Transportation Engineering 22 (2025) 100384

Contents lists available at ScienceDirect

Transportation Engineering

journal homepage: www.sciencedirect.com/journal/transportation-engineering

How does the built environment affect driver behaviours? A methodological framework for large-scale analysis that combines Onboard Diagnostic and Geographical Data to promote Road Safety

V. Nicolosi [a,*], M. Mameli [b], S. Shiralizadeh [b], I.G. Coltea [b], M. D'Apuzzo [c], G. Cappelli [a,c]

[a] University of Rome Tor Vergata, Depatment of Enterprise Engineering „Mario Lucertini", Via del Politecnico 1, 00133 Rome, Italy
[b] Key 2 Business S.r.l., via del Viale dell'Esperanto 71, 00133 Rome, Italy
[c] University of Cassino and Southern Lazio, Department of Civil and Mechanical Engineering, Via G. Di Biasio 43, 03043 Cassino, Italy

ARTICLE INFO

Keywords:
On-board Diagnostic (OBD)
Geographic data
Driver behaviors
Road safety
Data fusion
Multi-layer perceptron (MLP) neural network
Shapley additive explanations (SHAP) analysis

ABSTRACT

Background: Driving behaviours are a set of actions made by drivers during their driving tasks and could be detected with the help of onboard sensors in vehicles. The interest from researchers in evaluating fuel consumption and enhancing overall safety has increased in recent years, thanks to the availability of algorithms and methods that can manage the large amount of data obtained from sensors.
Method: In this paper, both Onboard Diagnostic (OBD) data from the DDD20 dataset and open-source Geographic Data are used to study driver behaviours. After the preprocessing phases, a Multi-Layer Perceptron (MLP) classifier is used to classify driver behaviours, and a Shapley Additive Explanations (SHAP) Analysis is implemented to perform feature selection.
Results: Three models are created: the first combines OBD and geographical data, the second is based only on geographical data, and the last one contains a reduced subset of features retrieved from geographical data. The model with a reduced number of features shows good prediction accuracy, comparable with the full previous models. In addition to that, SHAP analysis highlights how the presence of schools, hospitals, bridges, parking, subways, cycleways, and footways increases the likelihood of having aggressive driver behaviour.
Conclusions: This study aims to show how the external context influences driver behaviours and to create a methodological framework for future developments in road safety. The model that uses only open-source geographical data with a reduced number of features is particularly suited for large-scale analysis in the context of road safety.

1. Introduction

Identifying the driver profile, which allows for categorizing driver behaviour [1,2], may ensure road traffic safety and reduce energy consumption. The terminology driver profile usually refers to a group of drivers that share similar driving characteristics and behaviours, while a driving pattern could be defined as a specific behavior that occurs often among one or more drivers [3].

Shahverdy et al. [4] define driving behavior as the set of actions exhibited by individuals during driving tasks, which can be categorized into safe, aggressive, distracted, drowsy, and drunk driving. To effectively perform this classification, both non-visual indicators (such as vehicle speed and acceleration [5–7] and visual cues (including eye movement patterns [8,9]) are considered valuable. As highlighted by the same authors [4], due to the high complexity of the classification task, alongside the large amount of data required, machine learning algorithms are among the most commonly adopted approaches for behavior classification, given their ability to process and interpret large volumes of data collected from vehicle sensors and onboard systems.

Thanks to the introduction of effective computer technologies and digitalized and connected vehicles, alongside machine learning algorithms and Information and Communication Technology (ICT), the mechanisms behind driver behaviours have been highlighted and discovered [10]. According to some authors, 90 % of road accidents could be attributed to driver behavior [1,11,12], so it is crucial to be aware that drivers make mistakes [13,14] and understanding road

---

[*] Corresponding author.

E-mail addresses: nicolosi@uniroma2.it (V. Nicolosi), m.mameli@key2.it (M. Mameli), s.shiralizadeh@key2.it (S. Shiralizadeh), i.coltea@key2.it (I.G. Coltea), dapuzzo@unicas.it (M. D'Apuzzo), giuseppe.cappelli1@unicas.it, cppgpp01@uniroma2.it (G. Cappelli).

https://doi.org/10.1016/j.treng.2025.100384
Received 20 May 2025; Received in revised form 28 August 2025; Accepted 28 August 2025

human behavior is key to improving road safety.

Another aspect that is related to road safety is fuel consumption. There is evidence in the current literature review that highlights how eco-driving profiles reduce both fuel consumption and risky driver behaviours [15–17].

According to Yen et al. [16], safe and low speeds, calm behavior during the driving task, in addition to appropriate route planning, are some of the most important contributory factors related to both safety and fuel consumption issues. An economical and “sustainable” driving style, characterized by smooth acceleration without sudden, huge, and frequent accelerations, may lower fuel consumption up to 4.83 l/100 km [18]. On the contrary, more frequent and higher acceleration, typical of a dynamic and “not-eco-friendly” driving style, could increase fuel consumption, reaching the mean value of 6.15 l/100 km. The key role of cinematic variables is also highlighted in the study of Graba et al. [19]: the authors found that the energy consumption increases by 0.15 J $(\text{kg m})^{-1}$ for each increment of 0.1 m/s^2 of the mean acceleration.

Several data sources could be used in the driver behavior profiling phase. Surveys and questionnaires, driving simulators, site-based observation, and naturalistic studies are the most common data sources for identifying driver profiling [20]. Surveys and questionnaires have often been used in the past and are based on self-reported data [21]. The main drawback related to such data is their subjectivity, and for this reason, they are not often considered valid by researchers [22,23]. On the other hand, a driving simulator represents a way to collect data in a safe and controlled environment [24]. As criticized by some researchers [25,26], participants tend to behave in a driving simulator differently than in a real driving scenario [25] and pursue more aggressive driving behaviour [26]. To close the gap between a simulated and real scenario, naturalistic driving studies are conducted in a real environment and under uncontrolled conditions [20,21]. In this setting, data on drivers, vehicles, and the external environment are collected [27]. Lastly, in site-based observation [21], cameras are mounted on the building or on the traffic lights to collect data on drivers [20]. As highlighted in the scientific literature, this data is not particularly precise, and the selection of specific sites could be detrimental to the generalization of driver behaviors [20,26].

The increased attention paid to driver behaviours could also be explained by the spread and improvements of sensors and equipment for the collection of the needed data [28–30]. Usually, these sensors are classified into three categories [10,31]. The first category refers to in-vehicle data recording systems based on Controller Area Network Bus (CAN bus) and On-board Diagnostic (OBD) interfaces [32]. Smartphones, and especially built-in sensors such as accelerometers, gyroscopes, GPS, and magnetometers, are often implemented in data collection phases due to their widespread availability and reduced costs compared to other technologies [33]. According to Okmi et al. [34], in their literature review, the authors find that recently there has been a huge growth of studies that use mobile phones (equipped with reliable sensors) to study human mobility in different contexts, from health monitoring to urban planning [35]. Lastly, it is worth mentioning behaviour detection methodologies that are based on real-time models, in particular, Machine Learning algorithms [24]. According to the current literature [36,37], due to the high amount of data generated by sensors and equipment, the most implemented analysis methods, after the pre-processing phase, are Decision Tree, Support Vector Machines (SVMs), Bayesian Learners, Ensemble Learners, Fuzzy Logic, K-Nearest Neighbour and Neural Networks to classify driver behaviours and styles.

Several functions in vehicles are monitored through electronic control units (ECUs) [38]. All these units interact with the area network (CAN) [39,40], and the onboard diagnostic module (OBD) gets all the data from the ECUs [41,42]. The data gathered from the OBD are often used to classify driver behaviours [42–44] and to evaluate fuel consumption [10,45,46]. The huge amount of data generated from sensors is usually handled with the implementation of Machine Learning Algorithms, which can face this complexity [10,47,48].

In line with the current literature, this paper aims to study and classify driver behavior using data collected through OBD. Specifically, the DAVIS Driving Dataset 2020 (DDD20) is utilized to fulfill this task. In addition, this dataset is enriched with geographical data to explore potential relationships between environmental characteristics and driver behaviors. Three different Multi-Layer Perceptron Neural Network Models are implemented to perform the classification: the first model (Model 1) combines OBD and geographical data, the second one (Model 2) only geographical data, and the last one (Model 3) only geographical data with a reduced number of features. In addition to that, Shapley Additive Explanations (SHAP) Analysis is performed to highlight the role of the single contributory factors and assess whether models that rely only on geographical data could be reliable in the driver behaviour prediction task.

2. Method

This section proposes the overall methodology applied to identify driving styles in a geographical area (see Fig. 1). The overall model is based on using OBD data from DDD20 to analyze and label driving styles. Road information is added to this dataset by using OpenStreetMap (OSM), and weather information can be retrieved from the OSM system [49] and via APIs from the selected meteorological portal [50].

After collecting all the information, different pre-processing procedures for each data source are applied to create the input structure, used to train a reliable model able to identify driving styles based on road context information. After the first training, the Shapley Additive exPlanations (SHAP) technique is applied as part of the feature selection process to identify the most influential features to check for the model’s ability to identify possible driving styles in the road context.

Finally, only the most influential features are selected to re-train the model, to make a light and accurate model with this reduced subset of features. The following sections will discuss in detail the main steps shown in Fig. 1.

2.1. Dataset description

As the first step of the method described in Fig. 1 in the previous section, data collection represents one of the most crucial aspects of every research. The structures, the dimensions, and the features that are included affect the methodology, models, and techniques to implement in the process analysis. In this paper, two main datasets are used to study driver behaviour. The first is data from OBD devices, which provide detailed vehicle measurements, including actual speed, engine rotation per minute (RPM), fuel level, and engine temperature. The second dataset is obtained from OpenStreetMap, which includes spatial information such as GPS coordinates, altitude, estimated speed, and contextual data on the road network (e.g., speed limits, type of road).

Regarding the OBD data, a comparative analysis of the main dataset available is performed to highlight and then choose the most adequate dataset. After this preliminary analysis, the DDD20 dataset was chosen [51]. The DDD20 represents a comprehensive end-to-end driving dataset specifically designed for neuromorphic event camera research in autonomous driving applications. This dataset was developed to address the critical limitations of conventional cameras in challenging lighting conditions, where traditional vision systems often fail to provide reliable visual information for driving decisions. The main parameters that are included in the dataset are reported in Table 1.

DDD20 contains 51 h of recordings collected over approximately 4000 km of different driving scenarios, representing a fourfold expansion from its predecessor DDD17. The dataset encompasses 1.3 terabytes of data captured using a DAVIS346B neuromorphic camera, which uniquely provides concurrent streams of both Dynamic Vision Sensor (DVS) brightness change events and Active Pixel Sensor (APS) intensity frames from the same optical pathway. The DVS component generates

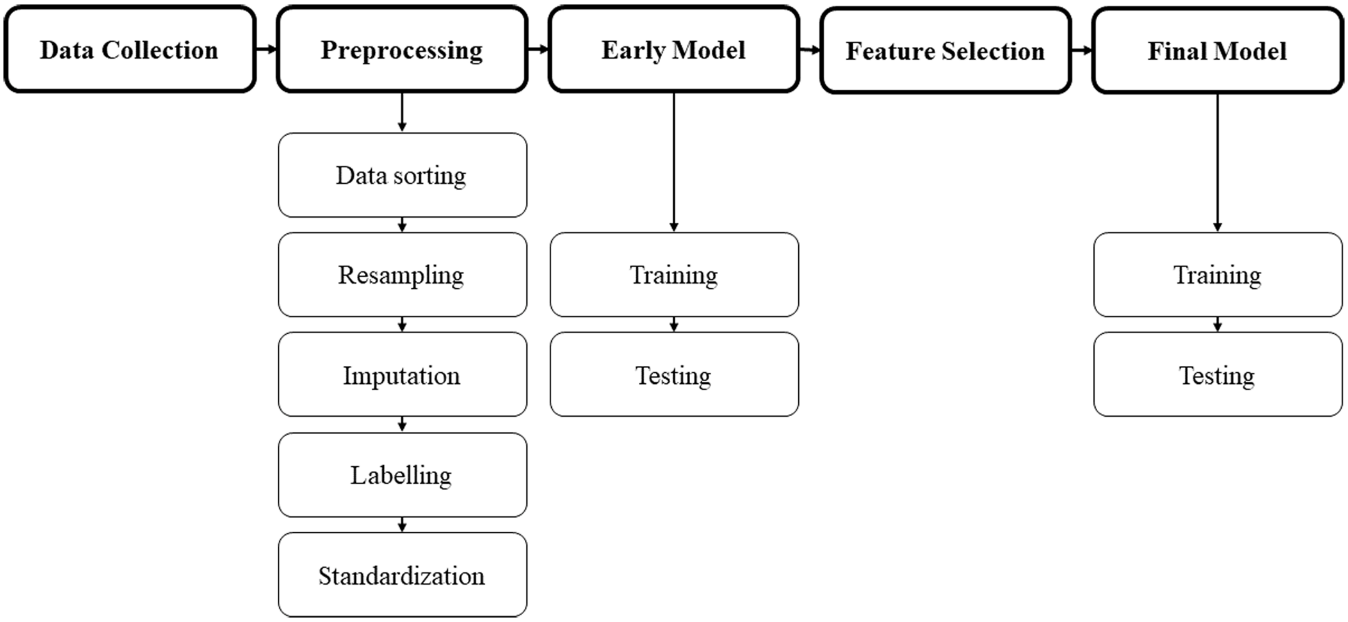

Fig. 1. Flow chart of the proposed methodology.

Table 1
Main available parameters in the DDD20 dataset with a description from the On-Board Diagnostics (OBD).

Feature	Description
Mean accelerator pedal position	Average percentage of accelerator pedal pressure
Mean engine speed	Average engine rotations per minute (RPM)
Mean vehicle speed	Average velocity of the vehicle
Mean torque at the transmission	Average torque delivered to the transmission system
Number of gear position changes	Count of transmission gear changes
Brake pedal status variations	Number of brake pedal state changes
Fuel consumption difference	Variation in fuel consumption
The sum of accelerator position variations	Cumulative changes in accelerator pedal position
The sum of vehicle speed variations	Cumulative changes in vehicle speed
The sum of engine speed variations	Cumulative changes in engine RPM
The sum of steering wheel angle variations	Cumulative changes in steering angle

asynchronous, timestamped address events triggered by local brightness changes at individual pixels, offering 120 dB dynamic range compared to the 60 dB range of conventional cameras. These events are transmitted with submillisecond latency and include pixel coordinates, brightness change polarity, and microsecond timestamps. The APS component captures standard grayscale intensity frames at variable rates between 8–50 fps using global shutter technology, with exposure duration optimized through auto-exposure algorithms targeting road surface visibility.

The data collection methodology employed a DAVIS346B camera with 346 × 260-pixel resolution mounted behind the vehicle windshield using a glass suction tripod. The camera utilized a 4.5 mm Kowa lens providing 71° horizontal field of view, optimized for capturing road features during turning maneuvers. Data acquisition was performed using custom Python software that synchronized multiple data streams through system timestamps, with camera data precision-timestamped at the microsecond level.

Vehicle telemetry data was concurrently captured at approximately 10 Hz through OpenXC Ford Reference vehicle interfaces connected to the CAN bus of test vehicles, including a 2015 Ford Mondeo for European recordings and a 2016 Ford Focus for US recordings. This telemetry encompasses comprehensive driving parameters including steering wheel angles up to ±720°, vehicle speed, accelerator pedal position, brake pedal status, engine speed, transmission gear position, and environmental indicators such as headlamp and windshield wiper status. Critically, the dataset includes precise geospatial coordinates through latitude and longitude measurements along with odometer readings, enabling spatial analysis of driving behaviors.

The recordings span diverse environmental and temporal conditions across multiple geographic regions. Data collection occurred across rural highway driving under challenging sunlight glare conditions, urban environments in Los Angeles and San Diego, and demanding mountain highway segments, including Colorado’s Lizard Head Pass and California’s Angeles Crest Highway. Temporal diversity was achieved through systematic collection during daylight, evening, and nighttime conditions, with recordings distributed across approximately 60 days of intermittent data capture. The dataset comprises 215 individual recordings with durations ranging from several minutes to over one hour, with a median length of approximately 700 s.

The integration of DDD20 with OSM geographic data significantly enhances the dataset’s analytical capabilities by providing rich contextual information about the road environment. This geographic enhancement enables correlation of driving behaviors with specific infrastructure characteristics, including road type classifications, posted speed limits, and geometric properties of the roadway. The OSM integration identifies the presence and locations of traffic control elements such as intersections and roundabouts, enabling detailed analysis of complex driving maneuvers and decision-making processes in challenging scenarios. Furthermore, the geographic context includes proximity analysis to significant points of interest, including schools, commercial areas, residential zones, and other landmarks that may influence driving patterns and behaviors. In Appendix A, all the features extracted from the OSM database and used as possible variables related to driver behaviour are reported and shown.

This enhanced geographic context transforms DDD20 from a pure sensory dataset into a comprehensive platform for understanding the relationship between road infrastructure, environmental conditions, and human driving decisions. The combination of high-precision neuromorphic visual data, detailed vehicle telemetry, and rich geographic context enables researchers to investigate how different road characteristics and environmental factors influence driving strategies, particularly under the challenging lighting conditions where event cameras demonstrate clear advantages over conventional vision systems. The dataset's proven effectiveness has been demonstrated through steering prediction experiments, where fusion of DVS and APS data achieved an explained variance of 0.88 for human steering prediction, significantly outperforming single-modality approaches and establishing DDD20 as a valuable resource for advancing neuromorphic vision research in autonomous driving applications.

2.2. Preprocessing

Data pre-processing ensures that data is clean, structured, and ready to be analyzed, improving the reliability and effectiveness of analyses and models. This section explores the main pre-processing operations needed to analyze data effectively, considering both the requirements of statistical algorithms and Machine Learning models.

When analyzing data from OSM and OBD devices, it is essential to ensure that the information is structured to preserve temporal consistency and optimize the use of different data sources. Two operations crucial for this objective are sorting and resampling.

Data sorting is particularly important when combining datasets from different sources, such as OSM and OBD, in particular synchronizing data sources, calculating time dynamics (e.g., accelerations or decelerations), and segmenting trips or driving sessions.

Resampling consists of realigning the data of the various features to the maximum frequency available, ensuring time consistency and improving the quality of subsequent analyses. Data from OSM and OBD may have different acquisition frequencies.

An unsupervised labeling method using the K-means algorithm has been employed to classify driving styles [52]. This technique was selected for its ability to uncover natural behavioral patterns in the data without predefined labels.

The labeling process involves two primary phases. In the first phase, K-means clustering is applied with $k = 3$ to categorize driving behavior into three main classes: normal, dangerous, and safe driving, as shown in Table 2. This initial classification groups similar driving patterns based on features derived from OBD data.

The second phase assesses the influence of each feature using a ranking system, where scores from 1 to 3 are assigned by comparing feature means across clusters. This process quantifies each feature's significance in distinguishing driving styles and ensures the classification's reliability.

Other preprocessing stages involve the use of One-Hot Encoding [53] to convert categorical variables, such as weather conditions, road types, and points of interest, into a numerical format suitable for Machine Learning algorithms. The MinMaxScaler algorithm [54] is then employed to normalize all features within a 0 to 1 range, enabling fair comparisons between variables with varying scales.

2.3. Multi-Layer Perceptron Neural Network

The model used to perform classification is a Multi-Layer Perceptron Neural Network (MLP-NN), which belongs to the class of Feedforward Neural Networks. Neural Networks are often used as supervised classification algorithms and work by emulating the brain's neural process. These neurons are connected and elaborate the information that is then transmitted as output or to the other neurons [55]. This network is constituted essentially by three layers: input, hidden, and output layers [56]. Such types of neural networks are among the first introduced and are characterized by the fact that the information goes forward through several layers, without any feedback [57]. The MLP-NN works with a superposition of non-linear functions and which allows that the multilayer structure can approximate nonlinear relationships between dependent and independent variables [55]. MLP-NNs are often used as reliable models in pattern recognition tasks and especially when the decision boundaries are not linear [58].

Table 2
Principal Behavioural Classes used for the labeling approach based on the K-means algorithm.

Cluster	Description	Count
0	Safe Driving	79,788
1	Normal Driving	25,027
2	Dangerous Driving	46,226

2.4. Shapley additive explanations (SHAP) for feature selection

The Shapley Additive Explanations (SHAP) analysis [59–61] is an advanced technique for interpreting and explaining machine learning models. It is based on cooperative game theory and uses Shapley values to assign the importance of each feature to the result produced by the model. This methodology is beneficial for understanding the behavior of complex models (e.g., decision trees, neural networks, or ensemble methods) and guiding practical decisions, as well as selecting the most relevant features to train a lighter model that allows efficient use of computational resources.

High-performance Machine Learning models (for example, Gradient Boosting or Deep Learning) are often "black boxes" and difficult to interpret. SHAP explains how features contribute to a specific output, making the model more interpretable to users.

Through the SHAP values, it is possible to quantify the importance of each feature consistently and intuitively. This allows the identification of the most influential features, which can be used to build a lighter model with a reduced number of features, that is simpler and less computationally expensive compared to the previous one.

3. Results

3.1. Model with OBD and geographical data (Model 1)

The Multi-Layer Perceptron Model (MLP) is utilized as a classifier to achieve the classification objective in this paper. In Fig. 1, the pattern of model learning over the 40 training periods is proposed. The blue loss curve shows a rapid initial decrease, starting from about 0.30 and stabilizing around the value of 0.02 after about 10 epochs.Fig. 2

This behavior is indicative of an effective learning process, where the model rapidly improves its predictive ability in the early stages of training. The red curve of the validation accuracy confirms this positive trend (see Fig. 3), showing a constant increase starting from values around 0.95 and reaching performances above 0.98 after the first few periods, then remaining stable with slight oscillations until the end of the training.

Fig. 4 offers a complementary perspective showing the evolution of performance as a function of the number of training examples used. The blue curve represents the score on the training set, while the orange one indicates the cross-validation score. Both curves converge towards very high values (greater than 0.98) as the training examples increase, with the training curve slightly higher than the validation curve. This pattern suggests a good balance between bias and variance. The model shows a robust generalization capability, as evidenced by the small difference between training and validation performance.

The confusion matrix (see Fig. 5) provides a detailed analysis of the classifier's performance on the three classes: "dangerous", "moderate", and "safe". Overall, these results indicate that the MLP Classifier model achieved excellent performance in the classification task, with a very accurate discrimination capability between the three classes, reaching values above 97 % for the classification accuracy.

3.2. Model with only geographical data (Model 2)

The Neural Network's performance analysis with contextual data

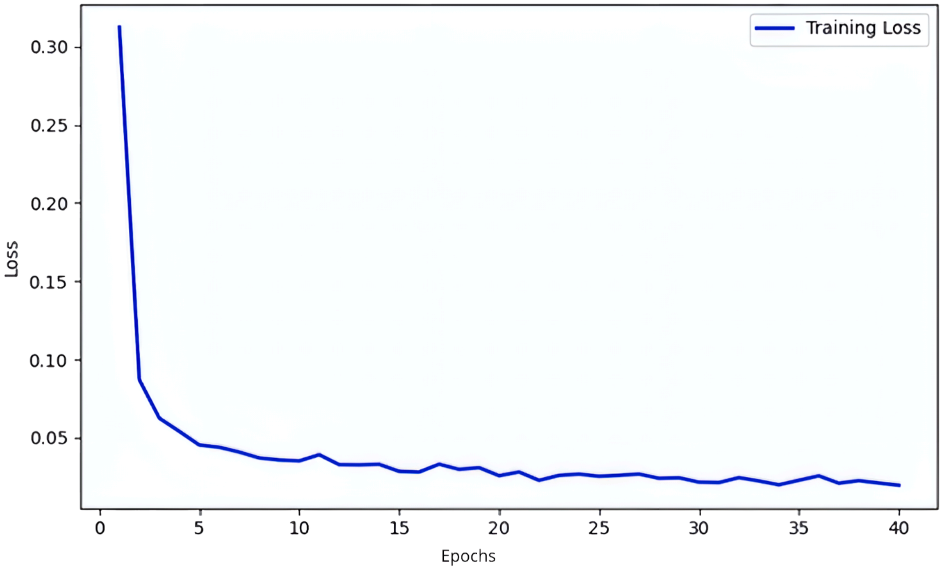

Fig. 2. Trend of the loss function during the model training phase during several epochs.

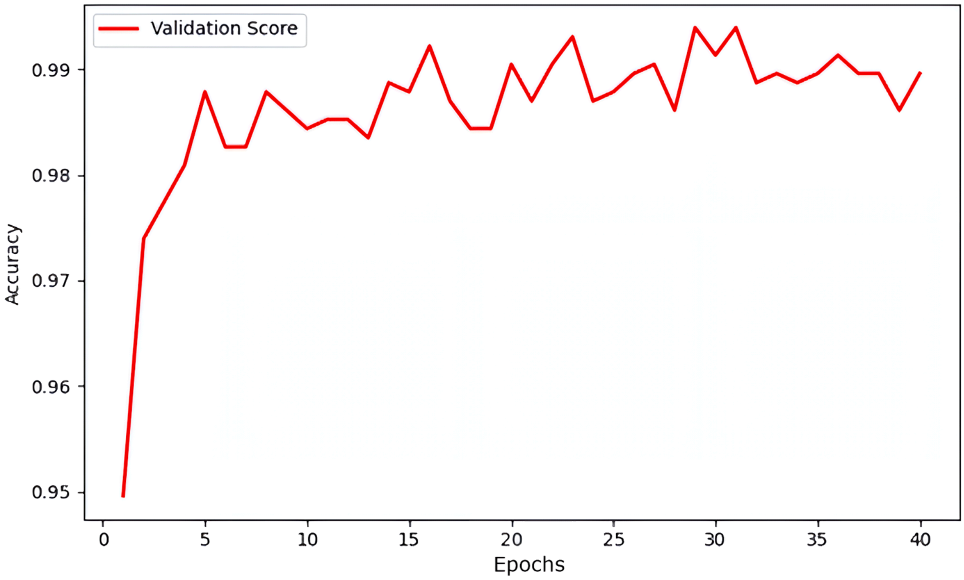

Fig. 3. Trend of the accuracy during the model training phase during several epochs.

alone demonstrates remarkable classification capabilities, suggesting that environmental and infrastructural features provide sufficient information for reliable driving style prediction.

The training process, illustrated in Fig. 6, shows characteristic learning patterns, with the loss function rapidly decreasing from 0.40 to 0.15 within the first 5 epochs, followed by gradual stabilization around 0.08. The validation accuracy demonstrates steady improvement (see Fig. 7), starting from approximately 0.92 and reaching a plateau above 0.96 in later epochs, with minor fluctuations indicating healthy model adaptability.

The learning curves, depicted in Fig. 8, reveal progressive convergence between training and cross-validation performance and an effective generalization without overfitting.

The confusion matrix analysis demonstrates robust classification performance despite the absence of vehicle telemetry data (see Fig. 9) with high values of accuracy rate, starting from 91.8 %.

3.3. Feature analysis

The proposed SHAP method provides a detailed analysis of the relative importance of different characteristics in the classification model. The graphical output of this feature selection algorithm is particularly informative because it shows not only the impact of each feature on the model, but also how that impact varies across different feature values (see Fig. 10).

The characteristics are sorted by decreasing importance, with "poi school" (that indicates the number of nearby schools) at the top showing the most significant impact on the model. The distribution of SHAP values for this characteristic is wide, ranging from about -0.2 to $+0.4$, indicating that the presence of a nearby school can significantly influence risk classification in both positive and negative directions.

The second most influential feature is "slope" (that stands for the slope of the ground), which has an interesting bimodal distribution. This suggests that the slope of the ground has distinctly different effects depending on its severity, with some values contributing positively to

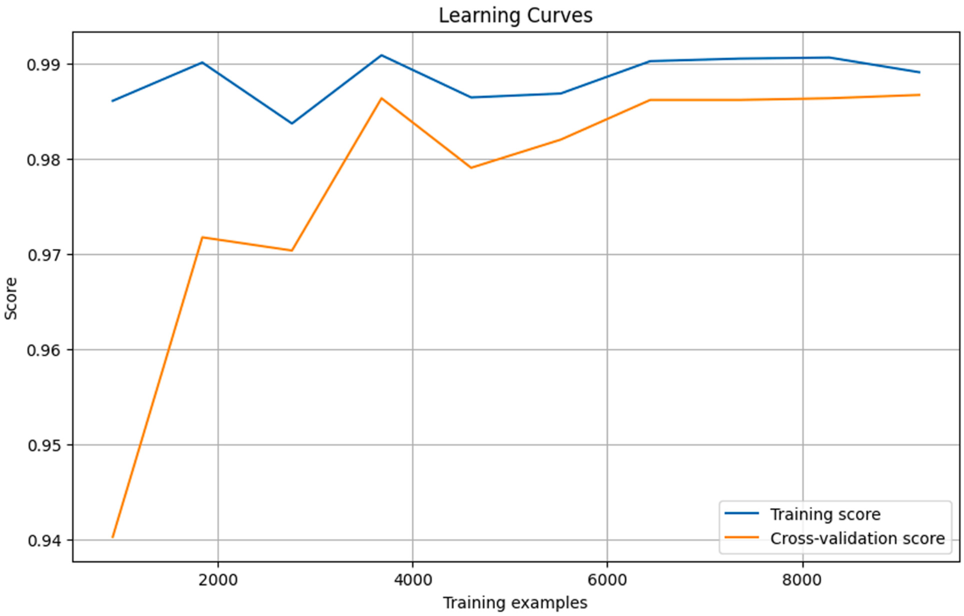

Fig. 4. Trend of the training set score (blue line) and cross-validation score (orange line).

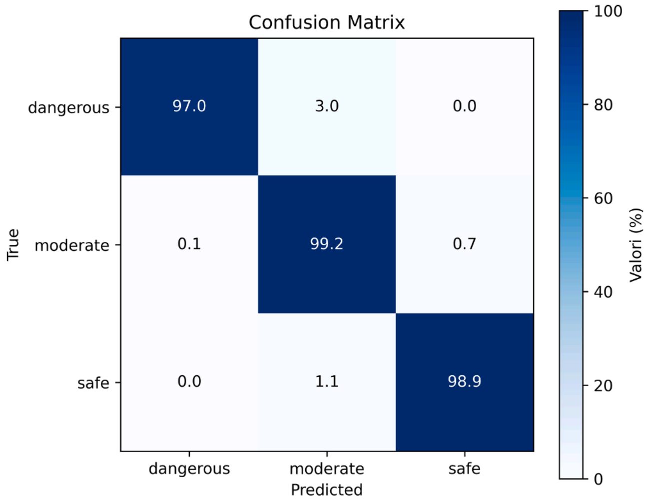

Fig. 5. Confusion matrix of the driving style classification model using a combination of OBD data and context data. The matrix shows that the model correctly classified 97 % of cases as "dangerous", 99.2 % as "moderate", and 98.9 % as "safe". Classification errors are relatively small: 3 % "dangerous" cases classified as "moderate", 0.7 % "moderate" cases classified as "safe", and 1.1 % "safe" cases classified as "moderate".

risk and others negatively.

The characteristics related to infrastructures (presence of bridges, hospitals, and parking show moderate but significant impacts, with asymmetric distributions that suggest that their presence tends to increase the probability of certain classifications.

It is interesting to note that the driving characteristics, such as differences in steering wheel angle, mean engine speed, and brake pedal status, are in the central part of the ranking, with more contained but still relevant impacts. This may suggest that while environmental and infrastructural conditions play a predominant role, vehicle behavior remains a significant factor in risk assessment.

The characteristics at the bottom of the ranking, such as wind degree and circuity, show more limited impacts on the model, with distributions more concentrated around zero, indicating a lesser influence on final predictions.

The blue-pink color in the graph adds an extra dimension to the analysis, allowing to identify how high (pink) or low (blue) values of each feature contribute to the final result. This is particularly evident for features such as "poi school" and "slope", where a clear separation between the effect of high and low values can be observed.

The SHAP analysis of the scenario where the model has been trained excluding OBD data (Model 2) shows significant differences and reveals how the model adapts using available information differently (see Fig. 11).

The presence of bridges ("bridge") emerges as the most influential factor in the model's predictions, with a distribution of SHAP values that extends significantly in both negative and positive directions. This suggests that bridges may represent both a risk and a safety element, probably depending on their specific structural and contextual characteristics.

The slope ("slope") is the second most important characteristic, showing an interesting distribution of SHAP values. The coloration of the dots indicates that higher gradients (in pink) tend to have a positive impact on risk, while more moderate gradients (in blue) help reduce it. This non-linear relationship effectively captures how the slope of the ground can affect road safety.

The presence of schools ("poi school") is the third most influential factor, with a distribution that suggests a significant impact on risk classification. The variability of SHAP values for this characteristic indicates that the specific context in which the school is located (distance, type of road, etc.) modulates its effect on risk assessment.

It is interesting to note that public service infrastructures ("poi hospital", "poi subway", "poi cycleway-footway") have a moderate but significant impact, with distributions suggesting differentiated effects according to the specific context. In particular, the presence of hospitals shows a bimodal distribution, indicating that they can influence risk in different ways depending on their location and accessibility.

3.4. Model with only geographical data and with a reduced number of features (Model 3)

In this section, a new model based on the results obtained after the optimization of the system through the selection of input variables by the SHAP methodology is proposed (Model 3).

Looking at the graph in Fig. 12, this trend suggests that the model, despite the reduction of features, maintains a good learning ability,

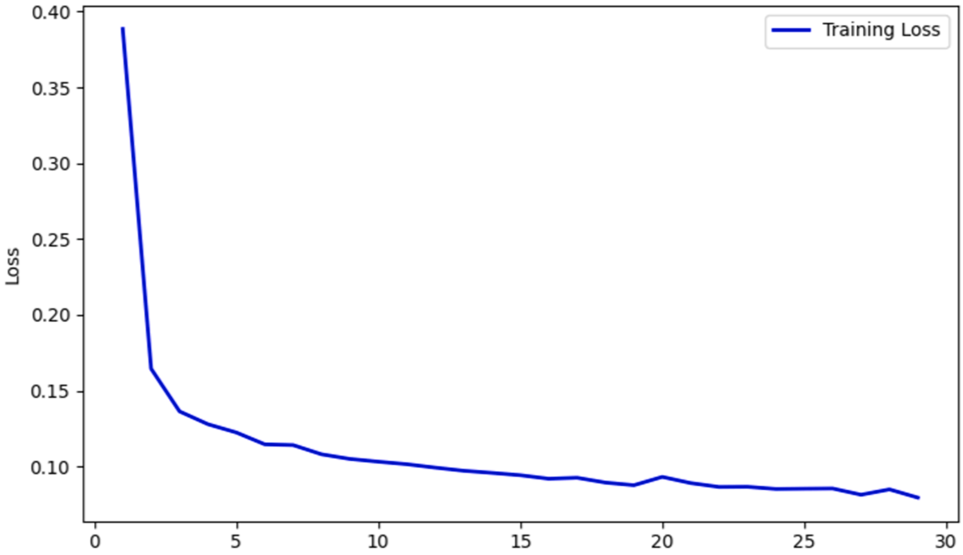

Fig. 6. Trend of the loss function during the model training phase with only context geographical data during the several epochs.

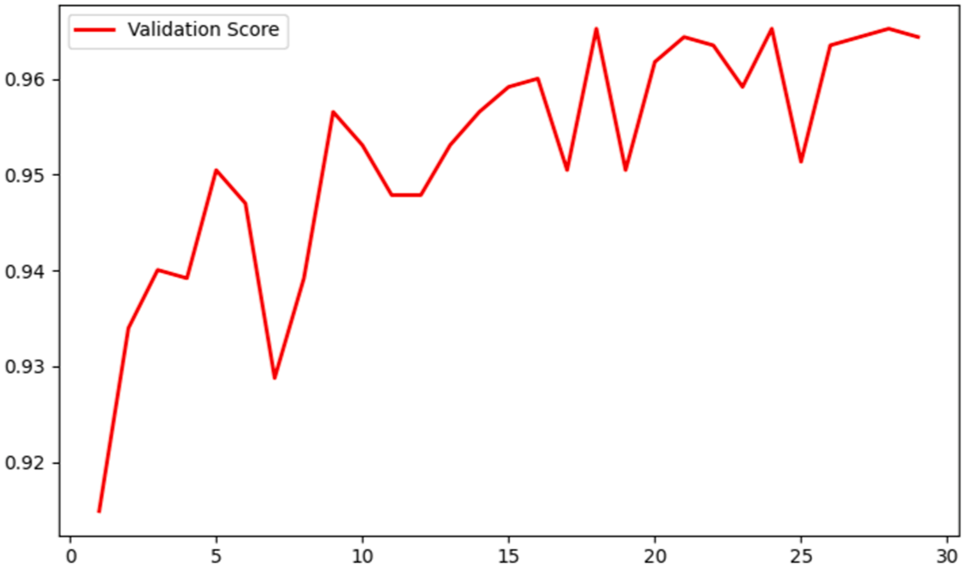

Fig. 7. Trend of the accuracy during the model training phase with only context geographical data during the several epochs.

being able to effectively extract relevant information from selected characteristics. At the same time, the validation accuracy curve (in red in Fig. 13) shows a progressive increase, starting from values around 0.86 and reaching a plateau above 0.90 in recent eras. The oscillations present in the validation curve, although more pronounced compared to the model with all features, remain small and do not indicate problems of instability in learning.

The Learning Curves in Fig. 14 provide a complementary perspective, showing the evolution of performance as a function of the number of training examples. It is particularly interesting to note that both curves reach values above 0.89, with the Training score remaining slightly above the Cross-validation score, indicating a good balance between bias and variance. This suggests that the model, even with the reduced set of features, maintains a good generalization capability.

The confusion matrix (see Fig. 15) provides a detailed analysis of the classifier's performance on the three risk classes. These results are particularly significant considering that they were obtained using only nine selected features. The model's ability to maintain performance above 90 % in all classes demonstrates the effectiveness of the selection of features. In particular, the "dangerous" class, critically important for road safety applications, maintains an accuracy above 91 %, confirming that the selected environmental and infrastructure characteristics are sufficient to effectively identify situations of greatest risk. The slight degradation in performance compared to the full model is largely compensated by the greater simplicity and practical applicability of this reduced approach.

The analysis of the SHAP chart for the model with a reduced set of features provides a perspective on the relevance and impact of each of the nine selected characteristics in the model's decision-making process (see Fig. 16).

The presence of schools ("poi school"), hospitals ("poi hospital"), slope ("slope"), and bridges ("bridge") maintains a significant influence, as it is shown in the previous section.

It is interesting to note that parking areas ("poi parking") show a moderate but well-defined impact, suggesting that these transition zones between vehicular and pedestrian mobility require special attention in risk assessment.

The weather conditions, represented by "wind_kph", show a more

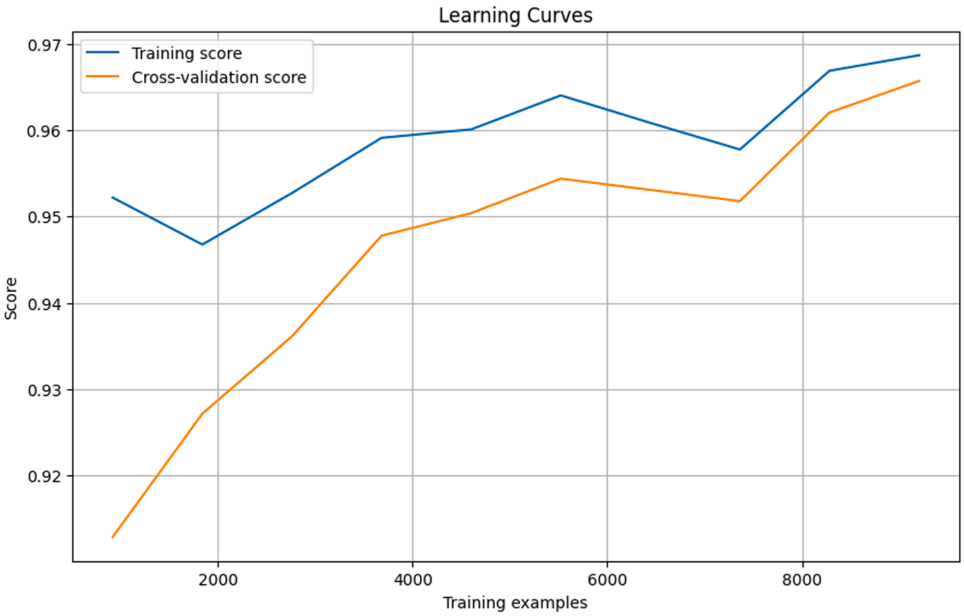

Fig. 8. Neural model learning curves showing performance evolution during the training phase. The graph shows the parallel trend of training score (blue line) and cross-validation score (orange line) as the number of examples used for training increases. The performance gap between training (blue line) and cross-validation (orange line) scores significantly narrows after 6000 training examples, with training scores oscillating around 0.96–0.97 and cross-validation scores achieving approximately 0.96.

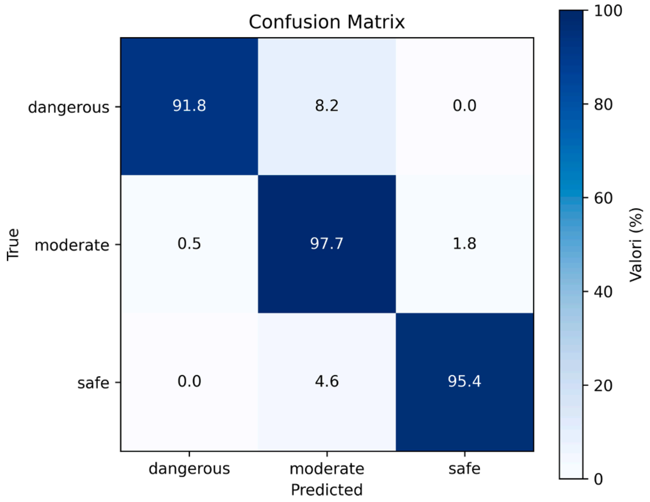

Fig. 9. Confusion matrix of the driving style classification model using a combination of OBD data and context data. The model accurately classified 91.8 % of cases as "dangerous", 97.7 % as "moderate", and 95.4 % as "safe". Classification errors remain reasonable, with 8.2 % of "dangerous" cases misclassified as "moderate", 1.8 % of "moderate" cases misclassified as "safe", and 4.6 % of "safe" cases misclassified as "moderate".

limited but still relevant influence, confirming the importance of considering environmental factors in road risk assessment.

The characteristics related to public transport infrastructure ("poi subway") and soft mobility ("poi cycleway-footway"), together with the "diameter" of roads, complete the picture showing more moderate but still significant impacts for a full risk assessment.

4. Discussion

The comparative analysis of the two SHAP graphs (see Fig. 10 and Fig. 11) offers interesting reflections on the relevance of the different characteristics in the two models and, above all, on the practicality and applicability of the model that does not use OBD data.

In the second model (Model 2), which does not use OBD data, environmental and infrastructure characteristics emerge as determining factors for risk classification. "Bridge”, "slope", and "poi school" are confirmed as the most influential features, with SHAP distributions showing a significant impact on both positive and negative predictions (see Fig. 11). This suggests that the model can capture effectively the complexity of the road context based solely on physical and environmental elements.

In the first model (Model 1), which includes OBD data, there is a redistribution of the importance of characteristics. While "poi school", "slope", and "bridge" hold prominent positions, variables such as "diff_steering_wheel_angle" and "mean_engine_speed" are introduced, which provide information on the behaviour of the vehicle (see Fig. 10). However, it is significant to note that these OBD characteristics do not dominate the ranking of importance but integrate into the model with a comparable or lower influence than the main environmental characteristics.

A particularly relevant aspect emerges from the analysis of the magnitude of the SHAP values. In the model without OBD (Model 2, see Fig. 11), the values are distributed in a range from about -0.1 to $+0.3$, while in the model with OBD, the range is slightly wider, from -0.2 to $+0.4$ (Model 1, see Fig. 10). This relatively small difference suggests that the addition of OBD data does not lead to a drastic improvement in the predictive capacity of the model.

The model with a reduced set of features (Model 3) shows similar performance to the full model (Model 2). It is particularly relevant to observe how the reduced model maintains a pattern of errors similar to the full model, with a tendency to commit "conservative" errors. In fact, in both cases, the most frequent errors occur between adjacent classes (dangerous-moderate and moderate-safe), while there are practically no errors that jump directly from the class "dangerous" to "safe" or vice versa.

These results are extremely encouraging because they show that it is possible to achieve very similar performance using a considerably reduced set of features. The fact that the reduced model can maintain accuracy above 90 % for all classes, with a maximum loss of performance of the order of 7 % in the worst case, suggests that the nine

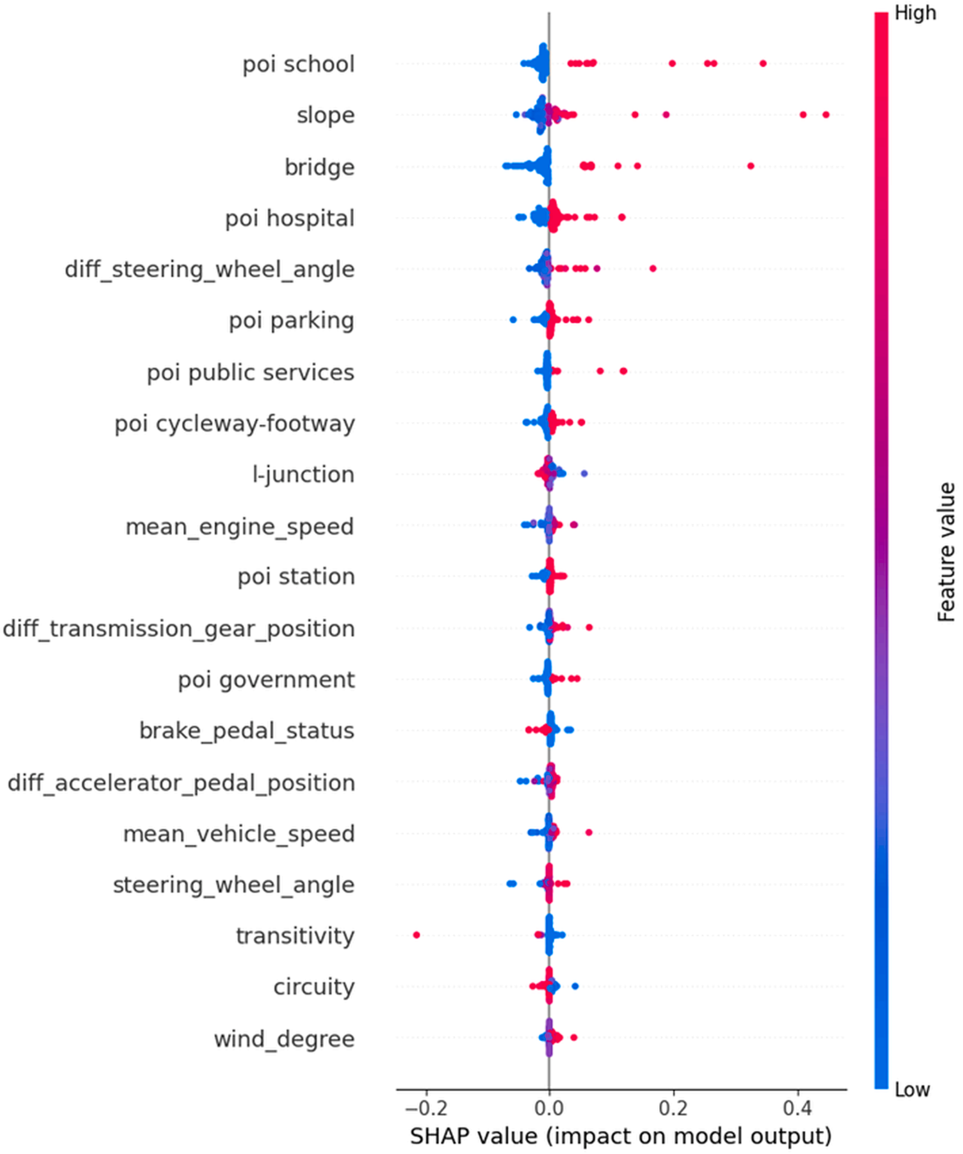

Fig. 10. Swarm plot for SHAP Analysis applied to the model based on OBD and geographical data (Model 1). This SHAP analysis provides valuable insights into understanding the model's decision-making process and could be used to identify priority areas for action in road risk management or to optimize future iterations of the model.

selected features effectively capture the essence of the road risk classification problem. This makes the small model a viable alternative to the full model, offering the advantage of greater simplicity in implementation and more immediate practical applicability, without significantly compromising the quality of the classification.

The comparative analysis of performance between the two models reveals a particularly significant aspect from the perspective of sustainability and computational efficiency. The model with a reduced feature set not only maintains performance similar to the full model but offers substantial advantages in terms of energy and computational efficiency.

In the model with a reduced set of features (Model 3, see Fig. 16), there is a redistribution of the relative importance of the selected characteristics. For example, "poi school" emerges as the most influential feature, showing a distribution of SHAP values that extends significantly to the right (up to 0.6), indicating a strong positive impact on risk classification. In the complete model (see Fig. 10), while "bridge" occupied the highest position, "poi school" was in third position, while maintaining a significant impact.

One particularly interesting aspect is the range of SHAP values. In the lighter model, distributions of main characteristics show slightly wider ranges than in the full model. This suggests that the reduced model has "compensated" for the lack of other features by giving more weight to the remaining characteristics, but without significantly altering the direction of their impact on classification.

The characteristics related to critical infrastructures ("poi hospital", "bridge", "poi parking") maintain a prominent position in both models, with similar distributions that suggest robustness in their predictive importance. This confirms that the selection of features has effectively preserved the most informative elements for road risk assessment.

The environmental conditions, represented by "wind_kph", show an interesting behaviour: in the reduced model, this characteristic maintains a well-defined influence, similar to that observed in the complete model. This suggests that the meteorological information captured by this single feature is sufficient to represent the impact of environmental conditions on risk.

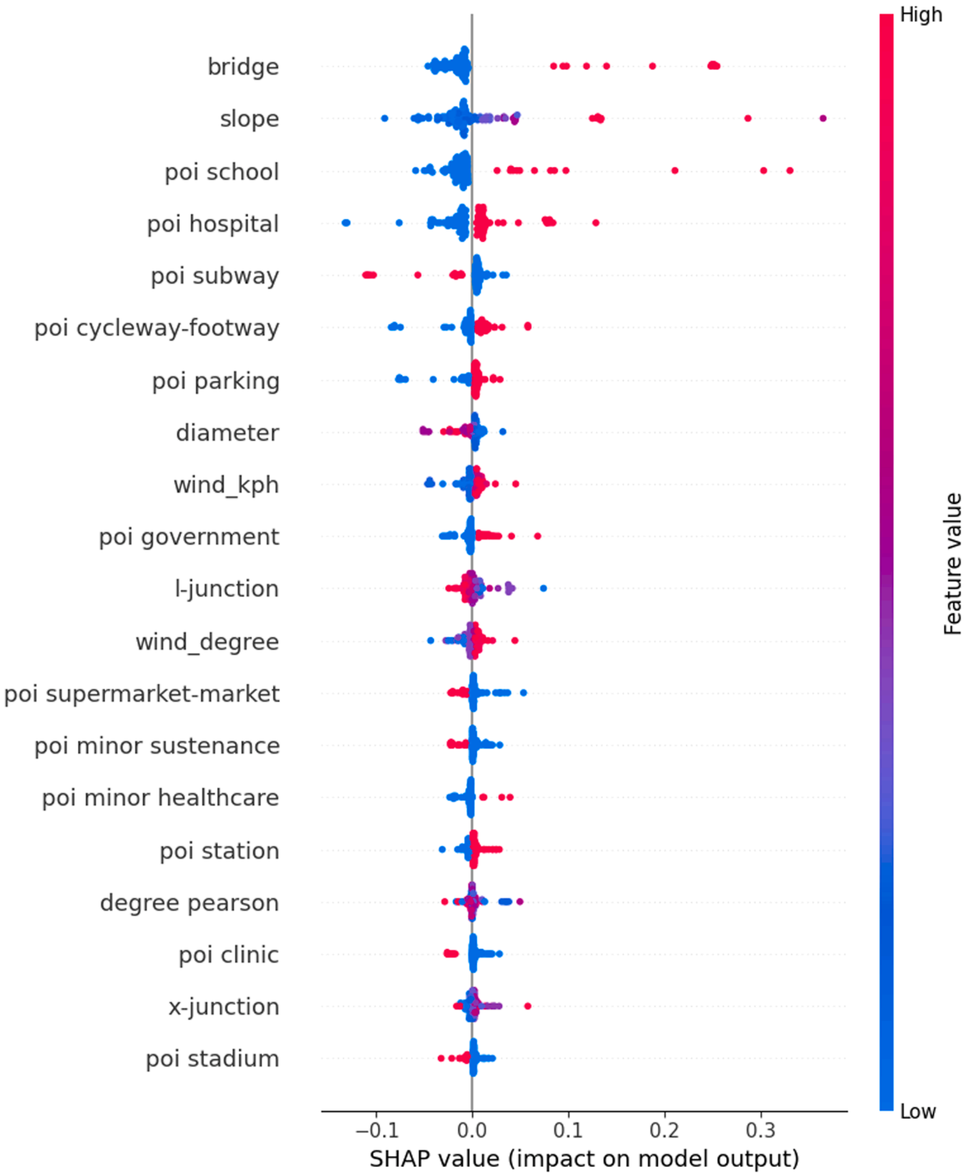

Fig. 11. Swarm plot for SHAP Analysis applied to the model based only on geographical data (Model 2).

5. Future works

The merging of OBD and geographic data enables the creation of a relationship between driving behavior and geographical features, as discussed in the previous section. The use of the DDD20 dataset, retrieved in a specific context and with the help of high-performance sensors, is one of the main limitations of this work.

Intending to study and correctly classify driver behaviours to improve sustainability and safety conditions in the road environment, the main challenge is to deploy affordable and reliable sensors to tackle this limitation. As it emerges from the literature [34,35], smartphones are widely implemented in mobility problems, and this is mainly due to their large availability and the high specifications with which they are built. Their integration with OBD could be a less expensive solution for creating a dataset, such as the DDD20 used in this paper. Another limitation that emerges from the current study is that the data collection was only for a single context. Although the contribution of the study regards both results (the role of geographic characteristics in predicting driver behaviours) and especially the methodological framework (due to the possibility of replicability), extending this type of analysis and methodology to other contexts could be doable and advisable. Thanks to the method's suitability for large area analysis, extending data collection in a European context or to European cities could be an additional and important contribution to this field. When the focus is on a single city or single area, usually other data sources can be implemented. From this perspective, data enrichment could be another important contribution. In addition to geographic information retrieved from OSM, the dataset could be enriched with traffic-related information, such as vehicular flows and speeds. These two dataset sources could give a framework on what happens in the road context, and if there is any correlation with aggressive driver behaviour.

By following this framework, an integration with road crash data should be an important future step. Usually, driver behaviours are classified by establishing a threshold for speeds and especially accelerations: so, if the driver's acceleration goes beyond an established value, this could be classified as aggressive, and aggressive behaviour leads to higher fuel consumption than other driving styles. The relationship between aggressive behaviour and road safety seems not to have been adequately demonstrated. This research gap that emerges from the literature could be deeply studied by comparing road crash frequency or severity with aggressive driver behaviour, and checking if there is a link or a recurrent pattern between aggressive driving behaviour and a lack

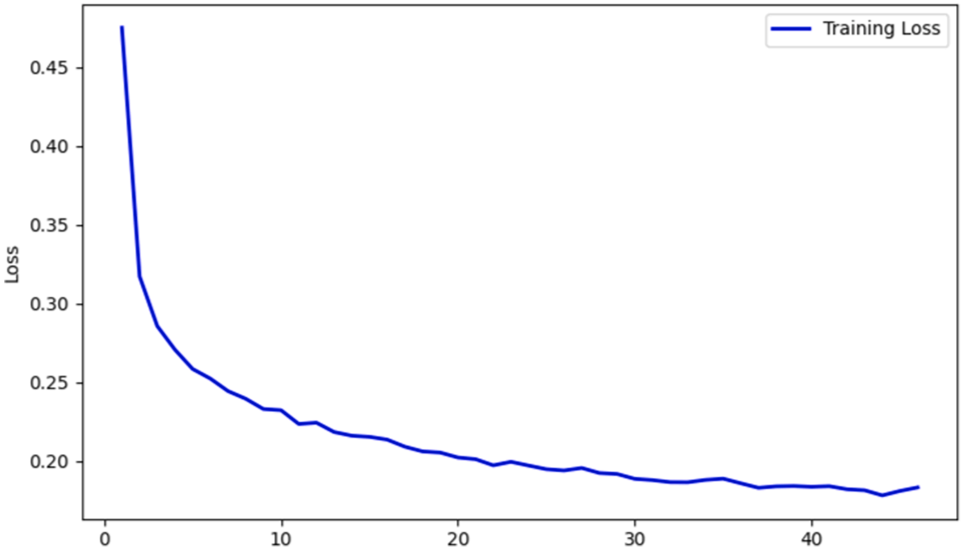

Fig. 12. Trend of the loss function during the model training phase with only context geographical data and with the reduced number of features. The Loss curve (in blue) shows a rapid initial decrease from about 0.45 to 0.25 in the first 5 epochs, followed by a more gradual reduction that brings the value to stabilize around 0.18–0.19 towards the end of the training.

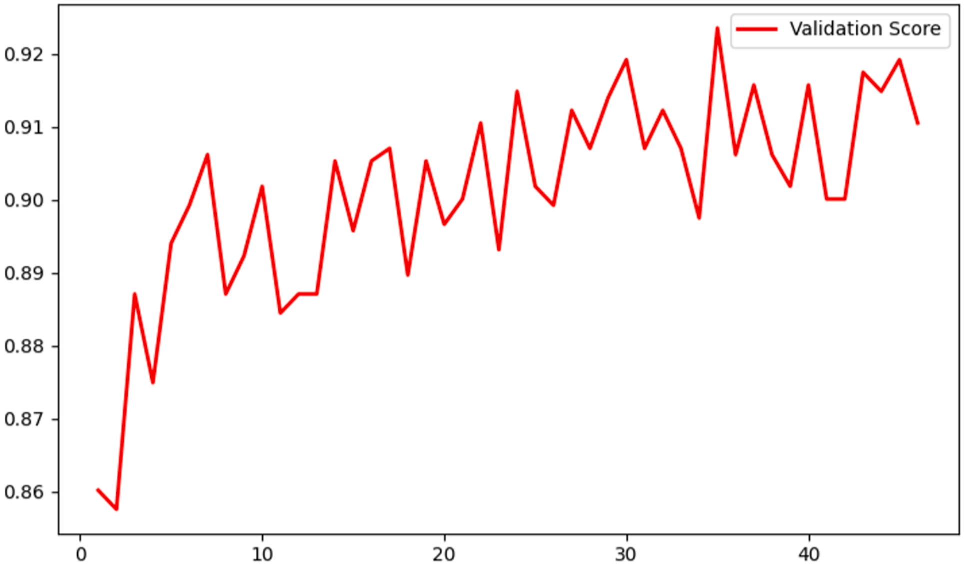

Fig. 13. Trend of the accuracy during the model training phase with only context geographical data and with the reduced number of features.

of road safety. This could be even crucial for studying interaction with vulnerable road users, especially pedestrians and cyclists [62,63]. As it emerges from recent studies on pedestrian and cyclist severity [64,65], driver behaviour seems strongly related to pedestrian severe crash severity [64], and vehicular flow and speeds increase the likelihood of having a severe outcome in crashes involving cyclists [65].

From a modelling perspective, given the high number of features that are used in the model and the possible future enrichment with other variables, integrating an attention mechanism will allow to have a better overview of which features the model will be focused on, trying to smooth the black box nature of these Machine Learning Models. Another possible future work, especially when new data in a different context is available (from a smartphone or from OBD), could be the introduction of other Recurrent Neural Networks (RNNs), such as Long Short-Term Memory (LSTM) [66,67]. The reason why these models could be adopted concerns the vanishing gradient problem, i.e., the impossibility for RNNs to integrate distant observations into a temporal sequence. Understanding that a driver is aggressive does not just require a single acceleration data point, but the entire time pattern of accelerations. As the last step, deploying the model in a real-time mobile application (it is worth noting that it is now released on the cloud due to the high amount of data) could be downloaded and used by all road users or integrated into other applications to improve eco-sustainability and safety.

6. Conclusion

The paper presented provides an in-depth analysis of the development and validation of an artificial intelligence model designed to classify driving styles based solely on the road context. The methodological approach followed, from the definition of a complete methodology to the implementation and validation of three main models, allowed to identify the most effective solution to forecast driver behaviours

The research has gone through several critical stages, starting from

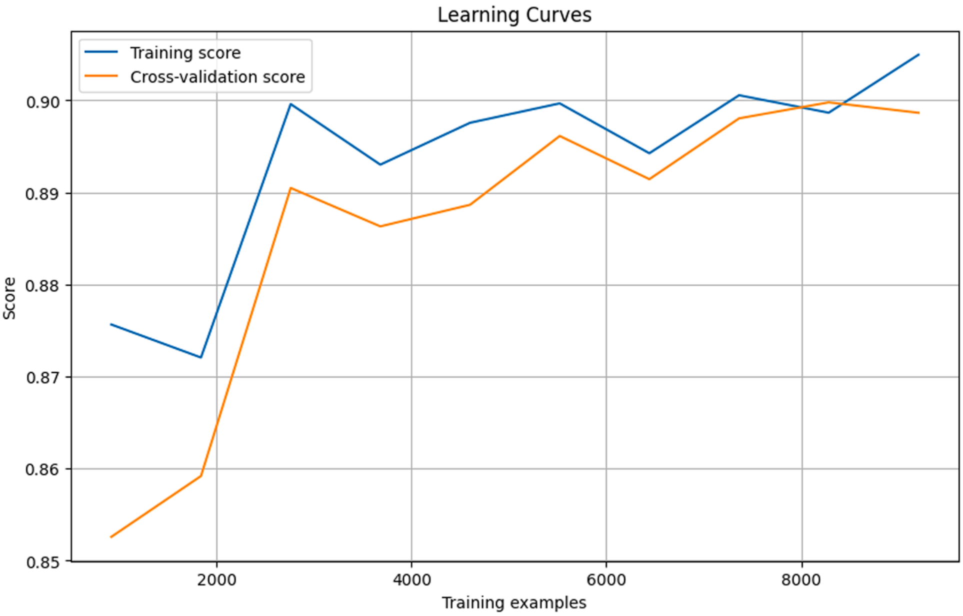

Fig. 14. Neural model learning curves showing performance evolution during the training phase. The graph shows the parallel trend of training score (blue line) and cross-validation score (orange line) as the number of examples used for training increases. The blue curve (Training score) and the orange one (Cross-validation score) show a converging trend, with the gap between the two narrowing significantly after 6000 training examples.

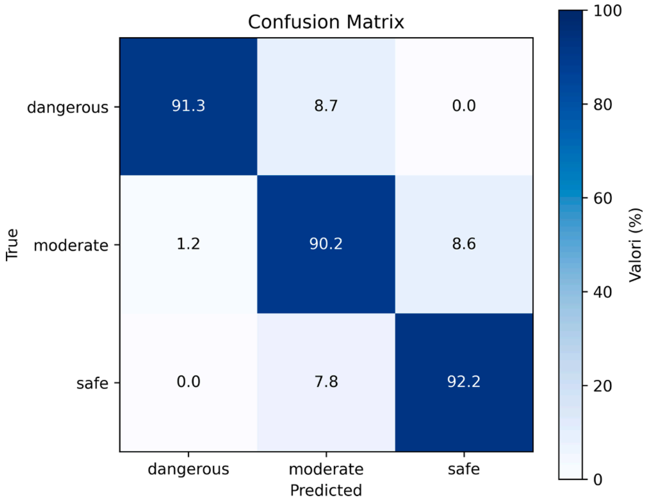

Fig. 15. Confusion matrix on the classification performance of the MLPClassifier model using only selected features with SHAP for training. The matrix shows that the model correctly classified 91.3 % of cases as "dangerous", 90.2 % as "moderate" and 92.2 % as "safe". The most significant errors are found in 8.7 % of "dangerous" cases classified as "moderate", 8.6 % of "moderate" cases classified as "safe", and 7.8 % of "safe" cases classified as "moderate".

data preparation using advanced pre-processing techniques, including temporal sorting and resampling of data, to accurate management of missing values. One particularly significant aspect was the inclusion of contextual factors in the labelling, which allowed elements such as weather conditions, the presence of points of interest, and road characteristics to be taken into account in the risk assessment.

The implementation of the MLP-NN model has proven its effectiveness through three main models: the first model (Model 1) use both OBD and contextual data, achieving an accuracy of more than 97 %; the second one (Model 2) showed that by using only contextual data, it is possible to obtain comparable performances with an accuracy of more than 90 %; the third model (Model 3) validated the effectiveness of a small subset of selected features through SHAP analysis, maintaining

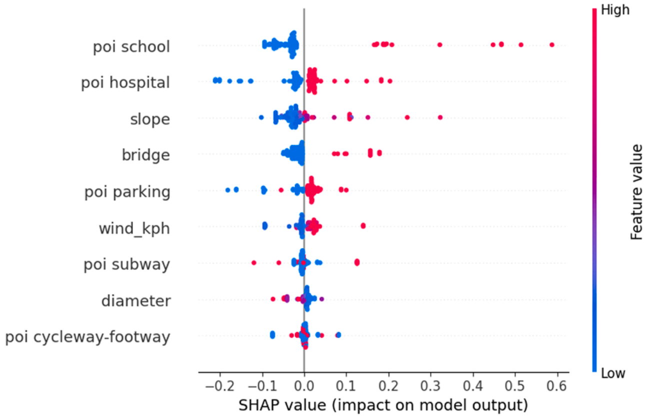

Fig. 16. Graph of the SHAP analysis applied to the model that uses fewer features for training and only geographical data. The SHAP analysis in this context is used to verify that the selected features continue to have the same order of importance on the new model.

high performance with only nine key characteristics.

The innovative application of SHAP analysis has played a fundamental role in the selection of the most relevant features, allowing for to development of a leaner, but also an equally effective model. The final selection of the nine most significant features (including the presence of schools, hospitals, road slope, and bridge characteristics) has shown that it is possible to achieve accurate classification of driving styles using a minimum set of environmental and infrastructural characteristics.

The real strength of the model without OBD data lies in its immediate applicability. This approach allows for the performance of road risk assessments without relying on the presence of instrumented vehicles, analyzes planned new areas or infrastructure changes, supports urban planning decisions and road safety interventions, and easily scales applications to different geographies without technological constraints

In particular, the model's ability to rely solely on environmental and infrastructure characteristics makes it a valuable tool for preventive road risk assessment. The comparison of the two models shows that the inclusion of OBD data, while providing additional information, is not necessarily crucial to an accurate road risk assessment. The model without OBD demonstrates that it can effectively capture the most relevant risk factors, while maintaining greater simplicity of implementation and wider practical applicability than the full model. This feature makes it particularly suitable for real-world applications, where the availability of OBD data may be limited or inconsistent.

One particularly relevant aspect of the research is the computational and energy efficiency of the reduced model. The reduction in the number of features resulted in: reduced memory and computing consumption, a significant reduction in energy consumption during inference, increased scalability and edge deployment capabilities, and reduced overall system carbon footprint

This paper contributes to the selection of the most appropriate road-context analysis model, providing a solid scientific basis for future design decisions. Comparative analyses of the results allow a full understanding of the capabilities and limitations of each tested approach. In particular, the demonstration that accurate classifications can be obtained using only contextual and geographic data, without relying on information from vehicles, opens new perspectives for the development of more accessible and effective road risk monitoring and prevention systems.

The results show that the optimized model with nine features (Model 3) represents the best compromise between predictive accuracy, computational efficiency, and applicability. This solution allows for a preventive assessment of road risk in different geographical areas without the need for specific vehicle instrumentation, making the system immediately usable for urban planning and road safety applications.

In conclusion, the work presented not only provides an advanced technical solution for driving style analysis but also establishes a robust methodological framework for future developments in road safety. The ability to achieve high performance using only geographic data, combined with the computational efficiency of the small model, makes this solution particularly suitable for large-scale implementations, contributing significantly to the objective of making roads safer through preventive and systematic risk analysis.

Funding sources

The research reported in this paper was developed in the EMOTIVEs Research Project, supported by the “FESR Lazio 2021–2027 Program - RSI Competitive Repositioning Notice”.

CRediT authorship contribution statement

V. Nicolosi: Writing – review & editing, Visualization, Supervision, Project administration, Methodology, Investigation, Funding acquisition, Conceptualization. **M. Mameli:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **S. Shiralizadeh:** Writing – review & editing, Resources, Methodology, Conceptualization. **I.G. Coltea:** Writing – review & editing, Supervision, Resources, Methodology. **M. D’Apuzzo:** Writing – review & editing, Methodology, Conceptualization. **G. Cappelli:** Writing – review & editing, Writing – original draft, Visualization, Methodology, Investigation, Conceptualization.

Declaration of competing interest

The authors declare that they have no competing and relevant interest(s) to disclose.

Supplementary materials

Supplementary material associated with this article can be found, in the online version, at doi:10.1016/j.treng.2025.100384.

Data availability

Data will be made available on request.

References

[1] A.B. Ellison, S.P. Greaves, M.C. Bliemer, Driver behaviour profiles for road safety analysis, Accid. Anal. Prev. 76 (2015) 118–132, https://doi.org/10.1016/j.aap.2015.01.009.
[2] I. Silva, J.Eugenio Naranjo, A systematic methodology to evaluate prediction models for driving style classification, Sensors 20 (2020) 1692, https://doi.org/10.3390/s20061692.
[3] D.I. Tselentis, E. Papadimitriou, Driver profile and driving pattern recognition for road safety assessment: main challenges and future directions, IEEE Open J. Intell. Transp. Syst. 4 (2023) 83–100, https://doi.org/10.1109/OJITS.2023.3237177.
[4] M. Shahverdy, M. Fathy, R. Berangi, M. Sabokrou, Driver behavior detection and classification using deep convolutional neural networks, Expert Syst. Appl. 149 (2020) 113240, https://doi.org/10.1016/j.eswa.2020.113240.
[5] B.G. Pratama, I. Ardiyanto, T.B. Adji, A review on driver drowsiness based on image, bio-signal, and driver behavior, in: 2017 3rd Int. Conf. Sci. Technol.-Comput. ICST, IEEE, 2017, pp. 70–75.
[6] H.-B. Kang, Various approaches for driver and driving behavior monitoring: a review, in: Proc. IEEE Int. Conf. Comput. Vis. Workshop, 2013, pp. 616–623.
[7] S. Kaplan, M.A. Guvensan, A.G. Yavuz, Y. Karalurt, Driver behavior Analysis for safe driving: a survey, IEEE Trans. Intell. Transp. Syst. 16 (2015) 3017–3032, https://doi.org/10.1016/j.eswa.2020.113240.
[8] N. Pugeault, R. Bowden, Learning pre-attentive driving behaviour from holistic visual features, in: Eur. Conf. Comput. Vis, Springer, 2010, pp. 154–167.
[9] B. Cyganek, S. Gruszczyński, Hybrid computer vision system for drivers’ eye recognition and fatigue monitoring, Neurocomputing 126 (2014) 78–94, https://doi.org/10.1016/j.neucom.2013.01.048.
[10] N. Peppes, T. Alexakis, E. Adamopoulou, K. Demestichas, Driving behaviour analysis using machine and deep learning methods for continuous streams of vehicular data, Sensors 21 (2021) 4704, https://doi.org/10.3390/s21144704.
[11] E. Papadimitriou, C. Schneider, J.A. Tello, W. Damen, M.L. Vrouenraets, A. Ten Broeke, Transport safety and human factors in the era of automation: what can transport modes learn from each other? Accid. Anal. Prev. 144 (2020) 105656 https://doi.org/10.1016/j.aap.2020.105656.
[12] E. Petridou, M. Moustaki, Human factors in the causation of road traffic crashes, Eur. J. Epidemiol. 16 (2000) 819–826, https://doi.org/10.1023/A:1007649804201.
[13] What Is a Safe System Approach? | US Department of Transportation, (n.d.). https://www.transportation.gov/safe-system-approach (accessed August 18, 2025).
[14] M.N. Khan, S. Das, Advancing traffic safety through the safe system approach: a systematic review, Accid. Anal. Prev. 199 (2024) 107518, https://doi.org/10.1016/j.aap.2024.107518.
[15] N. Xu, X. Li, Q. Liu, D. Zhao, An overview of eco-driving theory, capability evaluation, and training applications, Sensors 21 (2021) 6547, https://doi.org/10.3390/s21196547.
[16] M.-H. Yen, S.-L. Tian, Y.-T. Lin, C.-W. Yang, C.-C. Chen, Combining a universal OBD-II module with deep learning to develop an eco-driving analysis system, Appl. Sci. 11 (2021) 4481, https://doi.org/10.3390/app11104481.
[17] D. Somayajula, A. Meintz, M. Ferdowsi, Designing efficient hybrid electric vehicles, IEEE Veh. Technol. Mag. 4 (2009) 65–72, https://doi.org/10.1109/MVT.2009.932546.
[18] M. Loman, B. Šarkan, T. Skrúcaný, Comparison of fuel consumption of a passenger car depending on the driving style of the driver, Transp. Res. Procedia 55 (2021) 458–465, https://doi.org/10.1016/j.trpro.2021.07.009.
[19] M. Graba, J. Mamala, A. Bieniek, Z. Sroka, Impact of the acceleration intensity of a passenger car in a road test on energy consumption, Energy 226 (2021) 120429, https://doi.org/10.1016/j.energy.2021.120429.
[20] W.A. Al-Hussein, L.Y. Por, M.L.M. Kiah, B.B. Zaidan, Driver behavior profiling and recognition using deep-learning methods: in accordance with traffic regulations and experts guidelines, Int. J. Environ. Res. Public. Health 19 (2022) 1470, https://doi.org/10.3390/ijerph19031470.
[21] H. Singh, A. Kathuria, Analyzing driver behavior under naturalistic driving conditions: a review, Accid. Anal. Prev. 150 (2021) 105908, https://doi.org/10.1016/j.aap.2020.105908.
[22] R.A. Blanchard, A.M. Myers, M.M. Porter, Correspondence between self-reported and objective measures of driving exposure and patterns in older drivers, Accid. Anal. Prev. 42 (2010) 523–529, https://doi.org/10.1016/j.aap.2009.09.018.
[23] S. Agramunt, L.B. Meuleners, M.L. Fraser, K.C. Chow, J.Q. Ng, V. Raja, First and second eye cataract surgery and driver self-regulation among older drivers with bilateral cataract: a prospective cohort study, BMC Geriatr 18 (2018) 51, https://doi.org/10.1186/s12877-018-0743-1.
[24] S. Bouhsissin, N. Sael, F. Benabbou, Driver behavior classification: a systematic literature review, IEEE Access 11 (2023) 14128–14153, https://doi.org/10.1109/ACCESS.2023.3243865.
[25] A. Ziakopoulos, D. Tselentis, A. Kontaxi, G. Yannis, A critical overview of driver recording tools, J. Safety Res. 72 (2020) 203–212, https://doi.org/10.1016/j.jsr.2019.12.021.
[26] I. Zöller, B. Abendroth, R. Bruder, Driver behaviour validity in driving simulators–Analysis of the moment of initiation of braking at urban intersections,

Transp. Res. Part F Traffic Psychol. Behav. 61 (2019) 120–130, https://doi.org/10.1016/j.trf.2017.09.008.

[27] J.P. Ehsani, J.L. Harbluk, J. Bärgman, A. Williamson, J.P. Michael, R. Grzebieta, J. Olivier, J. Eusebio, J. Charlton, S. Koppel, Naturalistic driving studies: an overview and international perspective, Int. Encycl. Transp. (2021) 20–38, https://doi.org/10.1016/j.jsr.2019.12.021.

[28] W.A. Al-Hussein, M.L.M. Kiah, L. Yee, B.B. Zaidan, A systematic review on sensor-based driver behaviour studies: coherent taxonomy, motivations, challenges, recommendations, substantial analysis and future directions, PeerJ Comput. Sci. 7 (2021), https://doi.org/10.7717/peerj-cs.632 e632.

[29] T. Pholprasit, W. Choochaiwattana, C. Saiprasert, A comparison of driving behaviour prediction algorithm using multi-sensory data on a smartphone, in: 2015 IEEEACIS 16th Int. Conf. Softw. Eng. Artif. Intell. Netw. ParallelDistributed Comput. SNPD, IEEE, 2015, pp. 1–6.

[30] T.K. Chan, C.S. Chin, H. Chen, X. Zhong, A comprehensive review of driver behavior analysis utilizing smartphones, IEEE Trans. Intell. Transp. Syst. 21 (2019) 4444–4475, https://doi.org/10.1109/TITS.2019.2940481.

[31] S.K. Alluhaibi, M.S.N. Al-Din, A. Moyaid, Driver behavior detection techniques: a survey, Int J Appl Eng Res 13 (2018) 8856–8861.

[32] N. Jain, S. Mittal, Review of computational techniques for modelling eco-safe driving behavior, Int. J. Automot. Mech. Eng. 20 (2023) 10422–10440, https://doi.org/10.15282/ijame.20.2.2023.08.0806.

[33] J. Rios-Torres, J. Liu, A. Khattak, Fuel consumption for various driving styles in conventional and hybrid electric vehicles: integrating driving cycle predictions with fuel consumption optimization, Int. J. Sustain. Transp. 13 (2019) 123–137, https://doi.org/10.1080/15568318.2018.1445321.

[34] M. Okmi, L.Y. Por, T.F. Ang, W. Al-Hussein, C.S. Ku, A systematic review of mobile phone data in Crime applications: a coherent taxonomy based on data types and analysis perspectives, challenges, and future research directions, Sensors 23 (2023), https://doi.org/10.3390/s23094350.

[35] S. Taherinavid, S.V. Moravvej, Y.-L. Chen, J. Yang, C.S. Ku, L. Yee, Automatic transportation mode classification using a deep reinforcement learning approach with smartphone sensors, IEEE Access 12 (2023) 514–533, https://doi.org/10.1109/ACCESS.2023.3346875.

[36] C.M. Martinez, M. Heucke, F.-Y. Wang, B. Gao, D. Cao, Driving style recognition for intelligent vehicle control and advanced driver assistance: a survey, IEEE Trans. Intell. Transp. Syst. 19 (2017) 666–676, https://doi.org/10.1109/TITS.2017.2706978.

[37] Z.E. Abou Elassad, H. Mousannif, H. Al Moatassime, A. Karkouch, The application of machine learning techniques for driving behavior analysis: a conceptual framework and a systematic literature review, Eng. Appl. Artif. Intell. 87 (2020) 103312, https://doi.org/10.1016/j.engappai.2019.103312.

[38] E. Lattanzi, G. Castellucci, V. Freschi, Improving machine learning identification of unsafe driver behavior by means of sensor fusion, Appl. Sci. 10 (2020) 6417, https://doi.org/10.3390/app10186417.

[39] N. Fadzil, H.M. Fadzir, H. Mansor, U. Rahardja, Driver behaviour classification: a research using obd-ii data and machine learning, J. Adv. Res. Appl. Sci. Eng. Technol. (2024) 51–61, https://doi.org/10.37934/araset.56.3.147157.

[40] U. Fugiglando, E. Massaro, P. Santi, S. Milardo, K. Abida, R. Stahlmann, F. Netter, C. Ratti, Driving behavior analysis through CAN bus data in an uncontrolled environment, IEEE Trans. Intell. Transp. Syst. 20 (2018) 737–748, https://doi.org/10.1109/TITS.2018.2836308.

[41] H.A. Ameen, A.K. Mahamad, S. Saon, R.Q. Malik, Z.H. Kareem, M.A. Bin Ahmadon, S. Yamaguchi, Identification of driving safety profiles in vehicle to vehicle communication system based on vehicle OBD information, Information 12 (2021) 194, https://doi.org/10.3390/info12050194.

[42] K. Uvarov, A. Ponomarev, Driver identification with OBD-II public data, in: 2021 28th Conf. Open Innov. Assoc. FRUCT, IEEE, 2021, pp. 495–501.

[43] R. Kumar, A. Jain, Driving behavior analysis and classification by vehicle OBD data using machine learning, J. Supercomput. 79 (2023) 18800–18819, https://doi.org/10.1007/s11227-023-05364-3.

[44] E. Lattanzi, V. Freschi, Machine learning techniques to identify unsafe driving behavior by means of In-vehicle sensor data, Expert Syst. Appl 176 (2021), https://doi.org/10.1016/j.eswa.2021.114818.

[45] T. Abukhalil, H. AlMahafzah, M. Alksasbeh, B.A. Alqaralleh, Fuel consumption using OBD-II and support vector machine model, J. Robot. 2020 (2020) 9450178, https://doi.org/10.1155/2020/9450178.

[46] N. Rivera Campoverde, P. Molina Campoverde, J.Molina Campoverde, Estimation of areas with the highest accident rate on the Cuenca-Loja road according to the driving maneuvers, in: 2023: pp. 335–345. https://doi.org/10.1007/978-3-031-24327-1_29.

[47] A.E. Abdelrahman, H.S. Hassanein, N. Abu-Ali, Robust data-driven framework for driver behavior profiling using supervised machine learning, IEEE Trans. Intell. Transp. Syst. 23 (2020) 3336–3350, https://doi.org/10.1109/TITS.2020.3035700.

[48] V. Nicolosi, M. Mameli, S. Shiralizadeh, I.G. Coltea, M. D'Apuzzo, G. Cappelli, Interactions between drivers and road infrastructure characteristics: combining OBD and geographic data to classify driver behaviors with feed-forward neural network, in: Int. Conf. Comput. Sci. Its Appl, Springer, 2025, pp. 51–64.

[49] OpenStreetMap, OpenStreetMap (n.d.). https://www.openstreetmap.org/ (accessed August 18, 2025).

[50] Rain Viewer: Leading Weather Radar App for Accurate Rain & Snow Forecasts| Rain Viewer, (n.d.). https://www.rainviewer.com/ (accessed August 18, 2025).

[51] Y. Hu, J. Binas, D. Neil, S.-C. Liu, T. Delbrück, DAVIS Driving Dataset 2020 (DDD20), (2024). https://doi.org/10.3929/ethz-b-000707895.

[52] A. Coates, A.Y. Ng, Learning feature representations with k-means, Neural Netw. Tricks Trade Second Ed. (2012) 561–580.

[53] S. Okada, M. Ohzeki, S. Taguchi, Efficient partition of integer optimization problems with one-hot encoding, Sci. Rep. 9 (2019) 13036, https://doi.org/10.1038/s41598-019-49539-6.

[54] B. Komer, J. Bergstra, C. Eliasmith, Hyperopt-Sklearn: Automatic Hyperparameter Configuration for Scikit-Learn, Scipy, 2014, pp. 32–37.

[55] A. Ruangkanjanases, O. Sivarak, Z.J. Weng, A. Khan, S.C. Chen, Using multilayer perceptron neural network to assess the critical factors of traffic accidents, HighTech Innov. J. 5 (2024) 157–169, https://doi.org/10.28991/HIJ-2024-05-01-012.

[56] S. Mokhtarimousavi, J.C. Anderson, A. Azizinamini, M. Hadi, Factors affecting injury severity in vehicle-pedestrian crashes: a day-of-week analysis using random parameter ordered response models and Artificial Neural Networks, Int. J. Transp. Sci. Technol. 9 (2020) 100–115, https://doi.org/10.1016/j.ijtst.2020.01.001.

[57] M.E. Shaik, M.M. Islam, Q.S. Hossain, A review on neural network techniques for the prediction of road traffic accident severity, Asian Transp. Stud. 7 (2021) 100040, https://doi.org/10.1016/j.eastsj.2021.100040.

[58] O.I. Abiodun, A. Jantan, A.E. Omolara, K.V. Dada, A.M. Umar, O.U. Linus, H. Arshad, A.A. Kazaure, U. Gana, M.U. Kiru, Comprehensive review of artificial neural network applications to pattern recognition, IEEE Access 7 (2019) 158820–158846, https://doi.org/10.1109/ACCESS.2019.2945545.

[59] H. Sahlaoui, A. Nayyar, S. Agoujil, M.M. Jaber, Predicting and interpreting student performance using ensemble models and shapley additive explanations, IEEe Access 9 (2021) 152688–152703, https://doi.org/10.1109/ACCESS.2021.3124270.

[60] Y. Nohara, K. Matsumoto, H. Soejima, N. Nakashima, Explanation of machine learning models using shapley additive explanation and application for real data in hospital, Comput. Methods Programs Biomed. 214 (2022) 106584, https://doi.org/10.1016/j.cmpb.2021.106584.

[61] I.U. Ekanayake, D.P.P. Meddage, U. Rathnayake, A novel approach to explain the black-box nature of machine learning in compressive strength predictions of concrete using Shapley additive explanations (SHAP), Case Stud, Constr. Mater. 16 (2022) e01059, https://doi.org/10.1016/j.cscm.2022.e01059.

[62] G. Cappelli, M. D'Apuzzo, S. Nardoianni, V. Nicolosi, Exploring the influences of safety and energy expenditure parameters on cycling, Sustainability 16 (2024) 2739, https://doi.org/10.3390/su16072739.

[63] D. Santilli, M. D'apuzzo, A. Evangelisti, V. Nicolosi, Towards sustainability: new tools for planning urban pedestrian mobility, Sustainability 13 (2021) 9371, https://doi.org/10.3390/su13169371.

[64] G. Cappelli, S. Nardoianni, M. D'Apuzzo, V. Nicolosi, Pedestrian crash severity prediction and contributory factors analysis by using machine learning methods, in: Int. Conf. Comput. Sci. Its Appl, Springer, 2025, pp. 3–14.

[65] G. Cappelli, S. Nardoianni, M. D'Apuzzo, V. Nicolosi, Interpretable crash severity prediction models to improve cyclist safety, in: Int. Conf. Comput. Sci. Its Appl, Springer, 2025, pp. 319–334.

[66] C.S. Ku, J. Xiong, Y.-L. Chen, S.D. Cheah, H.C. Soong, L.Y. Por, Improving stock market predictions: an equity forecasting scanner using long short-term memory method with dynamic indicators for Malaysia stock market, Mathematics 11 (2023) 2470.

[67] Z. He, J. Yang, R. Alroobaea, L.Y. Por, SeizureLSTM: an optimal attention-based trans-LSTM network for epileptic seizure detection using optimal weighted feature integration, Biomed. Signal Process. Control 96 (2024) 106603.