driver. The vehicle's speed is a direct result of the driver's behaviour. It can depict whether drivers are speeding or maintaining safe driving conditions. The throttle position is used to quantify acceleration pedal movement. The engine requires a proper opening to operate efficiently. A throttle pollution setting that is too high or too low will result in incomplete fuel combustion and air pollution. While engine runtime specifies the duration of the engine's operation. This information will help improve the pre-processing of data.

3.3 Data Pre-Processing

The dataset was pre-processed by removing unnecessary columns. Next, additional data recorded before the commencement of experiments and following their completion were discarded. For instance, when at the beginning of an experiment the speed is at zero. This is because the acquisition system has already started recording while the driver has not yet begun driving, and vice versa when the driver has stopped driving. Finally, the missing data, null values, and duplicates were eliminated.

3.3.1 Relative scoring

Each row represented a sequential snapshot of the acquired driving data. If a driver failed to fulfil any of the safety standards mentioned in Table 1 at the specified engine runtime, the relevant column was given a score of 1 (aggressive), indicating that the driver was dangerous at that time. If not, the row receives a score of 0 (safe). This feature scoring method is seen to be used through these studies [13,15]. This feature scoring as shown in Figure 2 is implemented as a benchmark for our clustering method. Table 1 is constructed based on safe-driving suggestion metrics [17,18].

3.3.2 Data Labelling

The dataset lacks a class label. Hence, K-means clustering, and Silhouette Indexes are used to create different partitions describing the drivers' behaviour (safe/aggressive). Six (6) features are used to create the cluster: engine load, RPM score, speed, throttle position score, speed-RPM ratio score and acceleration.

Table 1
Features formula metrics

Feature name	Formula	Significance
Modified Speed *(km/h)*	$\frac{Speed}{MaxSpeed}$ The default max speed is 220 km/h	It is used to generate a rational speed-RPM ratio.
Modified RPM	$\frac{RPM}{MaxRPM}$ The default max RPM is 8000.	It is used to generate a rational speed-RPM ratio.
Speed-RPM ratio	$\frac{Mod.Speed}{Mod.RPM}$	A value between 0.9 to 1.3 is considered good in terms of gear performance
Acceleration *(m/s2)*	$\frac{(Speedt2 - Speedt1) \div 3.6}{t2 - t1}$	T1 and T2 represent the engine run time for 2 continuous driving indexes. 3.6 is used to convert km/h to *m/s2*

3.3.2.1 K-means

K-Means clustering is an unsupervised learning technique that clusters the unlabelled dataset into several clusters [19,20]. It allows us to cluster the data into several groups and provides a quick method for discovering the categories of groups in an unlabelled dataset without the requirement for training. It is a centroid-based technique in which each cluster has a corresponding centroid. This algorithm's primary objective is to reduce the total distance between each data point and its matching cluster. The method receives as input the unlabelled dataset, splits the dataset into k clusters, and continues the procedure until the optimal clusters cannot be identified. This labelling method enables the use of supervised machine learning algorithms and the evaluation of their performance on the provided dataset.

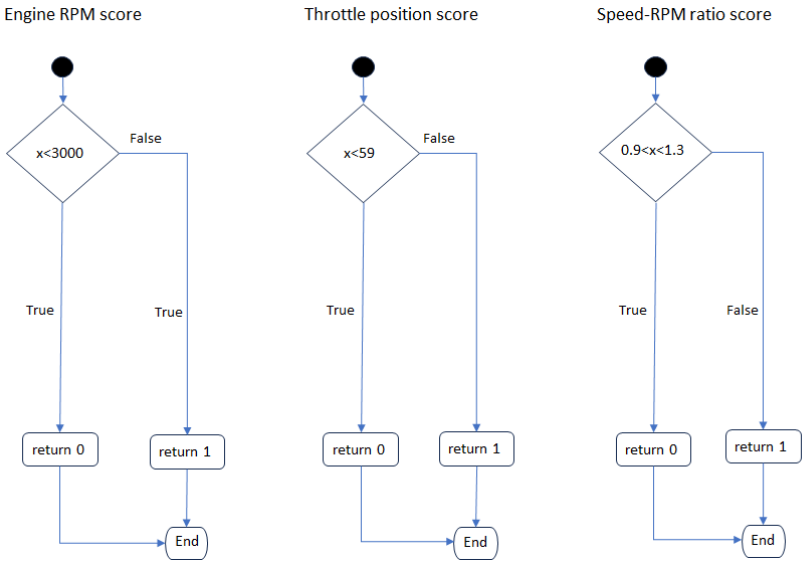

Fig. 2. Relative scoring figure

3.3.2.2 Silhouette indexes (SI)

Using silhouette analysis, one may examine the distance between the generated clusters. The silhouette plot provides a visual method for evaluating factors such as the number of clusters by displaying a measure of how near each point in one cluster is to points in surrounding clusters. Therefore, this analysis is used to validate the number of clusters resulting from the K-Means.

3.4 Classification using Supervised Learning Techniques

Using the labelled dataset created by the clustering procedure described in Section 3.3.2, the model is trained. Twelve (12) features are utilised to train the model: engine load, engine RPM, engine RPM score, speed, engine runtime, throttle position, throttle position score, modified speed, modified RPM, speed-RPM ratio, speed-RPM ratio score, acceleration and the clustered label (safe (0) /aggressive (1)). Multiple algorithms have been used to train the model, which is described below.

3.4.1 Decision tree (DT)

A decision tree (DT) is a supervised learning approach utilised for classification and regression issues, but mostly for classification problems. A decision tree divides a dataset based on certain conditions. Using a decision tree, a training model is created that can predict the class or value of the target variable by learning fundamental rules from training data.

3.4.2 Support vector machine (SVM)

Support Vector Machine (SVM) is a technique for supervised learning used to solve classification or regression problems [21]. It is used to determine the optimal line or decision boundary for classifying spaces. This optimal decision boundary is referred to as a hyperplane. This study uses the Linear SVM. This allows the algorithm to categorise the data into two distinct categories using a single straight line.

3.4.3 Multi-layer perceptron (MLP)

Multi-Layer Perceptron (MLP) is an example-learning artificial intelligence algorithm. This is one of the classification models that we have used to train the model. MLPClassifier is a library function that employs supervised learning. The layers of MLP are interconnected. Except for the input layer, all nodes' activation functions are nonlinear. Between the input and output layers, there might be nonlinear hidden layers. The input layer takes input, hidden layers give abstraction levels, and the output layer predicts outcomes.

3.4.4 Ensemble algorithms
3.4.4.1 AdaBoost (AB)

AdaBoost (AB) is a classic machine-learning classification technique. The ensemble method, also known as Adaptive Boosting, is a technique used in machine learning [22]. The approach used most frequently with AdaBoost is one-level decision trees or decision trees with a single split. It is referred to as Adaptive Boosting because the weights are reassigned to each instance, with more weights allocated to instances that were inaccurately categorised.

3.4.4.2 Random Forest (RF)

Random Forest (RF) is an established supervised machine learning technique employed for Classification and Regression. More trees make a forest more robust. Similarly, the accuracy and problem-solving skills of a Random Forest Algorithm improve with its tree count. Random Forest is a classifier that increases predicted accuracy by combining several decision trees on subsets of a dataset. Ensemble learning is used to address complex issues and improve model performance.

3.4.4.3 Linear combination (LC)

Using linear combination (LC) from the ensemble approaches, the decision tree algorithm and the SVM algorithm are combined into a single evaluation model. This step essentially mathematically averages the total of class probabilities obtained from both models using the predict prob() function. The following equation, Eq. (1) might be used to evaluate this method's ability to predict a score:

$$1n * f(x) + 1n * g(x) \quad (1)$$

where *f(x)* and *g(x)* are the scores obtained from both models predict proba() functions. And *n* denotes the number of models to be merged, which in this case would be two as two models are being utilised. This equation's yield result will be utilised as the second argument for determining the combined model's accuracy score.

3.4.4.4 Weighted linear combination (WLC)

The weighted linear combination (WLC) ensemble method was used, like the method above. The only difference is the weight multiplied by the predict_proba() score of each model, *f(x)* and *g(x)*. The weight of each model is determined by the training score provided by the .score() function when training data is used. Then, the scores of both models are normalised, and each normalised score corresponds to their weight. The formula of the study is as shown in Eq. (2):

$$w_1 * f(x) + w_2 * g(x) \quad (2)$$

where w_1 and w_2 represent the corresponding weights of each model, and *f(x)* and *g(x)* represent the scores obtained by the predict_proba() function for each model. Like the previous ensemble approach, the yield output will be employed as the second argument for calculating the precision score for this weighted linear combination model.

4. Results and Discussions

4.1 Clustering Model Evaluation

By doing the relative scoring, three new features including engine RPM score, throttle position score and speed-RPM ratio score are generated. These features along with other selected features such as engine load, speed and acceleration are then used during the clustering process by the K-Means algorithm. The best number of clusters "K" is 2, for all possible combinations of feature values validated by elbow methods as well as the Silhouette Index as shown in Table 2.

Table 2
Silhouette Index n values

Parameters	Silhouette Index
SI n=2	0.4501
SI n=3	0.4211
SI n=4	0.4089
SI n=5	0.4000

This score can help in determining the number of classes that should be selected as well. Hence, there are two different classes: safe (0) and aggressive (1). Even though the interpretability of the clustering procedure is poor due to the reason that not all drivers are always aggressive, considering the limitations of the dataset, we have printed out our inertia value (2947659.59) and made sure that all the samples were correctly labelled. The output clusters are more robust as it is based on the safety score criteria and simultaneously make sense of all relevant features.

4.2 Performance Evaluation of the Supervised Learning Techniques

The dataset is divided for training and testing purposes. The test size is set to 25% and the train size is set to 75%. After the models have been trained and tested, their accuracy is used to evaluate them. The confusion matrix from the sklearn.metrics library is used to calculate these scores. Detailed below are the relevant scores for the Decision Tree, SVM, MLP, AdaBoost, and ensemble algorithms. Table 3 shows the accuracy for Decision Tree and SVM, MLP, AdaBoost, Random Forest, LC and WLC. As a result, it has been determined that the SVM is less accurate than the Decision Tree. SVM uses a kernel approach to address non-linear problems, whereas decision trees construct hyper-rectangles in input space to solve the problem. For categorical data and addressing collinearity, decision trees are superior to SVM. As for MLP and Decision Tree, the results indicate that MLP is less accurate than Decision Tree. As both find non-linear solutions and have interaction between independent variables, the difference in accuracy is because decision trees perform better when there is a large number of categorical values in the training data, whereas MLP outperforms decision trees when sufficient training data is available.

The research then continues by evaluating the ensemble methods-obtained models. Between AdaBoost and Random Forest, shows that AdaBoost has better accuracy than Random Forest. This is because AdaBoost is often much better at making accurate classifications. Meanwhile, the accuracy of the training set for DT and SVM is used to obtain the accuracy for LC and WLC. The Linear Combination algorithm does slightly better than the Weighted Linear Combination algorithm, by roughly 0.0011 points. Therefore, it has been demonstrated that Ensemble Algorithms achieve higher accuracy test results than single models.

Table 3
Testing set accuracy scores

	DT	SVM	MLP	Ensemble			
				AB	RF	LC	WLC
Accuracy	98.72	96.38	91.02	99.48	99.36	98.43	98.54

5. Conclusion

The purpose of this study is to classify drivers according to their driving styles. We integrate supervised and unsupervised approaches to accomplish our learning objectives. In contrast to previous research, the dataset was labelled with relative scores using either row and segment

labelling or row scoring approaches. The novel contribution of this work is the application of relative scores prior to K-Means to validate the K-Means results. Based on the algorithms used, it is verified that ensembled models will generate different results than single models and that ensemble algorithms play a significant role in producing more accurate results. DT (98.72%) has the best accuracy score among single models, followed by SVM (96.38%) and MLP (91.02%). The ensemble model with the best accuracy is AB (99.48%), followed by RF (99.36%), WLC (98.54%), and LC (98.43%). Notably, the experiment carried out in this study has certain limitations. Since the primary objective of this study is to classify driving behaviours based on OBD-II data, the dataset is limited to information obtained solely from the OBD-II and excludes any external elements related to the drivers' behaviour. This paper has a significant capability and opportunity for improvement. Future works will include the collection and creation of our own dataset with additional sensors and suitable features for driving data. In addition, more neural network algorithms and accuracy score comparisons are required. Finally, we plan to develop a system that uses machine learning to analyse deep driving behaviour.

Concerning the United Nations Sustainable Development Goals (SDG), this study targets SDGs 11 as well as 12. The classification of driving behaviour was developed to assist in making cities and human settlements more inclusive, safe, resilient, and sustainable, as well as to reduce gas emissions that might lead to pollution. From the patterns of driving behaviours, behavioural issues such as road rage, drunk driving, reckless driving, and many more can be solved with communication/awareness on road traffic safety campaigns by the higher authority. Not only that, from the OBD-II dataset, integrating mathematical calculations and machine learning helps to reduce gas emissions by creating models to estimate gasoline fuel consumption. As a result, we believe that our initiative will contribute to a change in driving behaviour and have a good impact on the environment.

Acknowledgement
This research was funded by a grant from the Ministry of Higher Education of Malaysia (FRGS-RACER Grant ref no. RACER/1/2019/ICT03/UIAM//1 (no. RACER19-006-0006)).

References

[1] Holstein, Tobias, Gordana Dodig-Crnkovic, and Patrizio Pelliccione. "Ethical and social aspects of self-driving cars." *arXiv preprint arXiv:1802.04103* (2018). https://doi.org/10.29007/mgcs

[2] Ameen, Hussein Ali, A. K. Mahamad, S. Saon, M. A. Ahmadon, and S. Yamaguchi. "Driving behaviour identification based on OBD speed and GPS data analysis." *Advances in Science Technology and Engineering Systems Journal* 6, no. 1 (2021): 550-569. https://doi.org/10.25046/aj060160

[3] Suhaimin, Khairul Nizam, Wan Hasrulnizzam Wan Mahmood, Zuhriah Ebrahim, Halimaton Hakimi, and Syafiq Aziz. "Human Centric Approach in Smart Remanufacturing for End-Life-Vehicle (ELV)'s Stabilizer Bar." *Malaysian Journal on Composites Science and Manufacturing* 12, no. 1 (2023): 1-12. https://doi.org/10.37934/mjcsm.12.1.112

[4] Kabir, Afrida, Faiyaj Kabir, Saief Newaz Chowdhury, and AR M. Harunur Rashid. "Design, Simulation and Fabrication of an Ergonomic Handgrip for Public Transport in Bangladesh." *Malaysian Journal on Composites Science and Manufacturing* 13, no. 1 (2024): 98-111.

[5] Masrom, Maslin, Mohd Nazry Ali, Wahyunah Ghani, and Amirul Haiman Abdul Rahman. "The ICT implementation in the TVET teaching and learning environment during the COVID-19 pandemic." *International Journal of Advanced Research in Future Ready Learning and Education* 28, no. 1 (2022): 43-49.

[6] da Silveira Barreto, Cephas Alves, Joao C. Xavier-Júnior, Anne MP Canuto, and Ivanovitch MD da Silva. "A machine learning approach based on automotive engine data clustering for driver usage profiling classification." In *Anais do XV Encontro Nacional de Inteligência Artificial e Computacional*, pp. 174-185. SBC, 2018. https://doi.org/10.5753/eniac.2018.4414

[7] Kabra, Amit. "Clustering of Driver Data based on Driving Patterns." (2019).

[8] Mahadi, M. I., Nurulhuda Zainuddin, N. B. Shah, NurAsyira Naziron, and S. F. Rum. "E-halal restaurant recommender system using collaborative filtering algorithm." *Journal of Advanced Research in Computing and Applications* 12, no. 1 (2018): 22-34.

[9] Abdelrahman, Abdalla Ebrahim, Hossam S. Hassanein, and Najah Abu-Ali. "Robust data-driven framework for driver behavior profiling using supervised machine learning." *IEEE transactions on intelligent transportation systems* 23, no. 4 (2020): 3336-3350. https://doi.org/10.1109/TITS.2020.3035700

[10] Ferreira, Jair, Eduardo Carvalho, Bruno V. Ferreira, Cleidson de Souza, Yoshihiko Suhara, Alex Pentland, and Gustavo Pessin. "Driver behavior profiling: An investigation with different smartphone sensors and machine learning." *PLoS one* 12, no. 4 (2017): e0174959. https://doi.org/10.1371/journal.pone.0174959

[11] Peppes, Nikolaos, Theodoros Alexakis, Evgenia Adamopoulou, and Konstantinos Demestichas. "Driving behaviour analysis using machine and deep learning methods for continuous streams of vehicular data." *Sensors* 21, no. 14 (2021): 4704. https://doi.org/10.3390/s21144704

[12] Kumar, Anuj. "Driver Usage Risk Profiling by Analyzing Vehicle Driving Behavior using Machine Learning Model Based on Vehicular Cloud Telematics Data." PhD diss., Dublin, National College of Ireland, 2018.

[13] Al-Hussein, Ward Ahmed, Lip Yee Por, Miss Laiha Mat Kiah, and Bilal Bahaa Zaidan. "Driver behavior profiling and recognition using deep-learning methods: In accordance with traffic regulations and experts guidelines." *International journal of environmental research and public health* 19, no. 3 (2022): 1470. https://doi.org/10.3390/ijerph19031470

[14] Puchalski, Andrzej, and Iwona Komorska. "Driving style analysis and driver classification using OBD data of a hybrid electric vehicle." *Transport Problems* 15 (2020). https://doi.org/10.21307/tp-2020-050

[15] Fan, Yudong. "Eco-driving Analysis and Driving Style Feedback for Passenger Cars Using OBD-II Data and Machine Learning," (2022). https://www.cs.vu.nl/~versto/VU-CS-BSc-MSc-Theses/VU-CS-BSc-Thesis-Fan-Yudong-2022.pdf

[16] Kaggle. "OBD-II datasets." *Kaggle*. https://www.kaggle.com/datasets/cephasax/obdii-ds3

[17] Chen, Shi-Huang, Jeng-Shyang Pan, and Kaixuan Lu. "Driving behavior analysis based on vehicle OBD information and adaboost algorithms." In *Proceedings of the international multiconference of engineers and computer scientists*, vol. 1, pp. 18-20. 2015.

[18] Liu, Tianshi, Guang Yang, and Dong Shi. "Construction of driving behavior scoring model based on obd terminal data analysis." In *2020 5th International Conference on Information Science, Computer Technology and Transportation (ISCTT)*, pp. 24-27. IEEE, 2020. https://doi.org/10.1109/ISCTT51595.2020.00012

[19] Javatpoint. "K-Means Clustering Algorithm." *Javatpoint*. https://www.javatpoint.com/k-means-clustering-algorithm-in-machine-learning

[20] Mishra, Sanatan. "Unsupervised learning and data clustering." *Towards Data Science* 19 (2017). https://doi.org/10.1016/B978-0-12-811654-8.00003-8

[21] Javatpoint. "Support Vector Machine (SVM) Algorithm." *Javatpoint*. https://www.javatpoint.com/machine-learning-support-vector-machine-algorithm

[22] Choudhury, Ambika. "Basics of Ensemble Learning In Classification Techniques Explained," https://analyticsindiamag.com/basics-of-ensemble-learning-in-classification-techniques-explained