applied sciences

MDPI

Article

Driving-Style Assessment from a Motion Sickness Perspective Based on Machine Learning Techniques

Jon Ander Ruiz Colmenares *, Estibaliz Asua Uriarte and Inés del Campo

Department of Electricity and Electronics, Faculty of Science and Technology, University of the Basque Country UPV/EHU, 48940 Leioa, Spain

* Correspondence: jonander.ruiz@ehu.eus

Abstract: Ride comfort improvement in driving scenarios is gaining traction as a research topic. This work presents a direct methodology that utilizes measured car signals and combines data processing techniques and machine learning algorithms in order to identify driver actions that negatively affect passenger motion sickness. The obtained clustering models identify distinct driving patterns and associate them with the motion sickness levels suffered by the passenger, allowing a comfort-based driving recommendation system that reduces it. The designed and validated methodology shows satisfactory results, achieving (from a real datasheet) trained models that identify diverse interpretable clusters, while also shedding light on driving pattern differences. Therefore, a recommendation system to improve passenger motion sickness is proposed.

Keywords: motion sickness; ride comfort; driving style; ADAS; clustering

Citation: Ruiz Colmenares, J.A.; Asua Uriarte, E.; del Campo, I. Driving-Style Assessment from a Motion Sickness Perspective Based on Machine Learning Techniques. *Appl. Sci.* **2023**, *13*, 1510. https://doi.org/10.3390/app13031510

Academic Editor: José Salvador Sánchez Garreta

Received: 30 December 2022
Revised: 17 January 2023
Accepted: 18 January 2023
Published: 23 January 2023

1. Introduction

Research into advanced driver assistance system(s) (ADAS) has improved the activity of driving. Dynamic cruise control [1], automatic lane-keeping [2], risk prediction [3] among others, relieve the driver of certain duties, making driving a much more automatic and straightforward task, which translates into better driving conditions.

This rapid development of ADAS has led to comfort being an objective to be researched. Moreover, systems that combine all of these assignments are becoming the foundation of automated driving, and with high automation, drivers become passengers, which may negatively impact their riding comforts [4]. So, as the scope of automated driving widens, comfort should also be one of the considerations to be taken into account even in the early phases of development strategies or driving assistants [5].

Overall, the subjective sensation of comfort makes its evaluation complicated with external factors becoming important. Regarding comfort, age is a notable condition [6] and personality or personal driving preferences make each driver have different assessments for the same situations [7]. In [8], other agents that affect motion sickness are summarized (e.g., smell, sound). Recently, physiological data were 'tried' as predictors of levels of motion sickness. Forehead humidity [9] or complex wearable acquisition that combines brain activity and other physiological factors [10] shows promising results.

Current studies have attempted to accommodate comfort parameters by motion planning [11], speed, suspension control, uneven roads [12], or traffic sign information analysis [13]. Therefore, it is evident that the improvement of the ride comfort of passengers requires an exhaustive analysis of the origins of discomfort. Alternative approaches for motion sickness reduction have also been studied, such as designing an adaptive interface that reduces it [14] or training the visuospatial ability of drivers themselves [15].

Regarding driving style, a consensus driving styles evaluation was done in [16]. It is determined that speed, acceleration, and distances between vehicles are the most relevant differentiating variables. Moreover, research shows a more preferable approach to calmer and less dynamic driving styles [17].

A signal-based analysis was used to identify driving styles and drivers. Smartphone-based signal processing was used to create scoring criteria and study driver differences [18], driving patterns, and maneuvers within styles [19]. Moreover, different machine learning approaches have been utilized with classifier-based recognition of risk levels and styles in [20], or with clustering algorithms to label gathered data into aggressive and smooth classes [21], as a means of driver quality separation and classification using sensor data of driver actions [22,23] and as a predictor for fuel consumption in [24]. Nevertheless, the applications of these algorithms are not commonly researched from the perspective of comfort and motion sickness.

The objective of this research paper was to present an exhaustive methodology that, using measured car signals and different machine learning techniques, determines the origin of motion sickness in passengers during a journey. For that purpose, a methodology that identifies what kind of actions a driver can apply to his/her driving style in order to improve the sensation of motion sickness in the vehicle is proposed.

With this recommendation system in mind, these sub-objectives are presented:

1. A data processing methodology that combines different techniques on car signals in order to obtain an adequate database for machine learning algorithms.
2. Machine learning techniques, particularly cluster models that effectively recognize and group diverse and interpretable driving patterns from a comfort perspective.
3. A recommendation system to fix non-comfortable situations by adapting driver actions.
4. We validate this methodology by means of a real established database and then effectively interpret the results in order to identify and determine specific driver behaviors from a motion sickness perspective and recommend driving improvements.

Section 2 discusses ride comfort and motion sickness evaluation methods and proposals. Section 3 describes the Uyanik instrumented car's database and justifies its use in this work. Section 4 presents the selection of techniques used to preprocess the available data and the utilized machine learning algorithms. Section 5 explores the validation of each of the preprocessing and clustering steps. The results of the two best-obtained models are compared in Section 6 by analyzing cluster interpretability and evaluating their outcomes by comparing driver tendencies. This section also displays driver recommendation examples for each of the identified situations. Finally, conclusions are presented in Section 7.

2. Comfort Evaluation Methods

Comfort is a subjective feeling, and there are no unified and clear definitions for it in academia. The passenger comfort discussed in the present article refers to the discomfort of passengers during a car ride, specifically regarding motion sickness.

The level of ride discomfort is associated with the frequency of the vibration and is directly proportional to its intensity. Furthermore, it has also been observed that increasing the time of exposure to vibration means an increase in discomfort. With this in mind, we know that low-frequency vibrations, close to 1 Hz, are transmitted throughout the body by increasing malaise. Higher frequency vibrations are attenuated by the human body and reduce discomfort. Monotone continuous low-frequency vibrations increase fatigue, while transient vibrations produce stress [25].

The human body responds differently to vibration frequencies depending on the body part and in which direction the force acts. How the frequencies affect humans depends on the proportions of a person's body and the type of frequency that affects the person [26].

Thus, two types of discomfort are differentiated. In [27], discomfort (also named average discomfort, or simply discomfort) is defined as a general feeling of not well-being, while motion sickness is associated with dizziness, fatigue, and nausea. In this work, we will focus on the latter type—motion sickness.

Most of the time, comfort is evaluated using a subjective rating test and/or using electrical accelerometers combined with comfort evaluation parameters. Many of these parameters (vibration dose value (VDV), estimated dose value (eVDV), vibration num-

ber (VN), motion sickness dose value (MSDV), vomit rate (VR)) are part of the essential standards in this area: ISO-2631 [28] and British Standard 6841 [29].

Two main standards are used to evaluate comfort, the British Standard 6841 (BS 6841) and the International Standard 2631 (ISO 2631). Both standards describe ways to evaluate vibration exposure to the human body. They define methods for the measurement of vibrations as well as how to process measurement data to standardized quantified performance measures concerning health, perception, comfort, and motion sickness.

In this work, motion sickness-related frequencies will be studied; between the cited examples of the cited standards, only MSDV and VR can be used.

In the bibliography, there are some alternative indices, such as motion sickness incidence (MSI) [30], and fast motion sickness scale (FMS) [31]. Quantitative methods based on questionnaires can also be noted, such as the motion sickness susceptibility questionnaire (MSSQ) [32]. Nonetheless, given its relevance in this work, the motion sickness dose value is used (MSDV, Equation (1)).

MSDV is a measure of the likelihood of nausea and its evaluation is also determined in ISO 2631 Standard by the w_f filter seen in Figure 1. This standard only provides guidelines for the interpretation of the MSDV for the vertical direction, MSDVz.

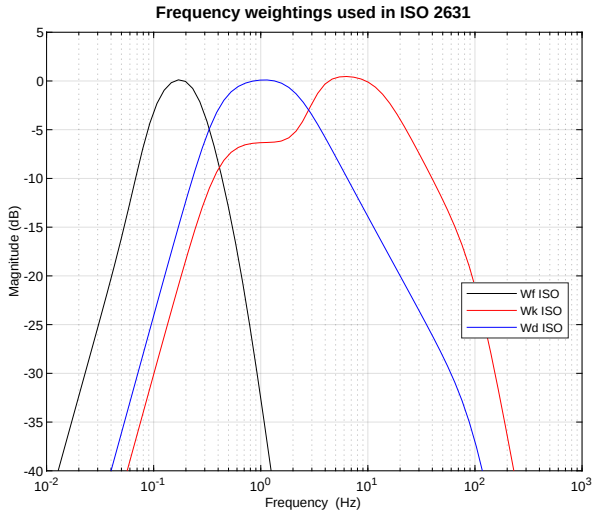


Figure 1. Acceleration filters proposed in the ISO−2631 standard for comfort.

$$MSDV_z = \sqrt{\int_0^T (a_{z,w_f}(t))^2\, dt} \tag{1}$$

Forstberg provides an interpretation of the MSDV for the lateral direction in [33] and our previous real-time experiments [34] have shown that MSDV values obtained in the X and Y axes are substantially bigger than the Z-axis [35], this is a direct consequence of accelerations in X and Y being bigger than the vertical ones when measured in m/s^2 without the added effect of gravity.

The British Standard 6841 (BS 6841) [29] attempts to combine the influences of different axes to obtain a general approach toward the origin of motion sickness by quantifying the lateral and longitudinal contribution as the sum of the parts as seen in Equation (2).

$$MSDV_{Total} = \sqrt{\int_0^T (a_{x,w_f}(t))^2\, dt} + \sqrt{\int_0^T (a_{y,w_f}(t))^2\, dt} \tag{2}$$

In this work, we propose an alternative approach, which also takes into account both X-axis and Y-axis vibrations, but follows the same reasoning as the general discomfort formula for Av proposed in the ISO 2631 [36]. This proposal has already been used in

motion planning [37] and it is observed in Equation (3). The objective is to find common ground in internal structures to derive both X-related and Y-related motion sickness while retaining the ability to identify causes with monodirectional origin. The vertical axis is excluded due to low impact.

$$MSDVxy = \sqrt{\int_0^T (a_{x,w_f}(t))^2\,dt + \int_0^T (a_{y,w_f}(t))^2\,dt} \tag{3}$$

As seen in Equations (1)–(3), motion sickness is cumulative. In order to avoid this effect, a windowing process is done, which allows a generalized approach.

3. Uyanik Instrumented Car Dataset

As highlighted in [34,38], road type, traffic conditions, and the car itself are crucial in the evaluation of passenger discomfort. In this work, given that the objective is to study the influence of driving patterns on comfort, or lack thereof, we considered a group of 18 drivers exhibiting different driving behaviors while driving the same car, along the same route, and in similar environmental conditions.

We used real-world signals from the Uyanik instrumented car [39,40], a Renault Mégane sedan car (seen in Figure 2) that complies with a large number of different signals and travels a fixed route around the city of Istanbul.

Figure 2. Instrumented car utilized in the collection of signals and images of the Uyanik database.

Regarding signals, both the data stream from its CAN-bus and the IMU XYZ accelerations are taken into account: accelerations in all three axes (ACCX, ACCY, ACCZ), rotational speeds (RR, YR, PR), and a variety of car sensor variables, such as vehicle speed (VS), engine RPM (ERPM), steering wheel angle (SWA), steering wheel angle relative speed (SWRS), brake pedal pressure (BP), gas pedal pressure (GP), and percent gas pedal (PGP) were recorded and analyzed. These last seven variables of car sensor signals will be named driver-controlled variables from now on.

In this study, all signals were handled jointly, which requires a re-sampling of the data streams to the highest frequency of 32 Hz. After a thorough analysis, the root means square and variance of the features were computed over 15-s windows (i.e., 480 samples) with a 7.5-s shift. That is to say, the overlapping between consecutive windows is 7.5 s (i.e., 240 samples). Moreover, the longitudinal acceleration signal was differentiated into positive and negative.

The route is presented in Figure 3, with 25 km of road and a duration of 40 min. It includes different types of road sections with different motion sickness predispositions (university campus, road entrances, city, highway, very busy city, and residential areas). The data were labeled with the road types, from type-0 to type-5.

The driver population was selected from academic partners and volunteer creators of the Uyanik dataset. From this database, many of the drivers had to be discarded in order to meet the same environmental conditions (no exceptional events, exactly the same route, no additional tasks). These exhaustive measures left us with 18 drivers. Out of the 18 drivers, 3 of them were female and 15 were male. The ages of the drivers varied, with the youngest driver being a 21-year-old female and the oldest a 61-year-old male.

Figure 3. Uyanik route and its geographical location through different zones of Istanbul.

Road Type Selection

After observing the complete compiled data, with the objective of having a similar driving pattern group, a single road type was selected. Of all possibilities, the main priority was to select the type of road with the greatest variety of motion sickness values. To do this, the data were separated into road types and categorized with colors to express different motion sickness levels (MSDVxy). The obtained results are shown in Figure 4. The Y-axis represents the windows that belong to different comfort levels within each type of road.

With the aforementioned goal, type-1 was selected. This road has all MSDVxy levels presented in the selected dataset, including the highest one, and it is the road with the best distribution of samples at each level. While type-4 also meets this condition, the extremely small sample size would make it impossible to use, so it was discarded.

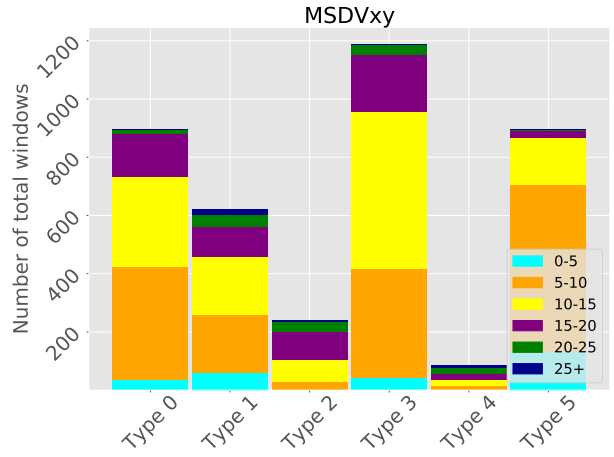


Figure 4. MSDVxy values for each road type. Criteria values are classified in five intervals that can be distinguished by color. The total number of windows depends on the length of each sub-route.

4. Data Processing Methodology

In order to analyze the driving style from a comfort-based perspective and, thus, be able to give driving recommendations for motion sickness improvements, a data processing pipeline was designed and implemented. The initial data are altered and augmented so that they could effectively be used in the training of machine learning techniques, specifically clustering techniques. The method is schematized in Figure 5 and it is summarized in this section.

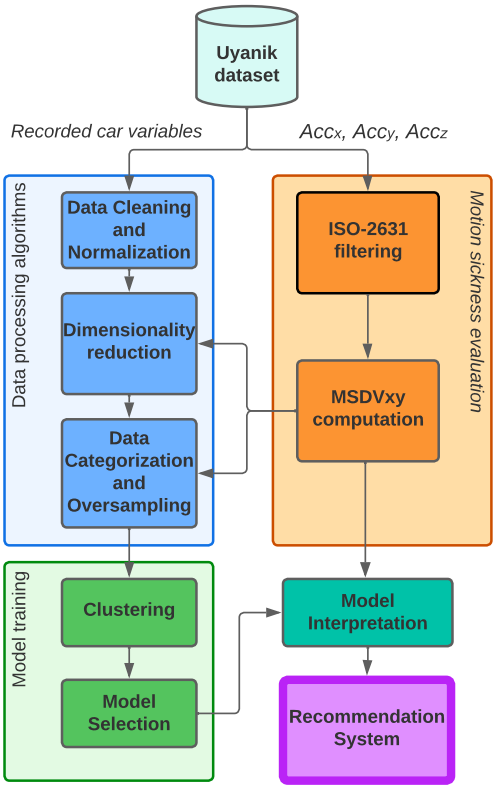


Figure 5. Comfort-oriented processing methodology applied to the Uyanik database.

The Uyanik database presented in Section 3 is used as the starting point. On the one hand, in the **motion sickness evaluation** step, accelerations are used to calculate the MSDVxy variable as explained in Section 2 in order to evaluate passenger motion sickness. On the other hand, all recorded data variables are processed in the **data cleaning and normalization** step, where outliers are analyzed and eliminated, and signals are normalized. Next, in order to select the most relevant variables of the dataset from a comfort perspective, in the **dimensionality reduction** step, different ways to do it are studied and implemented. In the **data categorization and oversampling** step, the MSDVxy value is used to categorize samples in different levels of motion sickness. Those categories will be used to augment the data by utilizing different oversampling methods, which will return a database that is uniformly distributed over motion sickness. Those four steps will output a database perfectly suited for model training. The objective of the **clustering** step is to find interpretable cluster systems that represent different driving patterns, so, the selected normalized and categorized variables of the previous steps are used to train clustering algorithms. Different clustering algorithms were 'tried' to assess the differences and the quality of the obtained results. From the obtained results, in the **model selection,** the best solutions will be chosen in order to understand them from a motion sickness perspective in the named **Model interpretation** step. Finally, the analysis of the obtained driving patterns will be the basis for the proposed **recommendation system**.

Next, the above-mentioned steps are detailed.

4.1. Data Cleaning

The first step is based on identifying and deleting possible error points or samples that are out of reasonable bounds. For this purpose, three different methods were used: manually identifying strange samples, the isolation forest algorithm [41], and the local outlier factor algorithm [42].

4.2. Dimensionality Reduction

In order to simplify the final system, dimensionality reduction will be studied. Different techniques that show improvements in comparison to not using feature selection at all were 'tried' in order to select the most important features [43–45]. The first selection method is based on the correlation analysis between car signal variables and the motion sickness evaluation criteria (MSDVxy). Secondly, recursive feature elimination (RFE) with a Random Forest regressor [46] is used. Moreover, in order to study an alternative approach that focused on obtaining results by starting just from driver-controlled variables, PCA was used to transform and reduce the dimensionality of the original data [47].

4.3. Categorization and Oversampling

MSDVxy is the motion sickness evaluation criterion, and although it is not used later in the unsupervised learning of the clustering algorithms, since it is important that the dataset becomes closer to a uniform distribution over motion sickness levels, it is the criterion for categorization. It is believed that the higher the motion sickness, the smaller the number of samples; thus, the purpose of this section is to generate more high-level discomfort samples to identify the different causes of medium/high discomfort.

Hence, categorization is done by separating motion sickness into different levels. On the other hand, for Oversampling, different algorithms (SMOTE [48], BorderlineSMOTE [49], and SVMSmote [50]) generate new synthetic samples following some simple rules.

4.4. Clustering Methods

Once the main data processing pipeline was applied to the starting data, our newly created database was used to effectively train machine learning techniques. In particular, different clustering algorithms were 'tried'. Algorithms that allowed complete control over the number of clusters became favored. In this situation, model creation and interpretation greatly benefit since a bigger number of clusters of our choosing subsequently create different and richer internal separations, perfect for better interpretations. These algorithms are K-Means [51], SpectralClustering [52], and AgglomerativeClustering [53].

4.5. Model Selection

As the combination of techniques and algorithms will generate many models, an objective criterion is needed. To do this, the silhouette coefficient [54] was employed, which quantified the distance between each point within a cluster and the average distance of all clusters. A data space with perfectly separated clusters scored a coefficient of 1 while a space with very adjacent clusters and an indistinguishable structure scored close to 0.

It is important to note that clustering algorithms are often used with labeled samples, which in turn assess the quality of the models. In this work, there is no real category attached to the samples; the grouping of samples (car signal values) will depend on the internal criteria of the used technique.

4.6. Model Interpretation

Once all models were evaluated, the subsets of models with the highest values of the silhouette coefficient were manually analyzed to identify those with the highest interpretability according to their driving variables. This slightly subjective approach aims to find the clustering solution that distinguishes different driving patterns and, hence, motion sickness levels, and the origin of this lack of comfort.

4.7. Recommendation System

After the combination of actions that generated motion sickness in each cluster was identified, we determined the driving patterns related to higher levels of motion sickness and recommend specific driving improvements that reduced it.

5. Experimental Results of the Applied Methodology

5.1. Data Processing Algorithms

First, in order to clean errors in the dataset, the local outlier factor algorithm was selected, eliminating 25 samples. It was observed that the isolation forest method had a very aggressive approach to data point elimination, which with our limited sample size brought problems down the line.

Concerning dimensionality reduction, first, the Pearson coefficient for correlation analysis was analyzed. In Figure 6, all variables presented in Section 3, are compared to the motion sickness criteria: MSDVxy. Secondly, recursive feature elimination (RFE) with a random forest regressor was applied to all variables until five variables were left. The results of both methods are summarized in Table 1.

In the first group, all variables with a higher factor than 0.5 to MSDVxy were declared of interest. In the second, only the variables which showed a correlation higher than 0.67 were selected. The third group contains the five variables obtained from applying the RFE algorithm. From the three options, the options with fewer variables are preferable, so, the two options that contained five variables from Table 1 were selected. Moreover, SWA showed a 0.9 correlation with a SWRS variable; thus, among the two five-variable options, the variables obtained with RFE were selected.

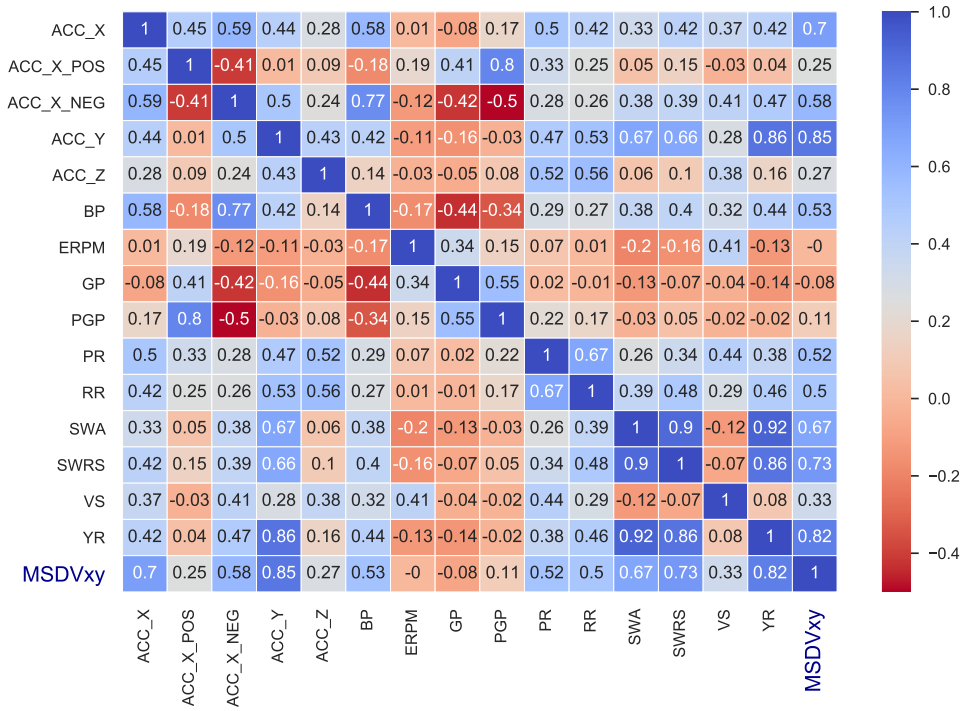


Figure 6. Correlations between all variables that belong to accelerations, rotational speeds, driver−controlled variables, and our discomfort criteria (MSDVxy).

Table 1. Different groups of variables after applying feature selection techniques.

	Variables
CA (Coeff > 0.5)	*ACCy, YR, SWRS, ACCx, SWA, BP, ACCx_neg, PR, RR*
CA (Coeff > 0.67)	*ACCy, YR, SWRS, ACCx, SWA*
RFE (N = 5)	*ACCy, YR, SWRS, ACCx, ACCx_neg*

Concerning PCA technique results, the starting seven driver-controlled variables were transformed into five new PCA variables (PCA_1, PCA_2, PCA_3, PCA_4, PCA_5). These five

variables are enough to explain 95% of the variance of the original dataset. The way that PCA functions is that it finds the direction with the highest percentage (%) of the variance of the data space, thus, PCA1 explains over 35% of the total variance, and PCA2, the second direction with the most variance, explains almost 25%.

The new dimensions of our PCA-created space are a combination of the initial variables, orthogonal, and they minimize the reconstruction error of the samples. Each new variable can be decomposed as a linear combination of the starting parameter used to train the following Equation (4). Coefficients for the PCA model can be found in Table 2. Since we can deconstruct the newly created axis, we can still maintain the high interpretability of the variables.

$$PCA^i = \beta_1^i X_1 + \beta_2^i X_2 + \cdots + \beta_N^i X_N \quad (4)$$

Table 2. Coefficients of the PCA model trained with driver-controlled variables.

	$\mathbf{PCA_1}$	$\mathbf{PCA_2}$	$\mathbf{PCA_3}$	$\mathbf{PCA_4}$	$\mathbf{PCA_5}$
BP	0.429	0.403	0.357	0.198	−0.69
ERPM	−0.343	0.570	−0.387	−0.546	−0.244
GP	−0.184	−0.03	−0.421	0.164	−0.256
PGP	−0.126	−0.078	−0.510	0.653	−0.211
SWA	0.561	0.039	−0.336	−0.164	0.225
SWRS	0.571	0.085	−0.402	−0.053	0.107
VS	−0.089	0.705	0.104	0.423	0.536

Regarding oversampling, after considering different possibilities, MSDVxy was divided into five equidistant categories, each representing a higher motion sickness sensation ranging from very low to very high as seen in Table 3. It is obvious that the MSDVxy distribution over all samples is not regular since there are many more samples with low MSDVxy than high MSDVxy.

Table 3. Categorization divisions and the original values of the utilized criteria.

***MSDVxy* (Normalized)**	***MSDVxy* (Original Values)**	***Category***	***Sample Number***
0–0.2	0–3.99	Very Low	371
0.2–0.4	3.99–5.71	Low	335
0.4–0.6	5.71–8.30	Normal	154
0.6–0.8	8.30–11.19	High	43
0.8–1	11.19–28.83	Very High	7

The three algorithms (SMOTE, BorderlineSmote, and SVMSmote) perform their tasks of generating high-motion sickness samples adequately, balancing the number of samples in each category. As a result, all categories had 432 samples after the algorithms were applied. It was concluded that the best results were achieved with BorderlineSmote and SVMSmote algorithms, with almost no difference between them. Applying any of those oversampling algorithms showed big improvements over not applying oversampling algorithms at all.

5.2. Model Training

The first batch of results studied the models created by the three clustering algorithms for the two groups of variables selected in the previous subsection. The process has handled a variety of parameters, such as the number of clusters, initializations, and different oversampling algorithms.

The obtained silhouette coefficients for the first group of variables (ACCy, YR, SWRS, ACCx, ACCx_neg) are presented in Table 4. Moreover, the results of the silhouette coefficients regarding the models created with the PCA variables (PCA_1, PCA_2, PCA_3, PCA_4, PCA_5) using the same combination of algorithms and parameters are shown in Table 5.

Concerning the clustering result evaluation, as the representation spaces used to evaluate the silhouette coefficients are not the same, coefficient results for both approaches cannot be directly compared. However, we can argue that K-means obtained the best results in both classes. The difference is especially high in Table 5 where the worst K-means model to appear in the Table is almost better than any model of the other algorithms. As a general rule, fewer clusters obtain better scores but since the final objective is to utilize the created clusters to identify driving patterns and styles, more clusters may be beneficial in pattern identification.

Table 4. Obtained silhouette coefficients for feature-selected variables.

Number of Clusters	Clustering Algorithm		
	Spectral	**Agglomerative**	**K-Means**
N = 5	0.3269	0.3581	0.3978
N = 6	0.2871	0.3492	0.3791
N = 7	0.2652	0.3159	0.3394
N = 8	0.2736	0.29246	0.3375
N = 9	0.2675	0.2995	0.3149

Table 5. Obtained silhouette coefficients for PCA-generated variables.

Number of Clusters	Clustering Algorithm		
	Spectral	**Agglomerative**	**K-Means**
N = 5	0.2727	0.3053	0.3133
N = 6	0.2468	0.2545	0.3171
N = 7	0.2714	0.2623	0.3034
N = 8	0.2763	0.2674	0.3051
N = 9	0.2798	0.2695	0.3048

Due to the nature of the data, a good amount of overlapping is expected, which explained the low silhouette scores, but the results allow us to objectively identify the best models.

6. Cluster Interpretation and Feasible Recommendations

Once the best models are identified, the final model selection is done manually by analyzing the best previously selected ten models of each approach. In this section, two models with high silhouette scores (with interpretable rich internal structures) were analyzed.

Although clustering was done with the selected variables detailed in Section 5.2, once clustering was completed, a statistical analysis of all driver-controlled variables for each cluster was computed. Relating this analysis with the level of MSDVxy, it is possible to diagnose the driving actions that cause motion sickness. Therefore, the interpretation of each cluster can be used to assist the car or driver with recommendations in order to improve the motion sickness of passengers.

6.1. Solution 1: Variables Selected by the RFE Method and Classified by K-Means Algorithm

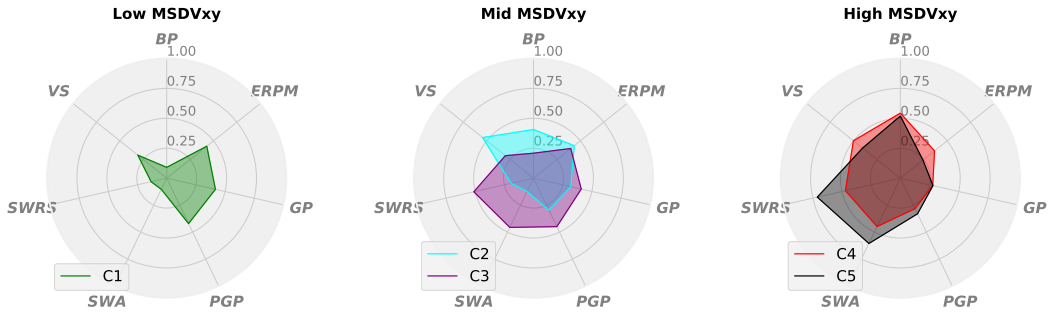
The model was created with the variables *ACCx, ACCx_neg, ACCy, YR, and SWRS*, which were selected using correlation and the recursive feature elimination method detailed in Section 4.2.

The mean value of MSDVxy for each cluster is displayed in Table 6, whereas the mean of the driver-controlled variables for each of the groupings is shown in Figure 7. The generated 5 clusters ordered from C1 (Green) to C5 (Black) are interpretable for the following reasons:

- C1 (Green): BP, SWA, and SWRS values are low. It also has one of the slowest VSs. It stands out that while ERPM and GP values are high compared to other clusters, motion sickness is the lowest of all. No recommendation would be given to a window of this color.
- C2 (Cyan): It contains the highest average VS of all clusters; the SWA and SWRS variables are very low (similar to those in the C1 cluster), but the BP variable shows moderate use. A more careful and smooth brake pedal usage or slower speeds would reduce the medium motion sickness of this cluster.
- C3 (Purple): While the motion sickness sensation is just a bit higher than that of C2, the causes are completely different; samples in the C3 cluster show a correct and small BP usage combined with low VS and similar motor and acceleration values to C1. The SWA and SWRS variables are the key difference. A reduction in sharp turns would reduce the steering wheel use, the most important factor in the motion sickness of this cluster.
- C4 (Red): A high BP usage with sharp SW actions. Red shows a high level of motion sickness. Reducing the brake pedal and the steering wheel's aggressive use would be the most direct recommendation.
- C5 (Black): It is the smallest cluster, and it is just a more extreme version of the red cluster, with similar BP and higher SWA and SWRS values. The same recommendations as the ones for C4 apply.

Table 6. The obtained MSDVxy average value when all samples of each cluster (C_i) of Solution 1 are considered.

Clusters (C_i)	*MSDVxy*
C1 (Green)	0.15
C2 (Cyan)	0.29
C3 (Purple)	0.36
C4 (Red)	0.49
C5 (Black)	0.75

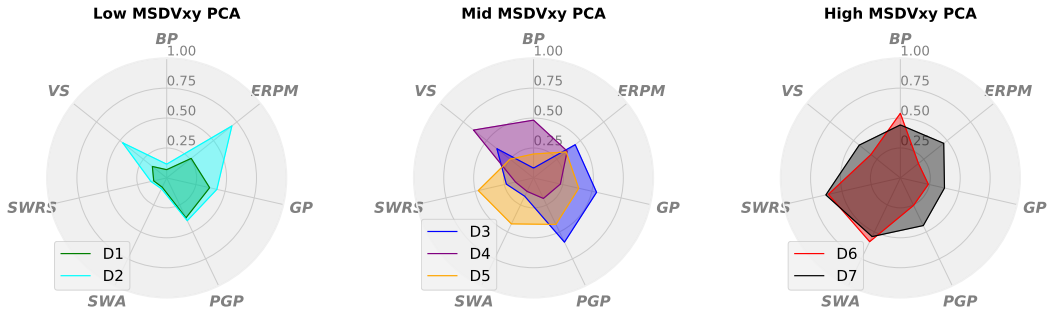


(**a**) Low MSDVxy (**b**) Medium MSDVxy (**c**) High MSDVxy

Figure 7. Spider graphs of low (**a**), medium (**b**), and high (**c**) MSDVxy clusters in the models trained with the RFE chosen variables. The visualized variables can be represented with values in the [0, 1] range since they are normalized. The combination of the mean of each variable gives us the characteristic form of each cluster.

6.2. Solution 2: Variables Obtained by the PCA Method and Classified by K-Means Algorithm

The model was created with the five PCA variables defined in Section 4.2. The mean value of MSDVxy for each cluster is displayed in Table 7, whereas the means of the driver-controlled variables for each grouping is shown in Figure 8. The different meanings and recommendations for the clusters are as follows:



(**a**) Low MSDVxy (**b**) Medium MSDVxy (**c**) High MSDVxy

Figure 8. Spider graphs of low (**a**), medium (**b**), and high (**c**) MSDVxy clusters in the models trained with PCA-transformed variables. The visualized variables can be represented with values in the [0, 1] range since they are normalized. The combination of the mean of each variable gives us the characteristic form of each cluster.

Table 7. The obtained MSDVxy average values when all samples of each cluster (D_i) of Solution 2 are considered.

Clusters (D_i)	*MSDVxy*
D1 (Green)	0.11
D2 (Cyan)	0.19
D3 (Blue)	0.26
D4 (Purple)	0.26
D5 (Orange)	0.31
D6 (Red)	0.49
D7 (Black)	0.58

- D1 (Green): Baseline cluster that shows very good driving with the lowest motion sickness. GP and PGP have medium values but do not translate into motion sickness. It is also characterized by very low values in everything else, BP pressure, SW, and VS. No recommendation is needed here.
- D2 (Cyan): With the second-lowest motion sickness value, this cluster group samples with very high ERPM and medium GP, PGP, and VS, but shares low BP and SW values in D1. No recommendation is needed here.
- D3 (Blue): The same motion sickness as D3, but instead of abrupt BP usage, D4 groups the samples with even higher GP values, combined with medium VS. Smoother accelerations would result in direct improvements.
- D4 (Purple): Medium motion sickness; holds the highest average VS with medium-high BP values. Other variables are either medium or low. Driving slower or smoother brake pedal usage would be ideal recommendations.

These first four clusters do not have high steering wheel usage, so recommendations need to be focused on speed and pedals.

- D5 (Orange): The first cluster with medium SWA and SWRS values; it does not have any high or distinguishable high-variable value. It has low BP and VS. A smoother steering wheel usage would significantly improve the generated motion sickness.

- D6 (Red): High motion sickness cluster related to very high SWA and SWRS values with aggressive BP management. The improvement of the problematic variables (steering wheel and brake pedal) would make the samples closer to the D5 cluster.
- D7 (Black): High motion sickness cluster related to high SWA and SWRS values with aggressive BP management and high VS. Comfort improvement could be achieved by smoother steering wheel usage, slower driving, and less aggressive braking. The main difference with D6 is the higher vehicle speed and ERPM.

While the cluster numbers are higher than in the previous Section 6.1, many of the generated structures from the previous solution also appear here. The worst clusters, red (C4 and D6) and black (C6 and D7), are very similar in both models sharing high BP, SWA, SRWS, and medium speeds. The C3 cluster of Solution 1 is very similar to the D5 of Solution 2 with medium-high SWA, SWRS, ERPM, GP, and PGP. The C2 cluster of Solution 1 is also similar to the D4 cluster of Solution 2, obtaining the characteristically high BP, VS, and little of everything else. The most comfortable cluster of Solution 1, C1 (Figure 7), also seems to be a mixture of the comfortable D1 and D2 clusters of Figure 8 with D1 obtaining the lower end of ERPM and VS, and D2 obtaining the higher end of these variables.

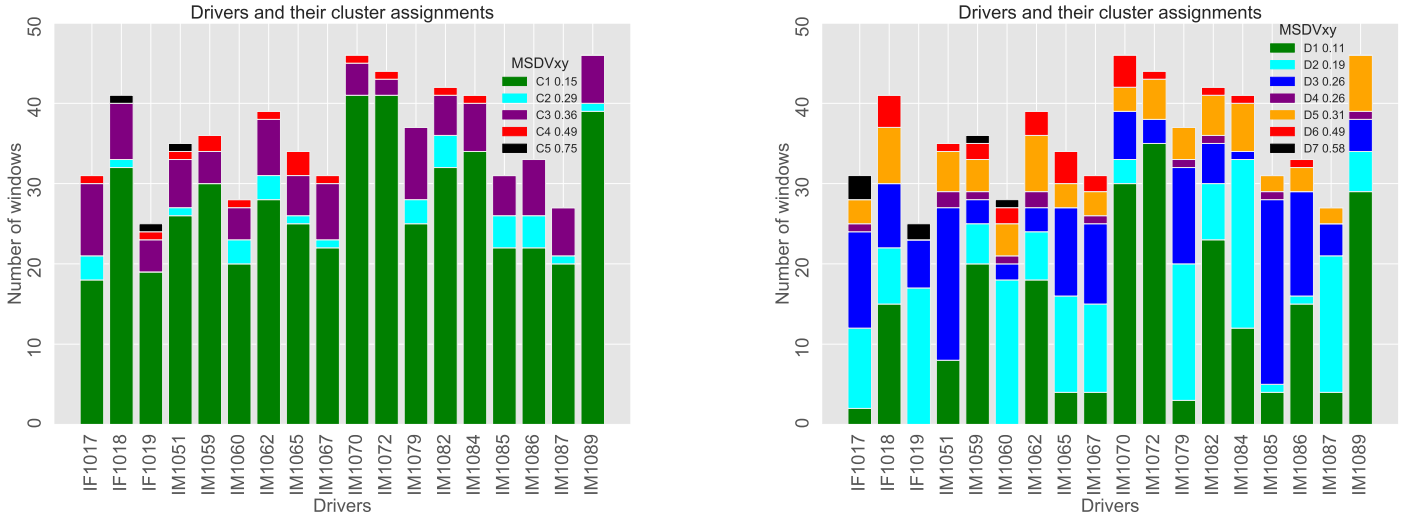
This last model shown in Figure 8 is used in the following sections, where driving style and temporal evolution are studied; thus, Table 8 summarizes the clustering driving patterns and driving advice recommendations in each case.

Table 8. Motion sickness level of each cluster, the characteristic that causes discomfort in passengers and the recommendations given to reduce it.

Cluster	*MSDV Level*	*Driving Pattern*	*Recommendation*
D1 (Green)	*Low*	*None*	*None*
D2 (Cyan)	*Low*	*High ERPM*	*None*
D3 (Blue)	*Medium*	*High GP, PGP, Medium VS*	*Reduce GP, PGP*
D4 (Purple)	*Medium*	*High VS, Medium-High BP*	*Reduce VS, BP*
D5 (Orange)	*Medium*	*Medium SW*	*Reduce SW*
D6 (Red)	*High*	*High SW and BP*	*Reduce SW, BP*
D7 (Black)	*High*	*High SW, Medium Everything*	*Reduce SW, VS, BP*

6.3. Driving Style Analysis

Driving styles can be deduced based on cluster assignment by utilizing Figure 9, which shows the cluster distribution for models created by the RFE variables (left, Ci clusters, *Solution 1* from Section 6.1) and PCA variables (right, Di clusters, *Solution 2* from Section 6.2). Those figures show the number of windows belonging to each cluster for each driver traveling the same route. The number of windows determines the duration of the route; faster drivers will have fewer windows.



(**a**) (**b**)

Figure 9. Cluster assignments from the chosen model created after using the RFE feature selection (**a**), C_i clusters of Solution 1) or PCA (**b**), D_i clusters of Solution 2) on the driver-controlled variables. Each vertical bar represents one driver, starting with IF1017 and ending with IM1089; the bar represents how many points from each driver belongs to what cluster. This allows us to visualize the distribution of each driver and its comparison to others.

As the model obtained in *Solution 2* offers a more diverse cluster result with seven distinct clusters, we will use this clustering to determine driver style differences. Table 9 summarizes the figure interpretation, driving style determination, and possible recommendations.

Many of these conclusions are applicable if we use the solution obtained by applying RFE (Figure 9a); however, since the models are different, some patterns are not completely represented and making direct links between the two models can be indicative but not decisive. Drivers that have the lowest amounts of C1 windows, also have low D1 and D2 cluster memberships. The only notable exception seems to be IM1051, which has 80% of C1 windows but only a small number of windows are in D1. We can safely say that both models have categorized the low MSDVxy samples of each driver in different yet similar ways. Concerning clusters related to high steering wheel usage of the PCA model (D5, D6, and D7) and RFE variables (C3, C4, C5), it seems that D5 is perfectly represented in the C3 cluster in driving style distribution. With C4 and C5, direct comparisons are not as useful since the error of a small number of windows could completely change our evaluation, but we can observe how all five drivers with no C4 and C5 clusters do not have D6 or D7 (except IM1086, which has a low D6).

Table 9. Driving style categorization, driver special features, and possible recommendations.

Drivers	Distinctive Clustering or Special Features	Possible Recommendations
IF1019	Fastest driver	-
IM1070, IM1072, IM1089	Slowest drivers	-
IM1072	Best driver Highest number of D1 No high-risk clusters	-
IF1019, IM1060, IM1079, IM1084, IM1087	High number of D2	Smoother gear change and ERPM
IF1017, IM1051, IM1085	High number of D3	Smoother GP, PGP and ERPM
IM1051, IM1062	High number of D4	Smoother VS and BP

Table 9. *Cont.*

Drivers	Distinctive Clustering or Special Features	Possible Recommendations
IF1018, IM1062, IM1089	High number of D5	Smoother SW
IF1018, IM1059, IM1060, IM1062, IM1065, IM1070	High number of D6	Smoother SW and BP
IM1059, IM1060	Presence of D7, D6 and D5	Smoother general driving is advised: SW, BP, VS reduction
IF1017	Presence of D7 but no D6	Smoother general driving is advised: SW, BP reduction
IF1019	Presence of D7 but no D6, D5	Smoother general driving is advised in exceptional cases

6.4. Temporal Evolution

Following the results and analysis of the previous sections regarding *Solution 2*, another way to visualize the distribution of the assigned clusters is presented in Figure 10. The figure examines the windows on a temporal basis, showing how the cluster assignment of each driver changes within the route.

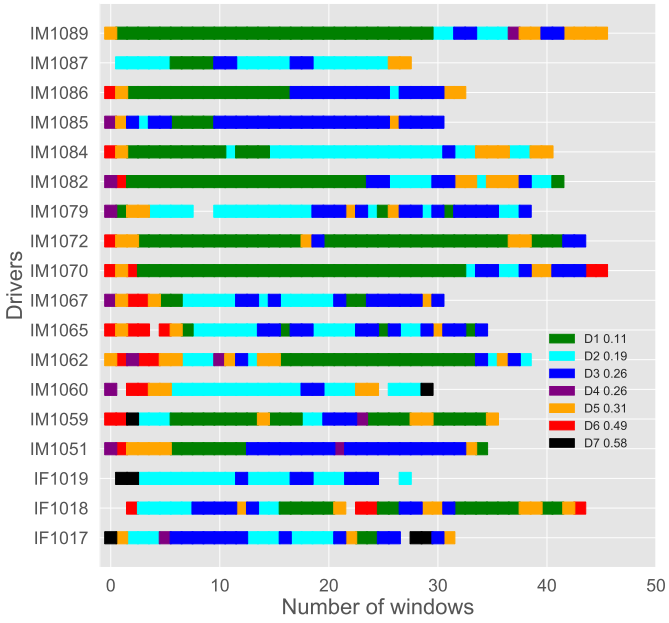


Figure 10. Temporal evolution of the cluster assignments of all drivers for Solution 2 clusters. Each horizontal bar represents a driver, starting with IF1017 and ending with IM1089. Each bar is composed of a number of windows, depending on how fast they completed the route. By observing the order of the assigned clusters, we can assume the nature of the road and observe how each driver approaches the route.

It is possible to observe some blank windows, for example, the first two windows of driver IF1018. These empty windows are those that were eliminated in the data cleaning step but are still represented so that the chronology is maintained.

By comparing the clusters and recommendations summarized in Table 8 with Figure 10, we can observe the path of all drivers. The first direct conclusion is that, as observed in our previous work [34], the route features play a key part in passenger comfort. Data support that the route has a rough start, with many of the D5, D6, and D7 windows located at the start. Moreover, while high-risk windows are located in the same parts, the number of windows, an indicator of the intensity of the pattern, can be clearly discerned. Thus,

drivers, such as IM1067, IM1065, IM1062, and IM1051, exhibit at least four medium/high-risk windows. Driver IM1089 only has one D5 window before moving into a continuous D1 assignment.

Furthermore, while changes between clusters do happen, most changes are restricted to clusters of the same motion sickness severity.

7. Conclusions

This work presents an exhaustive methodology that combines data processing techniques and clustering algorithms in order to determine which driver actions are responsible for the motion sickness in passengers.

After analyzing numerous data augmentation and transformation techniques, the best data processing pipeline was designed, and an exhaustive clustering analysis was conducted in order to obtain a good recommendation system that is comfort-based.

The pipeline consists of data cleaning by the local outlier factor (LOF), data normalization, a dimensionality reduction process based on PCA, and the data augmentation of either SVMSmote or BorderlineSmote, in conjunction with a K-means clustering algorithm; it is the most promising.

The method was validated with a subset of the established Uyanik database. The results show that a balanced number of clusters with distinct driving patterns and different levels of motion sickness is achieved. This cluster system can be used to distinguish driving styles and propose driving advice. Thus, the main objective of creating a recommendation system based on comfort (by studying driving patterns) was proven possible and was achieved.

The proposed methodology, algorithms, and analysis have good results on the selected road type, which sheds light on the generalization of the proposed approach.

The next reasonable step is the generalization of the methodology in order to apply it to different types of roads. A robust model that must take the road type into account to correctly assess different situations is proposed. Furthermore, instead of focusing on the motion sickness analysis, the generalization of the methodology is proposed in order to use it for general discomfort improvement. Moreover, the physical implementation of the obtained model in field programmable gate array (FPGA) hardware will be carried out. Finally, a real-time comfort analysis (before and after the driving recommendations) was executed and will be compared with passengers' answers from questionnaires in order to validate the complete system.

Author Contributions: Conceptualization and methodology, validation, formal analysis, and investigation, J.A.R.C. and E.A.U.; software, data curation and writing—original draft preparation, J.A.R.C.; writing—review, visualization, and supervision, E.A.U. and I.d.C.; project administration and funding acquisition, E.A.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Basque Government; partial support of this work was received from the project KK-2021/00123 Autoeval and the University of the Basque Country UPV/EHU, grant GIU21/007.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The utilized dataset is not publicly available and thus, we cannot share the data.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Eichelberger, A.H.; McCartt, A.T. Toyota drivers' experiences with Dynamic Radar Cruise Control, Pre-Collision System, and Lane-Keeping Assist. *J. Saf. Res.* **2016**, *56*, 67–73. [CrossRef] [PubMed]
2. Yenikaya, S.; Yenikaya, G.; Düven, E. Keeping the Vehicle on the Road: A Survey on on-Road Lane Detection Systems. *ACM Comput. Surv.* **2013**, *46*, 1–43. [CrossRef]

3. Arbabzadeh, N.; Jafari, M. A Data-Driven Approach for Driving Safety Risk Prediction Using Driver Behavior and Roadway Information Data. *IEEE Trans. Intell. Transp. Syst.* **2018**, *19*, 446–460. [CrossRef]
4. Elbanhawi, M.; Simic, M.; Jazar, R. Improved manoeuvring of autonomous passenger vehicles: Simulations and field results. *J. Vib. Control* **2015**, *23*, 1954–1983. [CrossRef]
5. Woolridge, E.; Chan-Pensley, J. *Measuring User's Comfort in Autonomous Vehicles*; Human Drive: Milton Keynes, UK, 2022.
6. Hartwich, F.; Beggiato, M.; Krems, J. Driving Comfort, Enjoyment, and Acceptance of Automated Driving—Effects of Drivers' Age and Driving Style Familiarity. *Ergonomics* **2018**, *61*, 1–55. [CrossRef]
7. Bellem, H.; Thiel, B.; Schrauf, M.; Krems, J.F. Comfort in automated driving: An analysis of preferences for different automated driving styles and their dependence on personality traits. *Transp. Res. Part F Traffic Psychol. Behav.* **2018**, *55*, 90–100. [CrossRef]
8. Karlsson, N.; Helena, T. *Motion Sickness in Cars: Physiological and Psychological Influences on Motion Sickness*; Department of Product and Production Development, Charlmers University of Technology: Gothenburg, Sweden, 2012.
9. Bando, S.; Shiogai, Y.; Hirao, A. Development of Evaluating Methods for Passenger's Motion Sickness in Real Driving Environment. *Int. J. Automot. Eng.* **2021**, *12*, 72–77. [CrossRef]
10. Tan, R.; Li, W.; Hu, F.; Xiao, X.; Li, S.; Xing, Y.; Wang, H.; Cao, D. Motion Sickness Detection for Intelligent Vehicles: A Wearable-Device-Based Approach. In Proceedings of the 2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC), Macau, China, 8–12 October 2022; pp. 4355–4362. [CrossRef]
11. Moazen, I.; Burgio, P. A Full-Featured, Enhanced Cost Function to Mitigate Motion Sickness in Semi- and Fully-autonomous Vehicles. In Proceedings of the Conference: 7th International Conference on Vehicle Technology and Intelligent Transport Systems, Online, 28–30 April 2021; pp. 497–504. [CrossRef]
12. Wu, J.; Zhou, H.; Liu, Z.; Gu, M. Ride Comfort Optimization via Speed Planning and Preview Semi-Active Suspension Control for Autonomous Vehicles on Uneven Roads. *IEEE Trans. Veh. Technol.* **2020**, *69*, 8343–8355. [CrossRef]
13. Tang, X.; Duan, Z.; Hu, X.; Pu, H.; Cao, D.; Lin, X. Improving Ride Comfort and Fuel Economy of Connected Hybrid Electric Vehicles Based on Traffic Signals and Real Road Information. *IEEE Trans. Veh. Technol.* **2021**, *70*, 3101–3112. [CrossRef]
14. Hwang, S.; Sama, M.R.; Kuhn, S.; Erusu, V.; Raiti, J. An Adaptive Tilting Interface to Alleviate Motion Sickness for Passengers in Vehicles. In Proceedings of the 15th International Conference on PErvasive Technologies Related to Assistive Environments; Association for Computing Machinery: New York, NY, USA, 2022; pp. 323–324. [CrossRef]
15. Smyth, J.; Jennings, P.; Bennett, P.; Birrell, S. A novel method for reducing motion sickness susceptibility through training visuospatial ability—A two-part study. *Appl. Ergon.* **2021**, *90*, 103264. [CrossRef]
16. Winner, H.; Hakuli, S.; Wolf, G. *Handbuch Fahrerassistenzsysteme*; Springer: Berlin, Germany, 2009. [CrossRef]
17. Roßner, P.; Bullinger-Hoffmann, A. *How Do You Want to be Driven? Investigation of Different Highly-Automated Driving Styles on a Highway Scenario*; Springer International Publishing: New York City, NY, USA, 2019; pp. 36–43. [CrossRef]
18. Castignani, G.; Derrmann, T.; Frank, R.; Engel, T. Driver Behavior Profiling Using Smartphones: A Low-Cost Platform for Driver Monitoring. *IEEE Intell. Transp. Syst. Mag.* **2015**, *7*, 91–102. [CrossRef]
19. Johnson, D.A.; Trivedi, M.M. Driving style recognition using a smartphone as a sensor platform. In Proceedings of the 2011 14th International IEEE Conference on Intelligent Transportation Systems (ITSC), Washington, DC, USA, 5–7 October 2011; pp. 1609–1615. [CrossRef]
20. Xue, Q.; Wang, K.; Lu, J.; Liu, Y. Rapid Driving Style Recognition in Car-Following Using Machine Learning and Vehicle Trajectory Data. *J. Adv. Transp.* **2019**, *2019*, 9085238. [CrossRef]
21. Bhoraskar, R.; Vankadhara, N.; Raman, B.; Kulkarni, P. Wolverine: Traffic and road condition estimation using smartphone sensors. In Proceedings of the 2012 Fourth International Conference on Communication Systems and Networks (COMSNETS 2012), Bangalore, India, 3–7 January 2012; pp. 1–6.
22. Kalsoom, R.; Halim, Z. Clustering the driving features based on data streams. In Proceedings of the INMIC, Lahore, Pakistan, 19–20 December 2013; pp. 89–94.
23. Constantinescu, Z.; Marinoiu, C.; Vladoiu, M. Driving Style Analysis Using Data Mining Techniques. *Int. J. Comput. Commun. Control* **2010**, *V*, 654–663. [CrossRef]
24. Ping, P.; Qin, W.; Xu, Y.; Miyajima, C.; Takeda, K. Impact of driver behavior on fuel consumption: Classification, evaluation and prediction using machine learning. *IEEE Access* **2019**, *7*, 78515–78532. [CrossRef]
25. Javier, G.S. Generation of Ride Comfort Index. Ph.D. Thesis, Universidad Politécnica de Barcelona, Barcelona, Spain, 2014.
26. Griffin, M.J. Discomfort from feeling vehicle vibration, Vehicle System Dynamics. *Veh. Syst. Dyn.* **2007**, *45*, 679–698. [CrossRef]
27. Svensson, L.; Eriksson, J. Tuning for Ride Quality in Autonomous Vehicle: Application to Linear Quadratic Path Planning Algorithm, Ph.D. Thesis, Uppsala Universitet, Uppsala, Sweden 2015.
28. *ISO 2631-1*; Mechanical Vibration and Shock—Evaluation of Human Exposure to Whole-Body Vibration—Part 1: General Requirements. International Organisation for Standarisation: Geneva, Switzerland, 1997.
29. *BS 6841*; Guide to Measurement and Evaluation of Human Exposure to Whole-Body Mechanical Vibration and Repeated Shock. British Standards Institution: London, UK, 1987.
30. Cepowski, T. The prediction of the Motion Sickness Incidence index at the initial design stage. *Sci. J. Marit. Univ. Szczec.* **2012**, *31*, 31–45.
31. Kamijo, K.; Tsujimura, H.; Obara, H.; Katsumata, M. Evaluation of Seating Comfort. *SAE Trans.* **1982**, *91*, 2615–2620.

32. Golding, J.F. Motion sickness susceptibility questionnaire revised and its relationship to other forms of sickness. *Brain Res. Bull.* **1998**, *47*, 507–516. [CrossRef]
33. Forstberg, J. Ride Comfort and Motion Sickness in Tilting Trains. Ph.D. Thesis, KTH Royal Institute of Technology, Stockholm, Sweden, 2000.
34. Asua, E.; Gutiérrez-Zaballa, J.; Mata-Carballeira, O.; Ruiz, J.A.; del Campo, I. Analysis of the Motion Sickness and the Lack of Comfort in Car Passengers. *Appl. Sci.* **2022**, *12*, 3717. [CrossRef]
35. Griffin, M.J.; Newman, M.M. An experimental study of low-frequency motion in cars. *Proc. Inst. Mech. Eng. Part D J. Automob. Eng.* **2004**, *218*, 1231–1238. [CrossRef]
36. *UNE-ISO 2631-1:2008*; Vibraciones y Choques Mecánicos. Evaluación de la Exposición Humana a Vibraciones de Cuerpo Entero. Parte 1: Requisitos Generales. International Organization for Standardization: Geneva, Switzerland, 2008.
37. Li, D.; Hu, J. Mitigating Motion Sickness in Automated Vehicles With Frequency-Shaping Approach to Motion Planning. *IEEE Robot. Autom. Lett.* **2021**, *6*, 7714–7720. [CrossRef]
38. Ericsson, E. Independent driving pattern factors and their influence on fuel.-use and exhaust emission factors. *Transp. Res. D. Transp. Environ.* **2001**, *6*, 325–345. [CrossRef]
39. Abut, H.; Erdoğan, H.; Erçil, A.; Çürüklü, B.; Koman, H.C.; Taś, F.; Argunśah, A.Ö.; Cośar, S.; Akan, B.; Karabalkan, H.; et al. Real-world data collection with UYANIK. In *In-Vehicle Corpus and Signal Processing for Driver Behavior*; Springer: New York, NY, USA, 2009; pp. 23–43. [CrossRef]
40. Abut, H.; Erdoğan, H.; Erçil, A.; Çürüklü, A.B.; Koman, H.C.; Tas, F.; Argunşah, A.Ö.; Akan, B.; Karabalkan, H.; Çökelek, E.; et al. Data Collection with "UYANIK": Too Much Pain; But Gains Are Coming. In Proceedings of the Biennial on DSP for In-Vehicle and Mobile Systems, Istanbul, Turkey, 17–19 June 2007.
41. Liu, F.T.; Ting, K.M.; Zhou, Z.H. Isolation Forest. In Proceedings of the 2008 Eighth IEEE International Conference on Data Mining, Pisa, Italy, 15–19 December 2008; pp. 413–422. [CrossRef]
42. Breunig, M.M.; Kriegel, H.P.; Ng, R.T.; Sander, J. LOF: Identifying Density-Based Local Outliers. *SIGMOD Rec.* **2000**, *29*, 93–104. [CrossRef]
43. Senan, E.; Al-Adhaileh, M.; Alsaade, F.; Aldhyani, T.; Alqarni, A.; Alsharif, N.; Uddin, M.I.; Alahmadi, A.; Jadhav, M.; Alzahrani, Y. Diagnosis of Chronic Kidney Disease Using Effective Classification Algorithms and Recursive Feature Elimination Techniques. *J. Healthc. Eng.* **2021**, *2021*, 1004767. [CrossRef]
44. Erkmen, B.; Yıldırım, T. Improving classification performance of sonar targets by applying general regression neural network with PCA. *Expert Syst. Appl.* **2008**, *35*, 472–475. [CrossRef]
45. Dash, R.; Mishra, D.; Rath, A.; Acharya, M. A hybridized K-means clustering approach for high dimensional dataset. *Int. J. Eng. Sci. Technol.* **2010**, *2*, 59–66. [CrossRef]
46. Liaw, A.; Wiener, M. Classification and Regression by Random Forest. *Forest* **2001**, *23*, 18–22.
47. Karl Pearson, F.R.S. LIII. On lines and planes of closest fit to systems of points in space. *Lond. Edinb. Dublin Philos. Mag. J. Sci.* **1901**, *2*, 559–572. [CrossRef]
48. Bowyer, K.W.; Chawla, N.V.; Hall, L.O.; Kegelmeyer, W.P. SMOTE: Synthetic Minority Over-sampling Technique. *CoRR* **2002**, *16*, 321–357. [CrossRef]
49. Han, H.; Wang, W.Y.; Mao, B.H. Borderline-SMOTE: A New over-Sampling Method in Imbalanced Data Sets Learning. In *Proceedings of the 2005 International Conference on Advances in Intelligent Computing—Volume Part I*; Springer: Berlin, Heidelberg, 2005; pp. 878–887. [CrossRef]
50. Nguyen, H.M.; Cooper, E.W.; Kamei, K. Borderline over-sampling for imbalanced data classification. *Int. J. Knowl. Eng. Soft Data Paradig.* **2011**, *3*, 4–21. [CrossRef]
51. MacKay, D.J.C. *Information Theory, Inference and Learning Algorithms*; Cambridge University Press: Cambridge, MA, USA, 2002.
52. Ng, A.Y.; Jordan, M.I.; Weiss, Y. On Spectral Clustering: Analysis and an Algorithm. In *Proceedings of the 14th International Conference on Neural Information Processing Systems: Natural and Synthetic*; MIT Press: Cambridge, MA, USA, 2001; pp. 849–856.
53. Nielsen, F. Hierarchical Clustering. In *Introduction to HPC with MPI for Data Science*; Springer International Publishing: Cham, Switzerland, 2016; pp. 195–211. [CrossRef]
54. Rousseeuw, P.J. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **1987**, *20*, 53–65. [CrossRef]