

Received November 2, 2021, accepted November 20, 2021, date of publication November 23, 2021, date of current version December 6, 2021.

Digital Object Identifier 10.1109/ACCESS.2021.3130400

Driving-Pattern Identification and Event Detection Based on an Unsupervised Learning Framework: Case of a Motorcycle-Riding Simulator

MOHAMED YACINE BOUAOUNI¹, RAYANE AIT ALI YAHIA¹,
AND ABDERRAHMANE BOUBEZOU²

¹Department of Electronics, Ecole Nationale Polytechnique, Algiers 16200, Algeria

²TS2-SATIE-MOSS, Gustave Eiffel University, 77454 Marne la Vallée, France

Corresponding author: Abderrahmane Boubezoul (abderrahmane.boubezoul@univ-eiffel.fr)

ABSTRACT Analysis of human driving behavior aims to inspect drivers' behavior in the real-world and in a virtual environment. The study of driving behaviors can be conducted in naturalistic situations or controlled experiments. Analyzing driving behaviors based on the data collected in naturalistic driving experiments or controlled experiments in the real-world or in a virtual environment is beneficial to fill in many of the knowledge gaps about driving behaviors and risk factors. The amount of data collected during complex experiments with many laps and many drivers tested under different experimental conditions and with different instructions can be huge. Analyzing such data can thus be considered challenging and time-consuming if done manually because it requires calling on experts in traffic psychology to inspect and understand various specific situations at a macroscopic scale involving different riders and at a microscopic scale for a particular rider on a specific lap. Also, it can be challenging in an unsupervised context to detect and match the same patterns in different laps to study similar patterns and spot important and risky events. This paper proposes a multi-step framework for analyzing driving behavior on both the macroscopic and microscopic scales. The core step of this framework is based on unsupervised machine learning algorithms applied to driving-pattern identification and the detection of critical driving events using anomaly-detection algorithms. The detected events are interpreted and described by computing their feature importance using graphs centrality measures. This provides new insight into driving behavior by identifying the motives behind the driver's actions. The present experimental study, based on a dataset collected from the Honda Riding Trainer (HRT) simulator was conducted in the context of the European project SimuSafe and demonstrates the effectiveness of the proposed methodology. These results argue in favor of the development of such methodologies in driving-behavior studies.

INDEX TERMS Time series analysis, time series segmentation, driving-pattern identification, motorcycle simulator, unsupervised learning, anomaly detection.

I. INTRODUCTION

The number of powered two-wheeler (PTW) users is on the rise, especially in cities because PTWs offer a solution to growing traffic congestion and parking problems. However, according to the global status report on road safety conducted by the World Health Organisation (WHO), PTW users are considered among the most vulnerable road users because PTWs offer less protection and stability, which often leads to fatal accidents [1]. Road crashes are complex events that

are highly unpredictable. Various parameters and factors may affect these types of events, e.g., the weather conditions, the geometric shape of the infrastructure, the vehicle's dynamic characteristics, and the driver's behavior. To reduce the number of traffic accidents, one must understand how and why they occur. There are many research tools available for this, such as naturalistic driving studies on open roads, and experimental studies in driving simulators or test tracks. A driving simulator is one such tool that can be seen as a way of tackling this problem. The behavior of real-world drivers can be modeled and introduced into the simulator. Different simulation scenarios can be generated, and the factors leading

The associate editor coordinating the review of this manuscript and approving it for publication was Chuan Zhou.

to a crash may be studied in greater detail. According to a study focusing on the analysis of traffic accidents [2], the personal responsibility of drivers is implicated in the majority of accidents. The driver's emotional, mental, and physical states (frustration, hurry, fatigue, illness, drunkenness, etc.) are some of the factors influencing drivers' decisions when they are interacting with the infrastructure and other road users. Human error, directly or indirectly, plays a central role in motorbike and motorcycle accidents [3], [4]. This may be because the powered two-wheeler users have to actively maintain the dynamic stability of their vehicle.

It is crucial to be able to describe a driver's behavior, both for traffic-collision prevention and traffic-model design. The design of a realistic traffic model could add value to driving simulators, which in turn could lead to the development of more realistic simulation models. The H2020 SimuSafe¹ project is aimed at designing a multi-driver driving simulator and refining Multi-Agent Simulators (MAS) so as to provide realistic interactions between the users (simulated or real) of a virtual road network. The goal is to design the digital twin of a potentially complex road network where simulators controlled by humans and simulated avatars of different road users (pedestrians, cyclists, motorcyclists, and car drivers) act and interact realistically.

Several research studies (see [5] and [6]) have been devoted to studying driving behavior in different contexts and evaluating the causal impact of taking medication on driving behavior. Other studies have focused on detecting riding patterns using supervised algorithms that model temporal dependency like LSTMs [7] and unsupervised algorithms that classify riding patterns using Gaussian Hidden Markov Models [8]. However, the model has been evaluated using supervised learning metrics (e.g. recall and accuracy), which state that even though the approach used was unsupervised, the data were labeled, so the methodology was not fully unsupervised and labeling data can be costly.

While unlabeled datasets are plentiful, labels can require a huge amount of effort by experts, and labeling is time-consuming and expensive. Many research studies focused on unsupervised methods for analyzing driving behavior and detecting dangerous driving situations. Some approaches rely on segmentation of the Inertial Measurement Unit (IMU) time-series data, then applying clustering to group the driving primitives into classes [9]. Another paper addresses this problem by building a high-level understanding of the driving behavior through decomposing the inertial data into linear segments and assigning each segment into a convex optimization of high-level driving behaviors [10]. Recently, an unsupervised approach was proposed in [11]; the proposed methodology is based on a two-step framework. The first step consists of applying Bayesian multivariate linear regression models to segment driving sequences into fragments. The second step applied extended latent Dirichlet allocation models to cluster the fragments into multiple

descriptive driving patterns. In this paper, we propose an end-to-end framework for studying the behavior of riders by better understanding their driving patterns and their interaction with the environment and the other road users (e.g. pedestrian, car) and detecting critical events deemed to be high-risk situations. The framework is fully unsupervised and used for characterizing driving behavior and detecting risky driving situations. Our approach is based on the Toeplitz Inverse Covariance-Based Clustering (TICC) algorithm that performs both segmentation and clustering simultaneously and considers the temporal dependencies. Moreover, we apply the anomaly detection algorithm, Isolation Forest, on the compressed statistics of the distinct driving patterns to detect potential risks (e.g. Abrupt accelerations in high curves). Furthermore, we take advantage of the Markov random field (MRF) matrices computed by the TICC algorithm to enhance the interpretability of these risky driving patterns by computing their feature importance using graph centrality measures applied to the MRF adjacency matrices. This gives a better understanding of the maneuvers that contributed the most to these risky situations.

This framework provides tools for road-safety researchers that help them analyze the driving behavior of powered two-wheeler users or any other road users such as cars and trucks. We tested the framework on sensory data collected from a driving experiment performed on a simulator by eleven drivers with distinct profiles, using various scenarios and instructions. The framework contains tools that we used for macroscopic behavior, which is the overall behavior of a given driver compared to other drivers. This simplifies the task of eliminating subjects that did not follow the instructions or finding subjects with a particular behavior that is worth analyzing in depth during the second stage, the microscopic study. For the second stage, we used a state-of-the-art time-series segmentation algorithm: Toeplitz Inverse Covariance Based Clustering [12] to study the behavior of a single subject, and detect and interpret the patterns in his/her driving style. We have demonstrated how this method outperformed other clustering methods like Hidden Markov Model and Mixture Model. We have also described a technique for matching similar patterns in large-scale datasets, so this automatic method will help find the various situations where the subject made the same manoeuvres.

Event detection is one of the most important parts of analyses of risky situations facing subjects. In this paper, we propose a method for detecting critical events by applying Isolation Forest [13] to the segmentation results of TICC. The critical events detected can be related to abrupt accelerations by taking curvatures with high vehicle dynamics or any other important event, since it is application-dependent, and the user of this framework can choose the features to provide to the anomaly-detection algorithm.

Detecting critical events is not enough to understand the origin of the risk incurred by the riders, thus, the need to inspect the effects of the rider's maneuvers and their importance in that particular situation. This is done in the present

¹<http://simusafe.eu/>

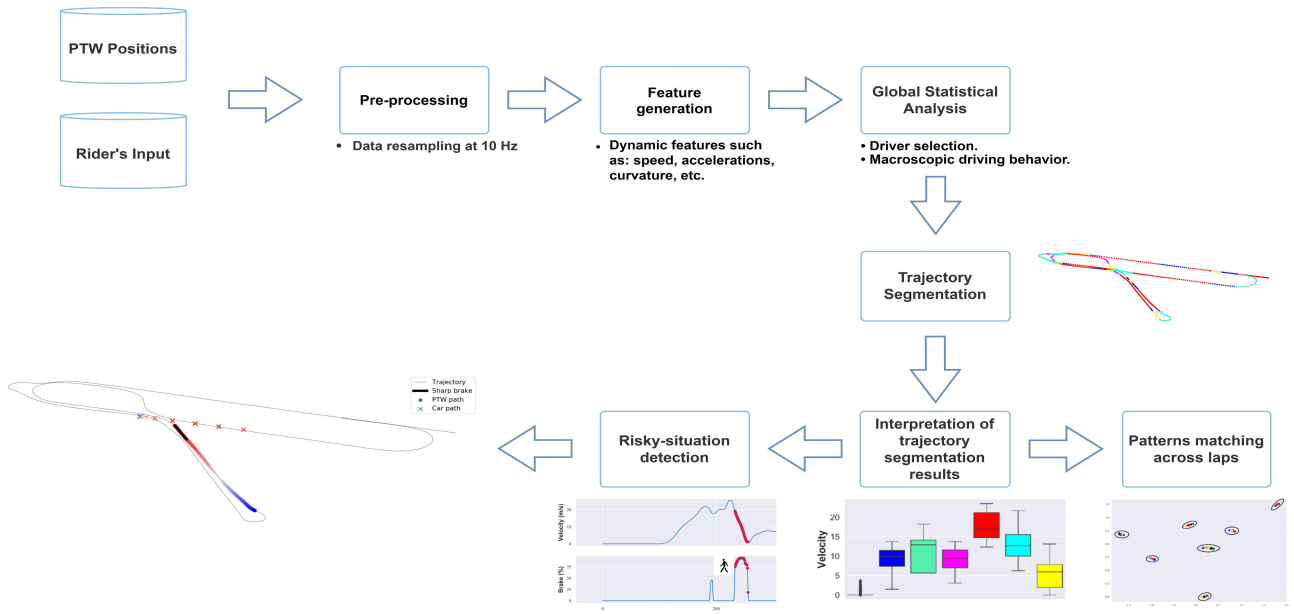


FIGURE 1. Synopsis of the proposed methodology.

paper using graph-centrality measures (e.g. betweenness centrality) applied to a graph of partial correlations between features. This method is used to compute feature importance in relation to a specific maneuver and to use it to interpret the risky situation.

The different steps of the proposed methodology are summarized in Figure 1

The main contributions of this paper are as follows:

- We propose a data-analysis framework that provides tools to road-safety researchers to help them analyze motorcyclists’ behavior or that of other road users such as car or truck drivers.
- We successfully apply the proposed framework to a dataset of sensor data collected during a simulator experimentation involving eleven subjects with different profiles.
- We use state-of-the-art unsupervised segmentation and clustering methods for driving-pattern identification and detection, and for event detection based on anomaly-detection algorithms, to spot critical driving events.
- We interpret the detected events by computing their feature importance using graph-centrality measures, which makes this manuscript a valuable reference for researchers interested in this research topic.

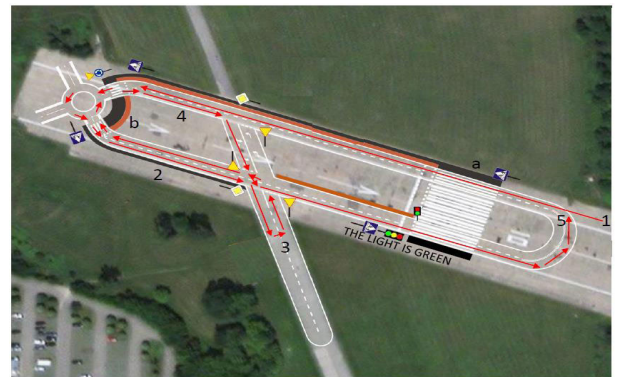


FIGURE 2. Simulation track [8].

TABLE 1. Experimental conditions and instructions.

| LAP | Traffic | Instructions |
|-----|---|--------------|
| 1 | No | No |
| 2 | No | No |
| 3 | No | No |
| 4 | One car and one pedestrian | No |
| 5 | No | Hurry |
| 6 | One car, one pedestrian, traffic cones, and a red light | Hurry |

this kind of situation is observed and his/her maneuvers are analyzed. Figure 2 displays the track.

In the following, let $\mathbf{R} = \{R_0, \dots, R_{10}\}$ denote the set of riders and $\mathbf{L} = \{L_1, \dots, L_6\}$ the set of laps performed by a specific rider. Each lap has particular experimental conditions related to the traffic, level of frustration, and speed instructions. The following table summarizes these scenarios:

TABLE 2. Notation of the features and their units.

| Notation | Description | Unit |
|----------|-------------------------|-----------------------|
| V | Velocity | $m \cdot s^{-1}$ |
| AT | Tangential Acceleration | $m \cdot s^{-2}$ |
| AN | Normal Acceleration | $rad \cdot s^{-2}$ |
| W | Angular Velocity | $rad \cdot s^{-1}$ |
| TH | Throttle Position | % |
| B | Front Brake Position | % |
| H | Handle-Bar Direction | Degree ($^{\circ}$) |
| C | Curvature | m^{-1} |

The raw sensory data are collected with a sampling frequency of 1 Hz for the position of the PTW and 10 Hz for the throttle position, steering-wheel direction, and brake position. We upsampled the spatial data to be able to fuse all variables with the same frequency. We maintained the high sampling frequency for performance reasons because we noted that segmentation algorithms gave better results for data with a high sampling rate, and this also helps avoid convergence errors in the Toeplitz Inverse Covariance-Based Clustering algorithm.

We extended the original data collected from the simulation procedure by computing and extracting new features based on motion physics. These features provided information about the vehicle's dynamics, the rider's behavior, and the infrastructure. These extracted features are velocity, angular velocity, tangential acceleration, and normal acceleration.

The set of features $\mathbf{F} = \{V, AT, AN, W, TH, B, H, C\}$ is used to analyse the macroscopic behavior of the riders across the six laps completed in the next section.

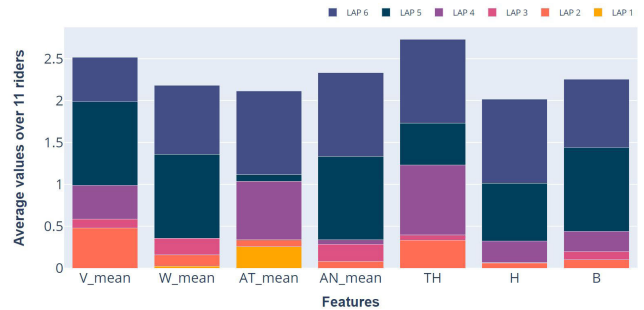
III. PROPOSED APPROACH

This section provides a high-level analysis of the data and the subjects' behavior. The set of features \mathbf{F} are used to study the global behavior across the subjects and the laps to describe the macroscopic behavior of the riders, detect those who did not follow the instructions, and select the riders that did respect to inspect their riding patterns on a microscopic scale. The methodology consists of the following points:

- 1) Compute feature statistics (e.g., max, mean).
- 2) Analyze the global trend across riders and laps.
- 3) Compress the feature statistics using t-SNE and cluster the compressed data into different groups with specific overall riding behaviors.
- 4) Interpret the clustering results and select a rider for the microscopic analysis.

A. DRIVER SELECTION

The experimental protocol defines a set of instructions and constraints that the riders must follow in each LAP, as described in Table 1. The method described here provides a general methodology for approaching comparative studies with different constraints, as in [5], and for obtaining an overview of the overall behavior of riders and spotting both

**FIGURE 3.** Feature statistics averages over the six laps.

the trends across laps and the outliers who did not follow the instructions.

The first step is to represent the data of each rider and each lap by computing the following feature statistics: Maximum and mean velocity, angular velocity, tangential acceleration and normal acceleration, mean use of throttle and front brake, as well as the mean of the absolute value of the handle-bar position and the inverse of the duration of the lap to stay in the same trend as the other statistics, in other words, higher values imply more dynamic behavior. These variables help in analyzing the overall trend in riders' maneuvers and the vehicle dynamics across laps and riders.

Figure 3 plots some of the above feature statistics over the 11 riders, where each color corresponds to a given lap. Note that the fifth and sixth laps have higher values for the vehicles' dynamics and the riders' maneuvers. The fourth lap has high values for throttle use and tangential acceleration, which can be explained by the fact that L_4 has denser traffic than L_5 because a car and a pedestrian are introduced in L_4 . This made the riders use more often the throttle to increase their speed after decreasing it, either to yield the right of way or to avoid the pedestrian. This is consistent with the instructions given in each lap (see Table 1).

Another statistical study had to be performed on each rider. This is the first step in the rider selection process. The main idea is to compute the average, over the six laps, of the mean and/or maximum of each feature, for each driver. This gives an overview of the rider's driving behavior. The results are shown in Figure 4, sorted according to the average of the normalized statistics. Note that some riders have very low vehicle dynamics and maneuvers. These riders are the ones framed in green and can be compared to riders in the red patch. In other words, the riders framed in red exhibit riskier driving behavior than the safe drivers framed in green.

The heat map in Figure 4 also highlights an important result in detecting anomalies. Riders R_{10} and R_8 have a very high maximum speed in comparison to their average speed; we observed the same thing for rider R_2 in terms of angular velocity. These visible results will be discussed further when we use an automatic method to detect the atypical behaviors.

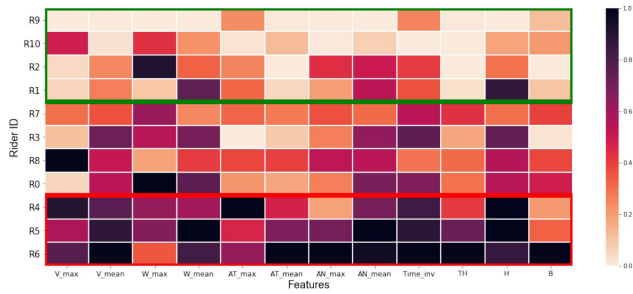


FIGURE 4. Feature statistics by rider.

TABLE 3. List of removed riders.

| Rider | Problem |
|----------|----------------------------------|
| R_1 | High velocity in L_1 and L_2 |
| R_7 | High velocity in L_1 |
| R_{10} | Very Low velocity in L_5 |

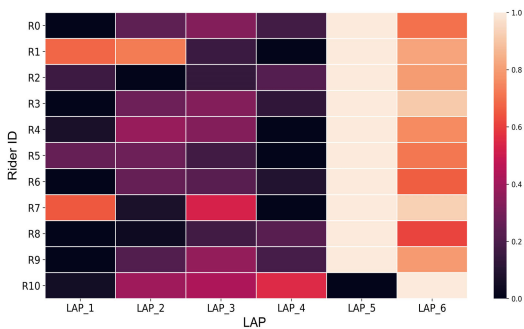


FIGURE 5. Average velocity on each lap for each rider.

This gave us two sets of riders, R_s for safe riders and R_r for potentially risky riders.

$$R_s = \{R_1, R_2, R_9, R_{10}\} \text{ and } R_r = \{R_4, R_5, R_6\}$$

The third step was to visualise the average velocity for each (R_i, L_i) pair to better spot the riders that did not follow the instructions in Table 1.

In general, L_1 had lower velocity than the other laps, because it is considered to be an adaptation phase in which the riders discover the circuit with a very low habituation level. In addition, L_5 and L_6 had the highest velocities because the riders were instructed to drive faster, with some additional constraints in L_6 (traffic cones and a red light) that slightly reduced the values.

However, some riders fell outside the trend, which means that they did not follow the instructions. These riders were removed from the dataset because they could skew the results. Table 3 lists the concerned riders as well as the reason for removing them.

Large datasets require automatic methods to spot anomalies in driving style. Our solution was to use a compression algorithm called t-Distributed Stochastic Neighbour Embedding (t-SNE) [14] followed by the Gaussian mixture mod-

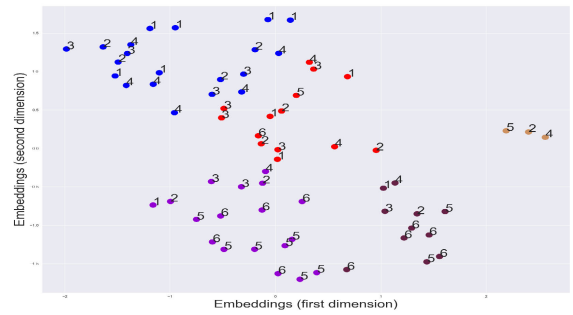


FIGURE 6. Compression obtained by t-SNE followed by clustering of laps using GMM. The numbers represent the laps and the colors are the corresponding clusters.

TABLE 4. The cluster of each (R_i, L_i) pair.

| Rider \ Lap | L_1 | L_2 | L_3 | L_4 | L_5 | L_6 |
|-------------|-------|-------|-------|-------|-------|-------|
| R_{10} | 0 | 0 | 0 | 0 | 4 | 3 |
| R_9 | 0 | 0 | 2 | 0 | 3 | 3 |
| R_8 | 0 | 0 | 0 | 0 | 2 | 2 |
| R_7 | 0 | 4 | 3 | 2 | 1 | 1 |
| R_6 | 0 | 2 | 2 | 0 | 3 | 1 |
| R_5 | 2 | 2 | 2 | 0 | 3 | 1 |
| R_4 | 1 | 1 | 1 | 1 | 1 | 1 |
| R_3 | 2 | 3 | 3 | 3 | 3 | 1 |
| R_2 | 2 | 2 | 2 | 4 | 3 | 3 |
| R_1 | 3 | 3 | 0 | 2 | 3 | 3 |
| R_0 | 0 | 0 | 0 | 0 | 3 | 3 |

els (GMM) clustering algorithm. The t-SNE algorithm is sensitive to its hyperparameter perplexity; its impact on the results is explained in [15]. After the tuning, we obtained good results by fixing the perplexity value at 7.0 and the number of GMM components at 5. Figure 6 and Table 4 present the results obtained by applying this approach. There are five distinct overall behaviors, with one group of laps considered as outliers either because the riders did not follow the instructions or due to a problem in the simulator data.

The box plots in Figure 7 show that the cluster 4 (outliers) has very high velocities and should therefore not be considered for naturalistic riding studies. Concerning the other clusters, they are divided into three categories based on their statistics: low dynamics (cluster 0), average dynamics (cluster 2), and high dynamics (clusters 1 and 3). The difference between the two clusters with high dynamics is the riders' use of both the throttle and the brake. This is quite high in cluster 1 due to the fact that half of the laps in cluster 1 are L_6 , and in L_6 the riders were instructed to hurry, in addition to having a car, a pedestrian, and traffic cones requiring the riders to use the brake. However, we can see that all the laps of rider R_4 have high dynamics. Figure 8 illustrates an example for comparison of this rider to another (R_9) who was classified as a "safe rider". We can see that R_4 used the brake more often. For this reason, we will use this rider in the next section

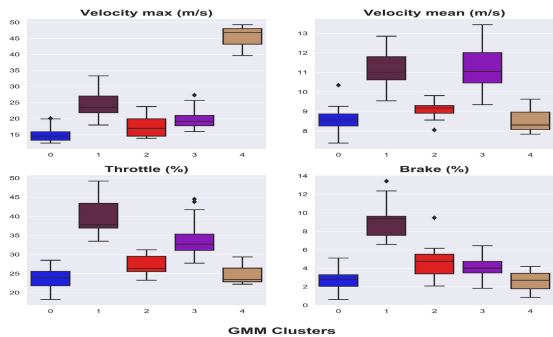


FIGURE 7. Velocity, throttle and brake positions distribution per cluster obtained in Figure 6.

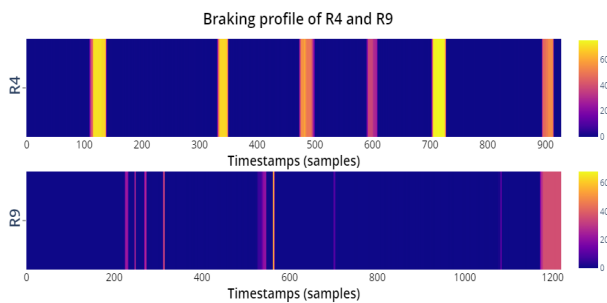


FIGURE 8. Braking profile of riders R_4 and R_9 illustrating the excessive use of the brake by R_4 .

to apply segmentation algorithms and extract the different events that caused this unusual driving style.

B. CLUSTERING AND TIME SERIES SEGMENTATION

1) TOEPLITZ INVERSE COVARIANCE-BASED CLUSTERING (TICC)

Reference [12] is a model-based clustering approach that performs both clustering and segmentation by categorizing time series subsequences into a fixed number of states \mathbf{K} . The states (clusters) represent the repeated patterns in the temporal data and are defined as a correlation network known as a Markov random field (MRF). This network characterizes the interdependencies between the different variables for a specific state across a window of size w . Formally, the learning of each cluster’s MRF is done by estimating a sparse Gaussian inverse covariance (Toeplitz) matrix.

The TICC method takes multivariate time series data as its input, denoted by $\mathbf{x} = [x_1, \dots, x_T]$, where T is the number of samples (observations) and $x_i \in \mathbf{R}^n$, where n is the number of features. The clustering is performed on short sequences of size $w \ll T$, so we refer to the new sequences as $\mathbf{X} = \{X_1, \dots, X_T\}$.

In the following, let $\mathcal{P} = \{P_1, \dots, P_k\}$ denote the point assignments and let $\Theta = \{\Theta_1, \dots, \Theta_k\}$, the Toeplitz matrices, where $\Theta_i \in \mathbf{R}^{nw \times nw}$. The optimization problem is as

follows:

$$\underset{\theta \in \mathcal{T}, \mathcal{P}}{\operatorname{argmin}} \sum_{j=1}^k \underbrace{[\|\lambda \circ \Theta_j\|_1]}_{\text{sparsity}} + \sum_{Y_i \in P_j} \underbrace{(\ell \ell (Y_i, \Theta_j))}_{\text{log likelihood}} + \underbrace{\beta \mathbf{1} \{Y_{i-1} \notin P_j\}}_{\text{temporal consistency}} \tag{1}$$

The algorithm has two regularization parameters: λ and β . Parameter λ controls the sparsity of the MRF matrices in each cluster. It has the same shape as the MRFs ($\lambda \in \mathbf{R}^{nw \times nw}$) but in practice, λ can be reduced to a single value to reduce the search process. The second parameter, β , characterizes the smoothness penalty that ensures temporal consistency and continuity between the adjacent subsequences. Increasing the value of β encourages the neighboring subsequences to belong to the same cluster.

TICC has two other parameters:

- **Window size (w)**, represents the number of observations in a given subsequence $X_t = [x_{t-w+1}, \dots, x_t]$, with $X_t \in \mathbf{R}^{n \times w}$. All the observations that belong to the same subsequence will be assigned to the same cluster.
- **Number of clusters (k)**, which corresponds to the number of patterns that need to be identified. The selection of this parameter can be done using BIC or the silhouette score. However, its value is often application-dependent.

Solving the optimization problem consists of randomly initializing the toeplitz matrices, Θ , and cluster assignments, \mathcal{P} , and then using a variation of the expectation maximization (EM) algorithm that alternates between subsequence assigning (update \mathcal{P}) and updating the clusters’ parameters (toeplitz matrices), Θ .

IV. EXPERIMENTAL METHODOLOGY

In this section, we first do a comparative study of different clustering algorithms based on different metrics and segmentation results. Then, we perform a segmentation of the different laps of a single driver and the results are interpreted using various techniques. Next, we extract the different driving patterns that the segmentation algorithm detected and we compute statistics related to each cluster in order to find important driving events. We also present a new technique for determining feature importance that will be used to interpret the events.

A. MODEL-BASED METHODS

In order to find the different driving patterns, a robust segmentation algorithm must be chosen. A comparison of several model-based and distance-based clustering methods was done in [12] on synthetic data. In our study, we used simulator data to compare TICC and other methods. We did not consider distance-based methods in this comparative study because they do not capture the temporal dependency in the data. Although a distance-based algorithm such as KMeans [16] can obtain good performance in terms of silhouette and Calinski-Harabasz scores, the results of the clustering will not reflect real driving patterns because this algorithm

TABLE 5. Comparison table regarding different metrics. The best results in terms of each metric are shown in bold.

| Method | Metric | | |
|------------|-------------|---------------|-------------|
| | Silhouette | Calinski | Davies |
| TICC (W=1) | 0.45 | 726.02 | 0.99 |
| TICC (W=2) | 0.43 | 609.38 | 1.07 |
| GMM | 0.29 | 382.16 | 1.40 |
| GHMM | 0.31 | 422.38 | 1.37 |

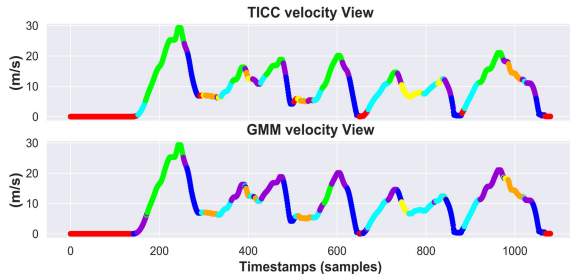


FIGURE 9. Segmentation of L_2 using TICC and GMM.

performs a clustering based on the distances in the data, which minimizes the within-cluster sum of squares and disregards the temporal dependencies.

The segmentation results obtained by three model-based methods (Gaussian of Mixture Model (GMM) [17], Hidden Markov Model with Gaussian mixture emissions (GHMM) [18], and TICC) are presented in Table 5.

The results presented in Table 5 give the performance levels of the different methods in terms of clustering. Toeplitz inverse covariance-based clustering outperforms other methods, and decreasing the window size could result in better scores. However, the Silhouette, Calinski-Harabasz, and Davies-Bouldin metrics are insufficient to comparing the methods on the segmentation task because they only reflect the properties of the different clusters and do not use temporal dependency. The visualization of the segmentation results shows that TICC was better at recovering the different driving patterns.

Figure 9 plots the L_2 -segmentation results for both TICC and GMM. Note that the GMM puts some accelerations and decelerations in the same cluster shown in purple. This is not the case for TICC, which was better at detecting the difference between accelerations and decelerations.

Figure 10 and 11 illustrates the segmentation results of a driving situation in L_2 using TICC and GHMM. The first figure shows that TICC recovered consecutive patterns better than GHMM did. Furthermore, we can see that GHMM did not perform the segmentation well because it assigned individual samples to a particular cluster without considering the temporal dependency of the driving situation. This was not the case for TICC, which has two hyper-parameters: β that encourages adjacent subsequences to belong to the same cluster, and window size, W , which enables TICC to cluster

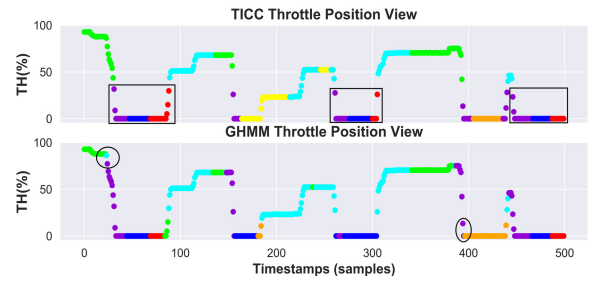


FIGURE 10. Segmentation of L_2 using TICC and GHMM.

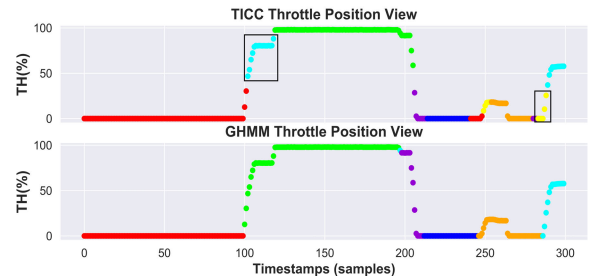


FIGURE 11. Segmentation of L_2 using TICC and GHMM.

a window of observations rather than a set of individual data points.

Another key difference between TICC and GHMM, which can be seen in Figure 11 is that the TICC algorithm detected more accurately the transition phases in the driving situations than GHMM did, as we can see in the framed segments.

The ability of TICC to spot transition phases, to cluster subsequences and to capture temporal dependencies allows it to outperform the other model-based methods presented in Table 5, in terms of both clustering and segmentation. For this reason, this algorithm was used to detect the different driving patterns in the next step of the pipeline.

B. SEGMENTATION RESULTS

In this section, we analyze the drivers' behavior at a microscopic level. We discuss the results of Toeplitz inverse covariance-based clustering and interpret the detected patterns. The parameters of the TICC $\{\lambda, \beta, W\}$ were tuned intensively to achieve good performance.

- Higher values of λ (e.g., 0.1) gave better segmentation results but the inverse covariance matrix was very sparse, so we could not perform centrality measures on the graph described by the inverse-covariance adjacency matrix, which will be described in a later section. Lower values of λ decreased performance and increased training time but provided a less sparse inverse covariance matrix which would help us interpret the results using graph theory tools like betweenness centrality.
- Higher values of β resulted in a model that did not detect some short-time patterns like transition phases

from straight lines to high curvatures or roundabouts. However, small values of β caused the model to catch noise in the data and this affected its ability to generalize. It should be noted that the parameter β is highly dependent on the dataset size; a large dataset requires higher values of β .

- The model was tested for $w \in \{1, 2, 5, 10, 15\}$. TICC proved robust with regard to its hyper-parameter W . However, a window size of 1 resulted in an MRF matrix of shape 7×7 , which made the task of interpreting the MRF using betweenness centrality impossible. For higher values of W , some transition phases were neglected. Thus, the best model we obtained in terms of segmentation and MRF interpretability was when $w=2$.
- K , the number of clusters, is a very important parameter to consider. In the case of an unsupervised approach, the selection of the value of this parameter can be done using the Bayesian Information Criterion (BIC) or the silhouette score [19], but these two metrics can lead to a value of K that is not suitable for the targeted application. Therefore, we determined the value of K by testing different values in the set $K \in \{5,6,7,8\}$. Higher values gave better scores but complicated the interpretation task. However, lower values caused the model to neglect the transition states and to treat different patterns as the same. A good trade-off between performance and interpretability was attained for $K=7$.








Figure 12 shows the results of the L_3 segmentation performed for R_6 .

We used the box plots in Figure 12 to interpret the different clusters in terms of the vehicle's dynamics and the rider's behavior. The TICC algorithm detected the smooth and abrupt accelerations and decelerations, very low dynamic states corresponding to situations where the rider had to stop or yield and sharp curves such as in roundabouts and on U-turns. Table 6 summarizes the different patterns and their respective colours.

The results also showed that the TICC algorithm could accurately detect sequential patterns, as seen in Figure 13.

- The first sequential pattern (blue-red) represents full acceleration, with the blue cluster representing a transition state from a low dynamic to an abrupt acceleration and a high velocity (red cluster).
- The second sequential pattern (cyan-yellow-purple) corresponds to the situation where the rider releases the throttle to start a smooth deceleration (cyan), and then presses the brake (yellow) to finally reach a state of low dynamics as in stopping or yielding.
- The third sequential pattern (purple-blue) illustrates situations in which the rider is in a state of low dynamics and then starts to press on the throttle smoothly. This kind of pattern can also be found after a sharp curve, where the riders start to accelerate along a straight line.

TABLE 6. Riding behavior by type of cluster.

| Color | Riding Behavior |
|---|--|
|  | Transition state from low dynamics to abrupt acceleration. |
|  | Abrupt acceleration with high velocity. |
|  | Average curvatures with high dynamics (velocity and throttle use). |
|  | Very low dynamics with no use of the brake or the throttle. |
|  | Abrupt braking. |
|  | Sharp curves as in roundabouts and U-turns |
|  | Transition state: smooth deceleration via release of the throttle. |

- The last sequential pattern (red-cyan) represents a transition from rapid acceleration to speed reduction by release of the throttle.

C. PATTERN MATCHING

The task of finding similar patterns in different laps, or datasets can be very difficult and time-consuming if done manually. In this part, we propose a method for finding the like patterns in different laps of rider R_6 . The method consists of computing the statistics (maximum, minimum, mean, median, interquartile range) for the set of features $\{V, AN, AT, W, TH, B\}$. All of the statistics are expressed on a scale between 0 and 1 using the formula:

$$s_{c,l} = \frac{s_{c,l} - \max(s)}{\max(s) - \min(s)}$$

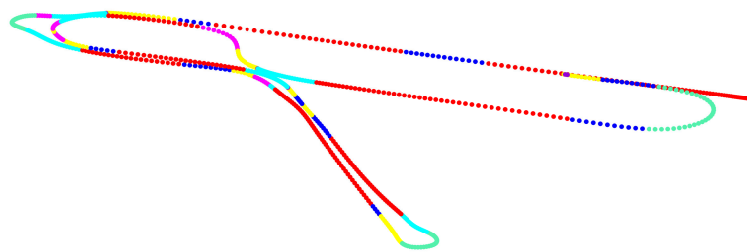
where $s \in S$, the set of all feature statistics.

c : Cluster number with $c \in \{0, 1, 2, 3, 4, 5, 6\}$.

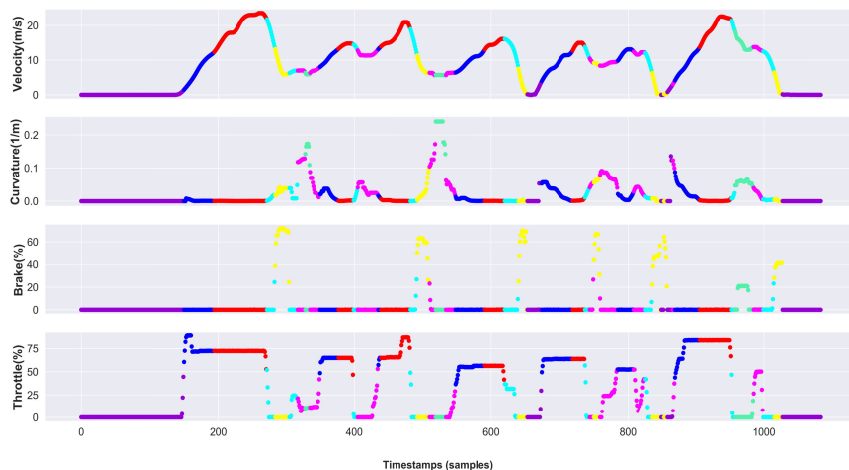
l : Lap number $l \in \{2, 3, 4, 5, 6\}$.

The statistical data was fed into a compression algorithm. We then did a comparative study of two of the widely used dimensionality-reduction techniques: t-SNE [14] and UMAP [20]. UMAP yielded better results and that can be explained by the fact that UMAP is better at retaining the overall structure of the data than t-SNE is, and it is less sensitive to its hyperparameter, the number of neighbors, than t-SNE is to perplexity.

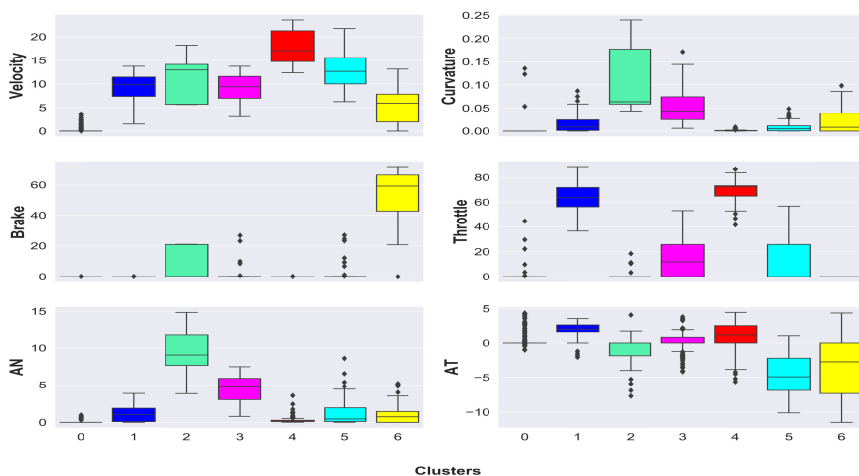
These results show that the TICC algorithm is robust to changes in the data. In other words, the new situations in L_5 and L_6 and changes in the experimentation conditions did not affect the ability of TICC algorithm to find like riding patterns that are similar (e.g., abrupt acceleration).



(a) Trajectory view.



(b) Feature profile.



(c) Box plot view.

FIGURE 12. Segmentation of (R_6, L_3) obtained by the TICC method in different views: (a) trajectory view, (b) feature profile and (c) box plot view, where the colors represent cluster assignment by the TICC algorithm.

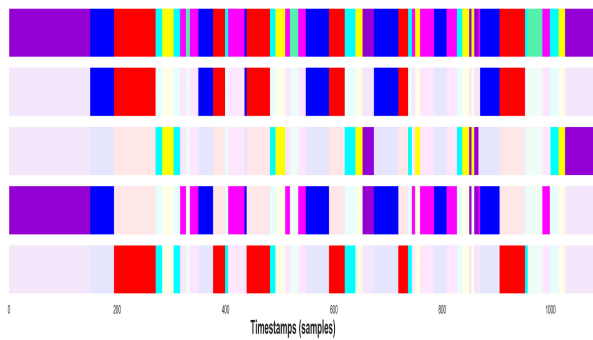


FIGURE 13. Sequential patterns detected by the TICC algorithm.

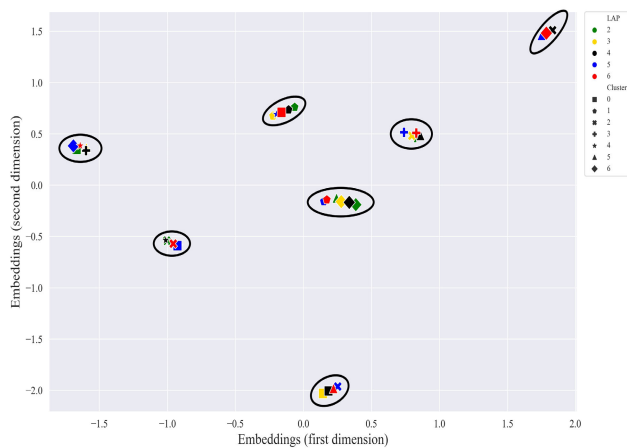


FIGURE 14. Embeddings of feature statistics obtained by UMAP.

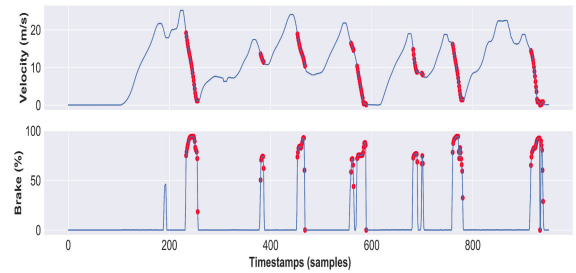
D. RISKY-EVENT DETECTION

One of the most important tasks in driving research is event detection. In this paper, we propose an application-dependent method to spot events using an anomaly-detection algorithm.

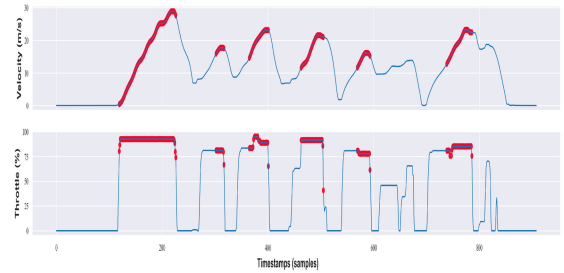
The first step is to choose the features that are the most closely related to the event. For example, to detect an abrupt braking event, the most relevant choice is to choose acceleration and brake position. Then, after selecting the appropriate features, we computed their statistics as described in the previous section and we fed the statistical data into the Isolation Forest algorithm, with a number of estimators equal to 55 and a contamination value of 0.1. The tuning of the contamination hyperparameter is very important for decreasing or increasing the number of detected events. Below we provide two examples of using this method for detecting abrupt acceleration and braking events.

1) SHARP ACCELERATION AND BRAKING DETECTION

In this example, we want to detect events where the rider accelerates in an unusual way. The features that best describe such a riding behavior are the throttle and the brake positions. We apply the Isolation Forest algorithm [13] on the max, mean, and median statistics to visualize the detected events.



(a) Abrupt braking events in L_6 .



(b) Abrupt acceleration events in L_5 .

FIGURE 15. Abrupt acceleration and braking events detected using the Isolation Forest algorithm.

In the present example, the Isolation Forest algorithm detected four events that belong to $\{L_4, L_5, L_6\}$. This is in line with the different scenarios and instructions for these laps, which introduce some elements that increase the frustration level of the rider (pedestrian, car) and prompt the riders to go faster.

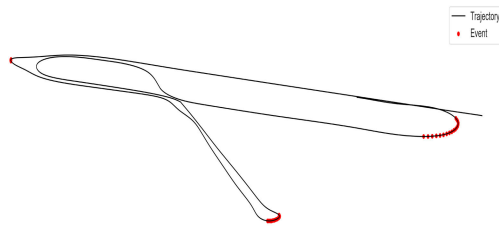
2) SHARP CURVES

Driving in some parts of the roads, like roundabouts and U-turns with inappropriate vehicle dynamics can be dangerous and considered risky. This example illustrates the detected events related to high curvatures with unsuitable vehicle dynamics. The most relevant features to consider are AN, W, H. The same steps were performed as the first example, while changing in the contamination parameter to 0.07 to reduce the number of events. Note that, as in abrupt accelerations, events were detected in L_5, L_6 which means that the rider followed the experimental instructions by going faster on these two laps. Figure 16 illustrates the detected event on the trajectory.

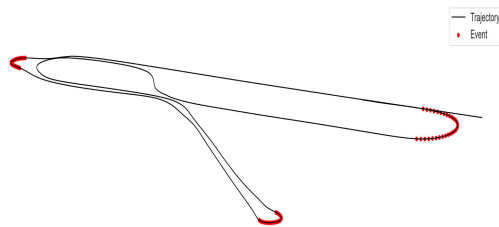
E. FEATURE CONTRIBUTION TO EACH DRIVING PATTERN

Another advantage of the TICC algorithm is its interpretability. The MRF of each cluster can be used to interpret the results using partial correlations between variables. We exploit this property by computing the following measures of centrality on the graph associated with the MRF.

- Betweenness centrality (BC) [21]. This is a measure of centrality in a graph based on the shortest path. For each vertex, the number of shortest paths between each



(a) Sharp curvature events in L_5 .



(b) Sharp curvature events in L_6 .

FIGURE 16. Sharp curvature events detected using the isolation forest algorithm.

of the two vertices that cross the selected vertex is the betweenness centrality measure. In our application, the vertices represent a feature at a certain position in the window W of the TICC. A high betweenness centrality indicates that the corresponding feature has more impact on the riding behavior of that particular cluster.

- Degree centrality (DC) [22]. This is defined as the number of links incident upon a node. This means that a node with a high degree centrality is regarded as a central node, and in our study, a more important feature.
- PageRank (PR) [23]. This algorithm is used by Google search to rank website pages and measure their importance. It relies on counting both the number and quality of the links to a page or node.

Before applying any one of these three algorithms, a threshold must be set to transform the graph into an unweighted graph. Edges that have weights below the threshold will be omitted and the ones that have values above the threshold will be considered as unweighted links. The threshold depends on the distribution of the MRF values and the algorithm. Choosing a very low threshold means that conditionally independent features that have a low partial correlation will be considered the same as the features with a very high partial correlation. This is due to the fact that all values that are above the threshold are considered in the same way. This parameter, then, must be carefully selected.

1) FEATURE IMPORTANCE FOR ABRUPT ACCELERATION

We used all three algorithms to quantify the importance of each feature in the event, abrupt acceleration in L_5 , shown in

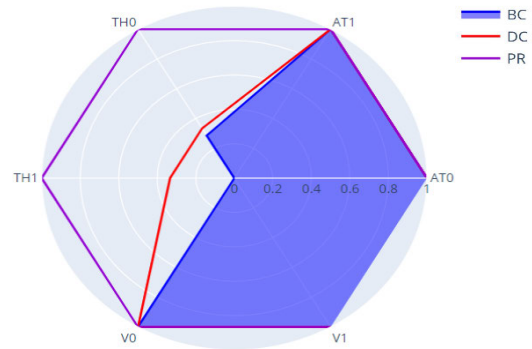


FIGURE 17. Feature importance for abrupt acceleration events using betweenness centrality, degree centrality, and PageRank.

Figure 17. The thresholds applied to the adjacency matrices used to compute betweenness centrality, degree centrality, and PageRank are 10^{-5} , 0.5, and 1, respectively. Figure 17 illustrates the results obtained by applying the three algorithms. Notably, the features with high centrality values are velocity, acceleration, and throttle position, which is consistent with the type of event: abrupt acceleration along straight lines.

F. DETECTION OF INTERESTING PATTERNS IN DRIVING BEHAVIOR

This section explains the link between the detected events and real-world situations. Riding a motorbike is a very complex task compared to that of four-wheeled vehicles. This complexity comes from the fact that the rider is intensely involved in the riding task to maintain her/his vehicle's dynamic stability. This task is even more complex when the rider interacts with other road users such as pedestrians or cars. We consider Lap 6 because of the different constraints (pedestrian, car, etc.) and most of the events we detected occurred on this lap. Figure 18 presents a situation where the PTW has a yield sign and a passing car. The Isolation Forest algorithm detected a rapid braking event, shown in black. We concluded that the methodology we propose provides the necessary tools to detect events related to critical driving situations.

The second situation consists of a pedestrian crossing the road, as illustrated in Figure 19. The anomaly-detection algorithm did not detect any events near the pedestrian. Studying the rider's behavior in such situations is thus a difficult task. Another aspect of this difficulty is related to the rider's anticipation of facing an unrecognized event during her/his journey, such as a pedestrian who is crossing. Capturing such anticipated events based only upon the data from the rider's actions and the PTW dynamics can be considered impossible. Note that this event would not be detected because the rider decelerated and did not brake abruptly because he/she anticipated the pedestrian's behavior. The same figure shows abrupt braking close to the traffic cones.

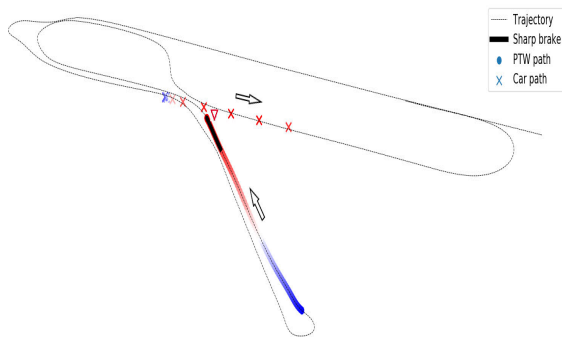
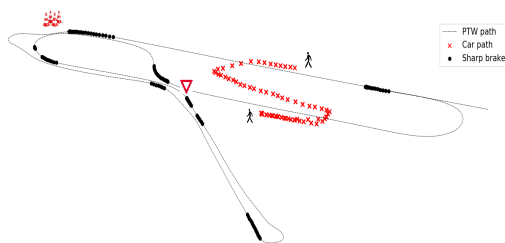
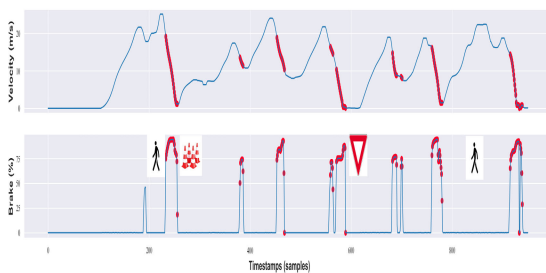


FIGURE 18. Situation 1: Rider has a yield sign and the car is passing.



(a) Pedestrian crossing the street.



(b) Velocity and brake profile.

FIGURE 19. Situation 2: Pedestrian crossing the street.

V. DISCUSSION AND CONCLUSION

In this paper, we proposed a set of techniques for analyzing the driving behavior of different riders at the macroscopic and microscopic levels using multivariate time-series data. The macroscopic analysis identified the riders that did not follow the instructions or displayed a particular driving behavior that needs to be analyzed at the microscopic level. To this end, an unsupervised approach was chosen to perform a clustering of the subsequences using Toeplitz Inverse Covariance-based Clustering (TICC) to extract and identify the different riding patterns. TICC algorithm proved promising because it outperformed both Gaussian Mixture Models (GMM) and Gaussian Hidden Markov Models (GHMM). We also proposed a technique for finding matching patterns across different laps using a dimensionality-reduction algorithm called Uniform Mani-

fold Approximation and Projection (UMAP). It demonstrated that TICC is robust and can detect like patterns in different laps.

The segmentation results are interpreted first using box plots to explain the driving behavior in the clusters, and then an event detection step was performed using an anomaly-detection algorithm called Isolation Forest to detect important events such as rapid acceleration, abrupt braking, and highly curved trajectories. We compute the feature importance for these events by applying the centrality algorithms betweenness centrality, degree centrality and PageRank on the inverse-covariance matrices to explain the relationship between each event and the features. This step requires carefully choosing the thresholds, which we did manually.

1) PRACTICAL APPLICATIONS

The framework proposed in this paper provides various techniques for analyzing large-scale datasets in the context of naturalistic driving in real or simulated environments and can therefore help researchers in the transportation domain to study driving behavior and the different patterns related to drivers interaction with other road users (e.g., pedestrians, cars). In addition, driving schools can benefit from this framework by showing trainees their trajectories and identifying their driving patterns based on the segmentation results. The event-detection techniques can be used to spot the high-risk situations and based on centrality measures, analyze the factors that contributed the most to making these situations risky.

The main objective of this study is to investigate the behavior during risky situations, which results from the behavior discrepancy between the rider's actions, vehicle dynamic, and the infrastructure. The riding behavior may differ for each subject according to different factors: riders' experience, interaction with infrastructure, riders' emotional state (frustration), and environmental factors (traffic). Consequently, road events detection in a general way may be a very arduous task because of the ambiguity around the definition of a driving event. Furthermore, this task can be challenging in the case of a powered-two-wheeler because of the driver's strong involvement in the driving task. Thus his experience could play a significant role in driving and on the perception of risk. To tackle that challenge, we propose a multi-step framework where, in the matching step, the algorithm was able to find the same patterns across different laps, knowing that each lap has its experimental conditions and instructions. Therefore, we could say that the TICC algorithm is robust to changes and can detect the same patterns. Moreover, the algorithm extracts high-level driving behavior from sensory data. These patterns are the most common in Naturalistic driving (going straight, sharp curves, etc.). The grouping step means that the clusters reflect the same driver's behavior, for example, accelerating/decelerating in the same way on the straight line or taking a curve sharply. This step also helps to distinguish the patterns presenting some particular behaviors.

By learning and identifying riding patterns, useful contextual information may be provided to intelligent transportation systems developed for PTWs, improving their effectiveness and riders' safety. The human riding behavior is a complex concept, and its characterization may lead to a better understanding of the rider's decisions when encountering different situations. This characterization will allow us to prevent collisions and design the riding models, which is one of the core algorithms that might make the future of self-riding motorbikes possible. Autonomous vehicles have to interact with other vehicles (even non-autonomous ones), and understanding their driving style can provide valuable information to avoid traffic collisions [24].

2) LIMITATIONS AND PERSPECTIVES

The proposed unsupervised framework provides an effective and efficient data mining tool to help researchers with a deep and comprehensive understanding of drivers' behavioral characteristics. The obtained results are promising, but if we want to continue along these lines, some parts of that methodology can further be improved in our future work. At first, the a priori setting of the number of patterns is the main limitation of the clustering algorithm (TICC) used in this framework. TICC can detect repeated behavior by assigning similar segments of data to the same pattern. However, the number of patterns is fixed and must be determined by the user, meaning that the number of different behaviors present in the data should be known a priori. Additionally, TICC is not directly suitable for streaming data, as it assumes that all data is available simultaneously. Secondly, the interpretation step highly depends on the threshold values that we set for the computation of the betweenness centrality, degree centrality, and page rank criteria. Using an automatic method to compute the threshold based on the distribution of the weights of the MRF graph would lead to better results and make the framework less dependent on this parameter. Moreover, the patterns matching step requires domain knowledge to choose the features that best describe the atypical behavior (anomaly) that the user of this framework is studying. That issue may also be considered an advantage because the user will extract the anomalies that depend on his application (e. g., abrupt braking). In contrast, he can also get a global atypical behavior by feeding all the features to the UMAP algorithm.

ACKNOWLEDGMENT

The authors would like to thank APTIV-POLAND Team for their contribution to the experimentation. They would also like to thank researchers of the Unit of Research in Traffic Psychology UCSC for the design of experimental protocol. They would like to express their deepest gratitude to the reviewers for their constructive comments and time spent to review this article.

REFERENCES

- [1] (2018). *Global Status Report on Road Safety 2018*. [Online]. Available: https://www.who.int/violence_injury_prevention/road_safety_status/2018/en/
- [2] L. Martín, L. Baena, L. Garach, G. López, and J. de Oña, "Using data mining techniques to road safety improvement in Spanish roads," *Proc. Social Behav. Sci.*, vol. 160, pp. 607–614, Dec. 2014.
- [3] J. R. Treat, "Tri-level study of the causes of traffic accidents: An overview of final results," in *Proc. Amer. Assoc. Automot. Med. Annu. Conf.*, vol. 21. Detroit, MI, USA: Association for the Advancement of Automotive Medicine, 1977, pp. 391–403.
- [4] L. Mallia, L. Lazuras, C. Violani, and F. Lucidi, "Crash risk and aberrant driving behaviors among bus drivers: The role of personality and attitudes towards traffic safety," *Accident Anal. Prevention*, vol. 79, pp. 145–151, Jun. 2015.
- [5] Q. Li, L. Zhao, Y.-C. Lee, Y. Ye, J. Lin, and L. Wu, "Contrast feature dependency pattern mining for controlled experiments with application to driving behavior," in *Proc. IEEE Int. Conf. Data Mining (ICDM)*, Nov. 2019, pp. 1192–1197.
- [6] Q. Li, L. Zhao, Y.-C. Lee, and J. Lin, "Contrast pattern mining in paired multivariate time series of a controlled driving behavior experiment," *ACM Trans. Spatial Algorithms Syst.*, vol. 6, no. 4, pp. 1–28, Aug. 2020.
- [7] M. Leyli-Abadi, A. Boubezoul, and L. Oukhellou, "Riding pattern recognition for powered two-wheelers using a long short-term memory network," in *Proc. IEEE 23rd Int. Conf. Intell. Transp. Syst. (ITSC)*, Sep. 2020, pp. 1–7.
- [8] M. L. Abadi, A. Boubezoul, and S. Espié, "Non-supervised trajectory segmentation and cross analysis of riders' dynamic behavior in a simulated riding platform," in *Proc. IEEE Int. Smart Cities Conf. (ISC2)*, Sep. 2020, pp. 1–8.
- [9] G. Agamennoni, S. Worrall, J. R. Ward, and E. M. Nebot, "Automated extraction of driver behaviour primitives using Bayesian agglomerative sequence segmentation," in *Proc. 17th Int. IEEE Conf. Intell. Transp. Syst. (ITSC)*, Oct. 2014, pp. 1449–1455.
- [10] A. Bender, G. Agamennoni, J. R. Ward, S. Worrall, and E. M. Nebot, "An unsupervised approach for inferring driver behavior from naturalistic driving data," *IEEE Trans. Intell. Transp. Syst.*, vol. 16, no. 6, pp. 3325–3336, Dec. 2015.
- [11] G. Li, Y. Chen, D. Cao, X. Qu, B. Cheng, and K. Li, "Extraction of descriptive driving patterns from driving data using unsupervised algorithms," *Mech. Syst. Signal Process.*, vol. 156, Jul. 2021, Art. no. 107589.
- [12] D. Hallac, S. Vare, S. Boyd, and J. Leskovec, "Toeplitz inverse covariance-based clustering of multivariate time series data," in *Proc. 23rd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2017, pp. 215–223.
- [13] F. T. Liu, K. M. Ting, and Z.-H. Zhou, "Isolation-based anomaly detection," *ACM Trans. Knowl. Discovery Data*, vol. 6, no. 1, p. 39, 2012.
- [14] L. Van der Maaten and G. Hinton, "Visualizing data using t-SNE," *J. Mach. Learn. Res.*, vol. 9, pp. 2579–2605, Nov. 2008.
- [15] M. Wattenberg, F. Viégas, and I. Johnson, "How to use t-SNE effectively," *Distill*, vol. 1, no. 10, p. e2, Oct. 2016.
- [16] J. A. Hartigan and M. A. Wong, "A K-means clustering algorithm," *Appl. Stat.*, vol. 28, no. 1, pp. 100–108, 1979.
- [17] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.
- [18] L. Rabiner, "A tutorial on hidden Markov models and selected applications in speech recognition," *Proc. IEEE*, vol. 77, no. 2, pp. 257–286, Feb. 1989.
- [19] P. J. Rousseeuw, "Silhouettes: A graphical aid to the interpretation and validation of cluster analysis," *J. Comput. Appl. Math.*, vol. 20, no. 1, pp. 53–65, Nov. 1987.
- [20] L. McInnes, J. Healy, and J. Melville, "UMAP: Uniform manifold approximation and projection for dimension reduction," Feb. 2018, *arXiv:1802.03426*.
- [21] U. Brandes, "A faster algorithm for betweenness centrality," *J. Math. Sociol.*, vol. 25, no. 2, pp. 163–177, 2001.
- [22] D. Sharma and A. Surolia, *Degree Centrality*. New York, NY, USA: Springer, 2013, p. 558.
- [23] L. Page, S. Brin, R. Motwani, and T. Winograd, "The pagerank citation ranking: Bringing order to the web," Stanford InfoLab, Stanford, CA, USA, Tech. Rep. 1999-66, Nov. 1999.
- [24] A. Ahadi-Sarkani and S. Elmalaki, "ADAS-RL: Adaptive vector scaling reinforcement learning for human-in-the-loop lane departure warning," in *Proc. 1st Int. Workshop Cyber-Phys.-Hum. Syst. Design Implement. (CPHS)*. New York, NY, USA: Association for Computing Machinery, May 2021, pp. 13–18.



MOHAMED YACINE BOUAOUNI received the Engineering degree in electronics and the master’s degree in signal and communication from the National Polytechnic School, Algiers, Algeria, in 2021, where he supported his final thesis on “Hybrid Deep Learning-Based Approach Applied on Speech Signal Separation” at the LDCCP Laboratory. He is currently pursuing the master’s degree in machine learning and data science with Paris-Saclay University, France, where he was awarded an IDEX Excellence Scholarship. His current research interests include pattern recognition, anomaly detection, machine learning, and source separation.



RAYANE AIT ALI YAHIA was born in Algiers, Algeria, in February 1999. He received the Engineering and master’s degrees in signal and communication from the National Polytechnic School of Algiers, Algeria, in 2021, through supporting his final year project in the LDCCP Laboratory based on Algiers titled “Hybrid Deep Learning-Based Approach Applied on Speech Signal Separation.” He is currently pursuing the master’s degree in data science with the Université de Versailles, Paris, France. His research interests include the area of source separation and anomaly detection.



ABDERRAHMANE BOUBEZOUL received the Ph.D. degree in computer science and mathematics from University Paul Cézanne (Aix-Marseille III), France, in 2008, and the master’s degree in virtual reality and complex systems from Evry Val d’Essonne University, France. Since 2008, he has been a Researcher with the IFSTTAR Institute and currently with Gustave Eiffel University. He has authored or coauthored over 50 papers in scientific journals and conference proceedings. His current research interests include statistical signal processing and machine learning applied to road transport systems.

• • •