Human Interaction and Emerging Technologies (IHIET-AI 2024), Vol. 120, 2024, 142–152
https://doi.org/10.54941/ahfe1004564

USID - Unsupervised Identification of the Driver for Vehicle Comfort Functions

Veljko Vučinić[1,2], Luca Seidel[2], Marco Stang[2], and Eric Sax[2]

[1]RA Consulting GmbH, Bruchsal, Germany
[2]Institute for Information Processing Technologies, Karlsruhe Institute of Technology (KIT), Karlsruhe, Germany

ABSTRACT

High comfort is one of the main demand for a modern vehicle. Comfort functions that are designed to ensure the comfort of modern vehicles are becoming more tailored to each driver. In order to maximize the effectiveness of comfort functions the driver must be precisely identified. The driver identification task can be accomplished by utilizing vehicle data from the standardized On-Board Diagnostic II system (OBD II). In this paper, the feasibility of precise driver identification was investigated based on unsupervised machine learning methods. The authors propose the USID (Unsupervised Identification of the Driver) concept for this purpose. The USID promises rich scalability since the unsupervised models don't use predefined classes to identify drivers. The unsupervised methods used in this work are K-Means, Autoencoders, Self-Organizing maps, and Density-based spatial clustering (DBSCAN). The models are trained and evaluated using the OBD II time series of 16 drivers driving the same vehicle. In the end, the experimental analysis of the USID was done that showed very good confidence of the concept in driver identification during the driving cycle of all 16 drivers.

Keywords: Driver identification, Machine learning, Vehicle comfort, OBD II

INTRODUCTION

Modern vehicles must meet the rapidly increasing customer requirements for in-vehicle comfort. An increase in comfort and safety is a main objective of car manufacturers as well as of their suppliers and a substantial distinguishing feature in the market (Bloecher, Dickmann, and Andres, 2015). Vehicle comfort functions are designed to ensure comfort with a more pleasant and convenient driving experience, including thermal comfort, seating comfort, noise reduction, and entertainment systems. Interaction with these systems plays an important role in ensuring comfort and elevating the overall driver experience, therefore they are becoming more tailored for each driver. In general, there is a considerable level of consumer acceptance for automated comfort functions for the highest levels of automation (4 and 5) (Guinea *et al.*, 2021). Thermal comfort represents one part of the comfort in the vehicle which is perceived as subjective and relative. In contrast to its subjectiveness, it has been successfully developed driver-specific (Lahlou, 2020, 2020; Schaut, 2019; Xie, 2020). Furthermore, the automatic drive seat positioning

and rear mirror setting in the ride comfort profoundly influence the driver's fatigue (Sharma *et al.*, 2021). Many in-vehicle functions can be controlled by the driver while the vehicle is in motion. With that in mind, a driver-specific user interface of an infotainment system can be applied on a vehicle to ensure seamless driver interaction (Kern, 2009; Stang, 2022). All of these comfort functions assume that the driver is identified successfully. Therefore, in order for a vehicle to uniquely recalibrate each comfort function, it is imperative that the driver is precisely identified.

Existing driver identification systems are functional, yet they exhibit limitations in precision or practicality. For example, a vehicle could detect the presence of a driver's smartphone but fail to distinguish whether the individual is the actual driver or a passenger. Similar approaches are proposed with the biometrics methods for driver detection, such as fingerprint, eye scan, face recognition, etc. Their high precision comes with the price of data protection endangerment, especially with oncoming connected vehicles, leaning them towards impracticality. The unique signature of each driver's style is encoded within the time series data collected from various vehicle sensors. Vehicle data contains valuable information allowing recognition and understanding of the unique driving patterns and behaviors of individuals. Therefore, there is no need for integrating novel driver identification sensors, as the identification task can be accomplished by utilizing vehicle data from the On-Board Diagnostic II system (OBD II). The numerous OBD II parameters available through the CAN interface lay a strong foundation for in-vehicle Machine Learning (ML) processing, besides their main OBD II purpose.

RELATED WORK

The ML techniques represent a viable choice for creating a driver identification system with data-based models when a valuable amount of quality data is available. The high quality, variety, and availability of the aforementioned OBD II data make it a valid choice to use these data-driven modelling algorithms as a foundation for this type of in-vehicle system. Supervised machine learning models are trained on labeled data, in this case, driver IDs, allowing them to detect specific drivers based on historical information fed to it during training phases. On the other hand, unsupervised machine learning deals with uncovering patterns within the OBD data itself, enabling the recognition of distinct driving styles without predefined driver IDs. Both approaches offer unique advantages in enhancing the accuracy and practicality of driver identification systems, contributing to increased comfort in the driving experience.

Considering the research done in the field of ML-based driver identification, supervised learning models seem more favorable. Table 1. lists the research covering driver identification with various purposes and ML solutions. All related work listed utilizes the real recorded data which proves the relevance of each work.

The related work proved that it is possible to precisely identify drivers using different supervised algorithms with high accuracy levels of above 90% in all mentioned research. Providing its advantages with high precision

achieved with labeled training data, the disadvantages lead towards low scalability and robustness of models. Unsupervised machine learning methods are barely used for driver identification, missing out on the benefits they can bring to enhance comfort functions.

Table 1. Summarized ML-based driver identification research.

Study	Purpose	No. Drivers	ML Type	ML Algorithms
Van *et al.* (2013)	vehicle safety	2	supervised, unsupervised	k-means, SVM
Khan *et al.* (2022)	driver behaviour	10	supervised	k-NN, SVM, logistic regression, NB, REP tree
Uvarov, Ponomarev (2018)	data processing	10	supervised	k-NN, decision tree, random forest, GB, SVM
Girma *et al.* (2019)	vehicle security	10	supervised	LSTM
Xun *et al.* (2019)	vehicle security	20	supervised	CNN, SVDD
Kwak *et al.* (2016)	vehicle security	10	supervised	SVM, random forest, NB, k-NN
Enev *et al.* (2016)	data privacy	15	supervised	SVM, random forest, NB, k-NN

USID ARCHITECTURE

In contrast to the related work, this research paper investigates the viability of precise driver identification based on unsupervised machine learning methods on a larger scale. The requirements for modern comfort functions are a precise, seamless, and scalable driver identification system. Unsupervised learning has the potential to cope with the mentioned requirements in the context of driver identification available OBD II data. For this purpose, the authors introduce the USID concept - Unsupervised Identification of the Driver. Due to its unsupervised characteristic, USID offers significant scalability, as its models don't rely on predefined classes to distinguish and identify drivers. This flexibility of USID is valuable when dealing with diverse driver styles that may not fit predefined categories.

The USID concept is inspired by the known Knowledge Discovery in Databases (KDD) process (Flawley et al., 1992). In general, a vehicle can be controlled with different drivers during its exploitation. Regardless of the driver, the OBD II system internally monitors and diagnoses different pollution and propulsion-related systems. This represents the input for the USID concept. Correlating to the KDD, the USID integrates different steps of data processing, such as data selection, data preprocessing, data transformation, and the model at the end (see Figure 1). In the first step, the scale of the data is reduced to the most relevant parameters, before being preprocessed. After preprocessing, the transformation of the data takes place according to the

unsupervised ML model later used. Finally, the transformed data goes into the model that gives the knowledge about the driver identification. Each step of data processing in USID contains the driver information, which is exploited using the data processing methodology previously noted for clear reasoning of driver identification.

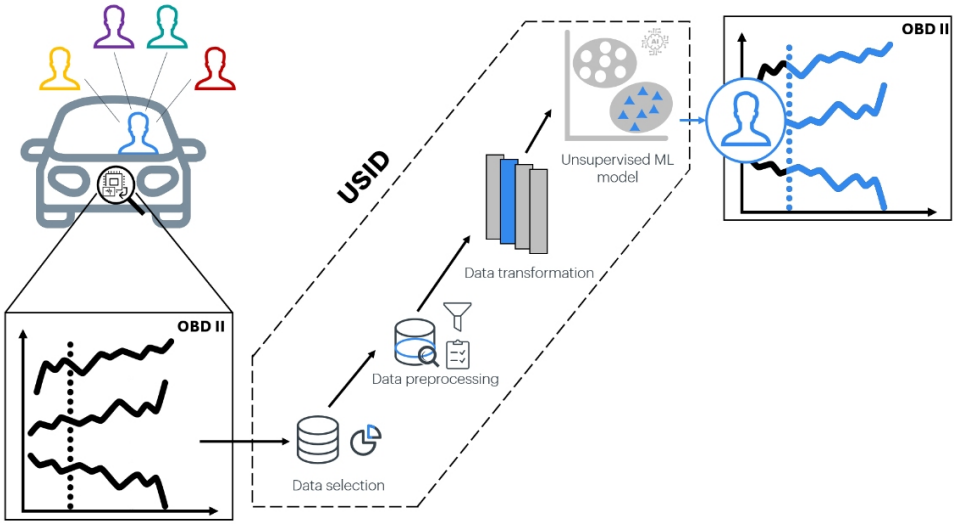

Figure 1: USID concept for unsupervised ML driver identification.

PROPOSED METHODOLOGY

The data used to prove the USID concept came from the open-source database published in Barreto (2018). The original data contains a total of 9,261 rows from 28 parameters recorded from the OBD II port, making it a viable source for the machine learning models. The vehicle Chevrolet S10 2.5l was used with 16 different drivers on one route with a length of 18.8km. The data has been already uniformly sampled. Additionally, the mentioned dataset was extended with additional features like longitudinal acceleration and jerk. Longitudinal acceleration and jerk of a vehicle are generated as a first and second time derivative of the vehicle speed signal. This was already proven to give good results in the related research (Enev *et al.*, 2016), therefore it is expected to give similar results with unsupervised driver identification models.

In the initial phase of the proposed process pipeline, the data quality checks were performed with the purpose of identifying and addressing outliers to avoid any potential distortions in model inferences. In the first iteration, basic preprocessing took place, unnecessary column information and rows with empty cells were removed, together with low-variance data. Following, the rough filtering was done with statistical outliers that were detected and removed for the majority of features using quantiles.

Next to the preprocessing, data normalization was done. In order to ensure the relevance of our time sequence comparisons across a diverse range of

driving contexts, drivers, and vehicles, we utilize global normalization techniques incorporating physical signal boundaries. With that in mind, the global physical limitation of the OBD II parameters, defined with the *SAE Standard J1979-DA* (2017), are combined with the min-max normalization method.

Feature selection is an important step in data processing using machine learning methods. In this case, it involves identifying and selecting the most relevant and informative OBD II parameters from the dataset, while discarding the features that could harm the performance of future USID unsupervised machine learning models. By selecting the most applicable features, redundant or irrelevant data are removed, reducing the dimensionality and computational heaviness of the USID. In this work, a filter-based feature selection method, the SelectKBest, was used. The goal was to select the 10 best features that can represent the 16 drivers during their driving cycles. The result of feature selection narrows down to engine coolant temperature, fuel level, ambient air temperature, latitude, longitude, altitude, engine rpm, air intake temperature, vehicle speed, and timing advance. To validate the meaningfulness of the feature selection, the feature correlation was checked using Spearman's rank correlation statistical method.

Upon choosing the most relevant features from the dataset, the data transformation involves altering the selected features to ensure they improve their interpretability later with the ML model. In this specific case, the Uniform Manifold Approximation and Projection (UMAP) method was used to reduce the dimensionality of the data, in order to streamline the dataset further, reducing noise and computational load while retaining the driver identification information. The UMAP is a nonlinear dimensionality reduction technique used for embedding high-dimensional data into a lower-dimensional space while preserving the underlying structure and relationships within the data points (McInnes, Healy and Melville, 2018).

Sliding windows represent a fundamental technique in time series analysis, enabling the transformation of continuous time series data into discrete manageable subsets suitable for both supervised and unsupervised learning approaches. A window is defined by its length w which moves over the dataset with a stride s. The temporal data stream is effectively segmented into coherent chunks. Local patterns are intricately encoded within the sequence of data, reflecting not just individual data points but the temporal context they inhabit. In this research, a window length of $w = 30$ with a stride of $s = 1$ is used. For the mentioned dataset this results in 6636 windows. Windows with undefined driver affiliations were neglected due to a lack of back traceability.

Besides the clustering itself, an autoencoder, presented in section Model Training, is used to compromise the time windows into one single temporal dimension. By encoding the temporal dynamics into a one-dimensional representation, the autoencoder reduces the convolution while preserving the essential information.

MODEL TRAINING

To ensure the highest possible performance of the USID, four different unsupervised machine learning methods were used, k-means, autoencoders, density-based spatial clustering, and self-organizing maps. These methods were chosen due to their diverse methodologies, in order to capture various data structures and driver patterns.

The k-means clustering is an unsupervised machine learning algorithm used to partition a dataset into K distinct, non-overlapping clusters. In the case of this research, the K represents the number of drivers that are to be identified, 16 in total. The output of the k-means model is the predicted driver ID, for each time sequence. The success of this model depends highly on the quality of data processing done previously, especially the data transformation step. If the dimension reduction technique manages to separate the OBD data of each driver in non-overlapping order, the k-means would find the fitting centers of each driver successful. The best results after multiple iterations of training and validating, gave the k-means with 16 centroids, training inertia set to 10 with 10,000 maximal iterations, and removed randomness of training.

Autoencoders' self-supervised nature makes them a suitable choice to extract relevant features or to cluster input sequences. Therefore, the autoencoder can be trained without any labels (Tavakoli *et al.*, 2020). With two one-dimensional convolutional neural network layers and one linear layer the encoder compromises the input dimension from (w, f) to (e), where w is the window length, f is the number of features and e is the dimension of the embedding space. Besides the direct clustering, the embedding layer can be used as input for following clustering approaches while containing the intricate patterns of a time series sequence. An embedding space e of 32 is used for embedding time series sequences. For direct clustering, e corresponds to the number of driver IDs and is therefore 16.

DBSCAN is a clustering algorithm designed to uncover clusters in noisy and irregularly shaped datasets. Unlike other clustering methods like k-means, DBSCAN does not rely on a fixed pre-set number of clusters. It defines two parameters, eps, and min-samples. Eps determines the maximum distances within data points that are considered neighbours. Min-samples specify the minimum number of data points within eps distance to define a dense region (cluster). From a seed data point clusters are iteratively expanded to include all direct or indirect reachable data points within the eps distance, forming dense regions. Data points that remain unreachable or do not meet the density criteria are considered outliers (Ester *et al.*, 1993). For the USID evaluation, 0.44 is used for eps and the number of min samples is 10. The distance is based on an Euclidean distance computed.

Self-organizing maps (SOMs) are a type of artificial neural network that excels in the field of unsupervised machine learning. Introduced by Teuvo Kohonen in the 1980s, SOMs are unique in their ability to transform high-dimensional data into a simplified, two-dimensional representation, often referred to as a map or grid (Ritter and Kohonen, 1989). For the purpose of driver identification, the shape of the SOM used is (16, 1). Each neuron in

the SOM is responsible for recognizing a cluster representing one driver. The training of the SOMs consisting of 250,000 iterations was conducted using specific hyperparameters chosen to achieve optimal results. The parameters include a sigma of 0.7, a learning rate of 0.25, and a neighbourhood function set to 'Gaussian'.

In order to transparently evaluate the models used for USID concept, various external evaluation metrics are used to compare real driver ID with assigned values output by classification models. These vary from pair counting metrics, entropy-based metrics, and set-matching-based metrics. Three pair counting metrics were used for evaluation, rand index, adjusted rand index, and Fowlkes-Mallows index. The entropy-based metric used in this research is normalized mutual information (NMI), and it represents the similarity of clusters by measuring how much information is shared between the clusterings. The last used metric is purity which quantifies the degree of homogeneity in the clusters, without investigating their relationship.

Compared evaluation based on the mentioned criteria is shown in Figure 2. The rand index and adjusted rand index vary from −1 to 1. Other metrics range from 0 to 1, with 1 denoting the highest achievable quality and the proximity to 0 indicating a decrease in each model's quality. The results from the figure show that the driver can be successfully identified with unsupervised ML models and promises the success of the USID concept. The two best models on this dataset turned out to be the k-means clustering and SOM, slightly ahead of DBSCAN. They dominate in every metric, except the purity of the clusters, where autoencoders are on the top. Regardless of their purity, the autoencoders didn't show great results in other metrics.

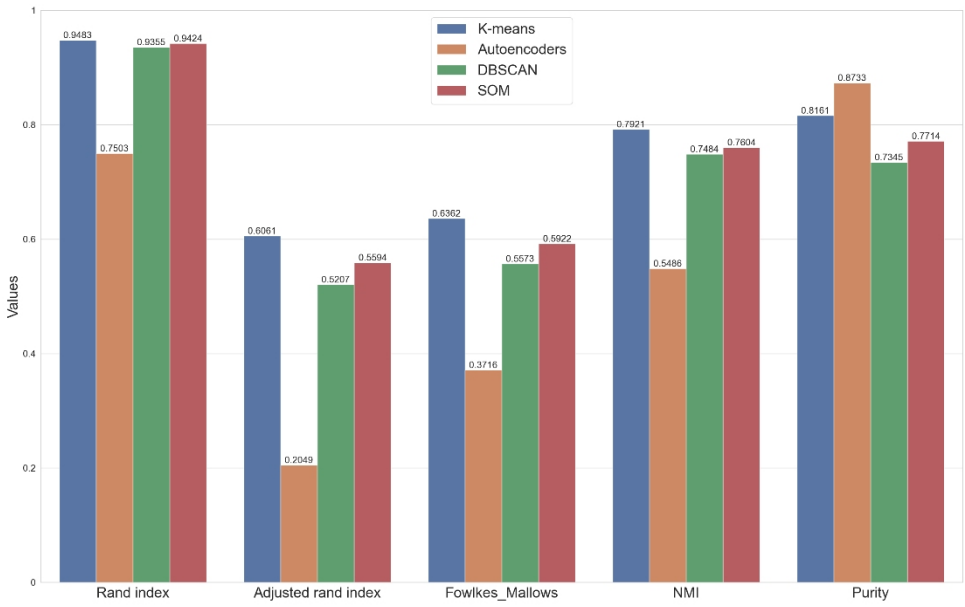

Figure 2: Comparation of performance of different unsupervised ML models used for USID.

EXPERIMENTAL ANALYSIS

The goal of the USID is to identify the driver with the data available already in vehicles. The main feature is its capability to assign the driver ID to the group of data, later interpreted as a driver identification, without using the label of the driver. Additionally, the concept is expendable to the any number of new drivers seamlessly without requiring new labeled data to be fed to the model, due to its unsupervised characteristics.

In order to do the quantification of the concept, besides the evaluation of the models shown previously, the authors have done the experimental analysis. For this purpose, the USID confidence coefficient (U_{cc}) is proposed. The USID confidence coefficient represents how confident the model is, varying from 0 to 1, in detecting the right driver during the driving cycle, see Equation 1. The equation represents the dependency of the coefficient from the OBD II data index x_d during the driving cycle. On the left side, the variable x_s is number of data indexes for which each driver has been successfully identified so far. The constant x_d^* is the specific data index after which the U_{cc} gains relevance. The x_d^* is to be selected according to the requirements of the comfort function used with USID, more specifically it defines how flexible the comfort function is towards the driver identification error. The filtering based on a first order system was used for the first x_d^* OBD II data indexes, explaining the right part of the equation. This was done to remove false confidence information at the beginning due to the low total number of messages processed, rendering every driver ID decision before x_d^* useless. The USID confidence coefficient is specially designed for use in comfort functions since the variable can be differently interpreted with various comfort functions. Comfort functions like ride comfort require a high degree of confidence in driver identification to apply their functionalities and not as early, while only a medium level of identification is needed in most thermal comfort functions at the start of the driving cycle.

$$U_{cc}\left(x_d\right) = \frac{x_s\left(x_d\right)}{x_d};\ x_d \geq x_d^* \ \wedge \ U_{cc}\left(x_d\right) = U_{cc}\left(x_d^*\right)\left(1 - e^{-\frac{4x_d}{x_d^*}}\right);$$
$$x_d < x_d^* \tag{1}$$

The experiment is done in such a way that the OBD II data stream from the driving cycles of each driver is inserted into the USID, and the change in U_{cc} is calculated. The SOM is used as the unsupervised model of USID in the experiment, and the value for x_d^* is picked 20. In this specific case, the frequency of data is 1Hz, and therefore x-axis can be represented as time in seconds with the same labels. The results show the variance of the confidence coefficient for each driver during the time of driving cycle is in the range from medium to high, see Figure 3. The U_{cc} for majority of drivers is in the range of 0.7 to 0.9, showing the very good confidence of models in detecting the driver and success of USID, even during the early stages of the driving cycle. Outside of this range are the perfect confidence in driver ID6 identification, and the lower confidence of drivers ID9 and ID16 between 0.4 and 0.6 after the 200th data index.

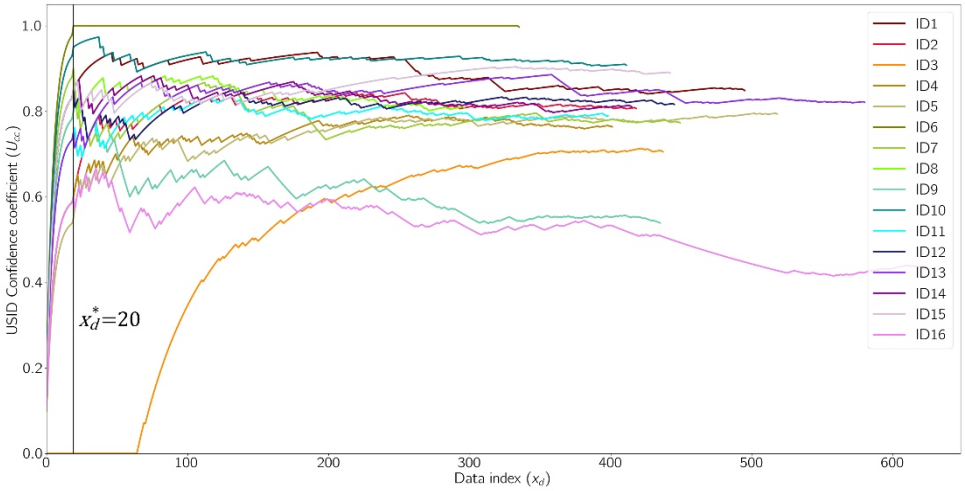

Figure 3: Results of experimental analysis.

CONCLUSION

This paper proposes the USID concept, which deals with the identification of the driver using unsupervised ML techniques, for the purpose of comfort functions. The architecture of the USID is shown in detail. The proposed USID seems suited to adapt to changes in driver behaviour over time, making it a more versatile approach for long-term driver identification. By not requiring labeled data, it simplifies data collection and model training, marking it a practical choice for applications such as comfort functions. These advantages position USID as a promising solution for driver identification in pursuing personalized comfort functions in vehicles.

The results revealed success in the evaluation of USID, with an OBD II dataset of 16 different drivers in one vehicle. Different models were used to validate the concept, where the best results gave k-means and SOM, ahead of DBSCAN and autoencoder models. Covering the 16-driver scenario proposed, the USID with SOM proved capable of coping with the driver identification requirements for comfort functions. The introduced USID confidence coefficient of the driver showed very good overall confidence of the predicted driver during the driving cycles of each driver. The findings outlined suggest that the practical implementation of the USID concept in identifying multiple drivers through unsupervised machine learning models with a high level of confidence is achievable.

REFERENCES

Barreto, C. A. S. (2018). “Kaggle OBD-II datasets.”, [Online]. Available: https://www.kaggle.com/datasets/cephasax/obdii-ds3 (visited on 09/29/2023).

Bloecher, H. L., Dickmann, J., & Andres, M. (2009, September). Automotive active safety & comfort functions using radar. In 2009 IEEE International Conference on Ultra-Wideband (pp. 490–494). IEEE.

Enev, M., Takakuwa, A., Koscher, K., & Kohno, T. (2016). Automobile Driver Fingerprinting. Proc. Priv. Enhancing Technol., 2016(1), 34–50.

Ester, M., Kriegel, H. P., Sander, J., & Xu, X. (1996, August). A density-based algorithm for discovering clusters in large spatial databases with noise. In kdd (Vol. 96, No. 34, pp. 226–231).

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. AI magazine, 13(3), 57.

Girma, A., Yan, X., & Homaifar, A. (2019, November). Driver identification based on vehicle telematics data using LSTM-recurrent neural network. In 2019 IEEE 31st International Conference on Tools with Artificial Intelligence (ICTAI) (pp. 894–902). IEEE.

Guinea, M., Stang, M., Nitsche, I., & Sax, E. (2021). Acceptance of Smart Automated Comfort Functionalities in Vehicles. In Human Interaction, Emerging Technologies and Future Applications IV: Proceedings of the 4th International Conference on Human Interaction and Emerging Technologies: Future Applications (IHIET–AI 2021), April 28-30, 2021, Strasbourg, France 4 (pp. 331–338). Springer International Publishing.

Kern, D., & Schmidt, A. (2009, September). Design space for driver-based automotive user interfaces. In Proceedings of the 1st International Conference on Automotive User Interfaces and Interactive Vehicular Applications (pp. 3–10).

Khan, M. A. A., Ali, M. H., Haque, A. K. M., & Habib, M. T. (2022). A Machine Learning Approach for Driver Identification Based on CAN-BUS Sensor Data. arXiv preprint arXiv:2207.10807.

Kwak, B. I., Woo, J., & Kim, H. K. (2016, December). Know your master: Driver profiling-based anti-theft method. In 2016 14th Annual Conference on Privacy, Security and Trust (PST) (pp. 211–218). IEEE.

Lahlou, A., Ossart, F., Boudard, E., Roy, F., & Bakhouya, M. (2020). A real-time approach for thermal comfort management in electric vehicles. Energies, 13(15), 4006.

Lahlou, A., Ossart, F., Boudard, E., Roy, F., & Bakhouya, M. (2020). Optimal management of thermal comfort and driving range in electric vehicles. Energies, 13(17), 4471.

McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. arXiv preprint arXiv:1802.03426.

Ritter, H., & Kohonen, T. (1989). Self-organizing semantic maps. Biological cybernetics, 61(4), 241–254.

Schaut, S., & Sawodny, O. (2019). Thermal management for the cabin of a battery electric vehicle considering passengers' comfort. IEEE Transactions on Control Systems Technology, 28(4), 1476–1492.

Sharma, R. C., Sharma, S., Sharma, S. K., Sharma, N., & Singh, G. (2021). Analysis of bio-dynamic model of seated human subject and optimization of the passenger ride comfort for three-wheel vehicle using random search technique. Proceedings of the Institution of Mechanical Engineers, Part K: Journal of Multi-body Dynamics, 235(1), 106–121.

Stang, M., Stock, S., Müller, S., Sax, E., & Stork, W. (2022, March). Development of a self-learning automotive comfort function: an adaptive gesture control with few-shot-learning. In 2022 International Conference on Connected Vehicle and Expo (ICCVE) (pp. 1–8). IEEE.

Tavakoli, N., Siami-Namini, S., Khanghah, M. A., Soltani, F. M., & Namin, A. S. (2020). Clustering time series data through autoencoder-based deep learning models. arXiv preprint arXiv:2004.07296.

Uvarov, K., & Ponomarev, A. (2021, January). Driver identification with OBD-II public data. In 2021 28th Conference of Open Innovations Association (FRUCT) (pp. 495–501). IEEE.

Van Ly, M., Martin, S., & Trivedi, M. M. (2013, June). Driver classification and driving style recognition using inertial sensors. In 2013 IEEE Intelligent Vehicles Symposium (IV) (pp. 1040–1045). IEEE.

Vehicle E E System Diagnostic Standards Committee (2017). "Digital annex of E/E diagnostic test modes," in SAE Standard J1979-DA. DOI: 10.4271/J1979DA 201702.

Xie, Y., Liu, Z., Liu, J., Li, K., Zhang, Y., Wu, C., Wang, P., & Wang, X. (2020). A Self-learning intelligent passenger vehicle comfort cooling system control strategy. Applied Thermal Engineering, 166, 114646.

Xun, Y., Liu, J., Kato, N., Fang, Y., & Zhang, Y. (2019). Automobile driver fingerprinting: A new machine learning based authentication scheme. IEEE Transactions on Industrial Informatics, 16(2), 1417–1426.