Behavioral Profiling of Vehicles in 5G-VANETs Using Anomaly Detection and Unsupervised Learning

Caterina Angelica Bergonzoni, Costantino D'Ambrosio, Gabriele Di Cesare

1. Introduction to the problem

The aim of the project is to profile the behavior of vehicles and detect anomalies in 5G-VANET networks, using unsupervised learning techniques and then consequently compare them to predefined labels that identify malicious behavior.
5G-VANET (5G Vehicular Ad-hoc Networks) are an evolution of traditional vehicular networks due to their low latency, high throughput and ability to support critical applications such as assisted driving, vehicle cooperation, traffic management and security services. Vehicles in VANETs are nodes in a moving network, whose configuration changes continuously as positions shift. These nodes exchange important information, such as vehicle speed, location, or traffic conditions. They allow vehicles to successfully communicate both with each other and with road infrastructure. This dynamic nature, however, also introduces new critical issues, particularly with regard to the security and reliability of communications. Anomalous or malicious behavior can compromise the integrity and functionality of the network, necessitating efficient anomaly detection strategies to mitigate such threats. Anomalies in VANETs manifest themselves as irregular patterns in communications, vehicle behavior, or interactions with network infrastructure. Advanced methodologies are required for their real-time detection, and the application of machine learning approaches.
The project therefore addresses the problem of characterizing vehicle behavior and identifying anomalies within a 5G-VANET network. The goal is to identify "normal" operating patterns and distinguish suspicious situations (e.g. related to attacks or out-of-distribution behavior) without necessarily depending on predefined labels, which in practice may be incomplete, expensive to obtain or unavailable.

1.1 Input Files

The analysis is based on a dataset structured at the level of time records per vehicle (each row represents the status/activity of a vehicle at a given point in time). For the construction of the dataset, a scan was carried out repeated 10 times, at regular intervals of 20 seconds, on a specific urban area, in order to sample the temporal evolution of the state of the vehicles. The dataset includes mobility variables, network context and performance indicators, as well as fields that can be used for validation. The features considered are:

- Vehicle_ID: Unique vehicle identifier.
- Vehicle_Type: type of vehicle (category or class, useful for distinguishing different mobility/consumption profiles).
- Timestamp: time reference of the record, used for dynamic analysis and trends over time.

- Latitude, Longitude: geographical coordinates of the vehicle, useful for describing mobility and spatial distribution.
- Speed_kmh: Vehicle speed (km/h), direct display of driving behaviour and movement dynamics.
- Direction_deg: direction of travel in degrees, useful for estimating trajectories and sudden changes.
- Congestion_Level: level of congestion (traffic condition or local load), which affects both mobility and communication performance.
- Vehicle_Density: vehicular density in the area, context indicator that can impact channel contention and interference.
- Energy_Used: Energy consumed in the period/record (or consumption associated with network and motion activities).
- Remaining_Energy: residual energy of the vehicle, useful for assessing sustainability and energy behavior over time.
- Delay_ms: Communication delay (latency) in milliseconds.
- Throughput_Mbps: Communication throughput (Mbps), a measure of actual transmission capacity.
- Packet_Delivery_Ratio: Packet delivery ratio (PDR), synthetic indicator of communication reliability.

There are also information fields related to the presence of malicious activity:

- Is_Malicious: A label that indicates whether the record is associated with malicious (binary) behavior.
- Attack_Type: type of attack, when present (categorical).

Network-related anomalies: These include **DoS**, **Wormhole**, **Black Hole** attacks, each designed to breach network integrity through communication.
In this work, these fields are mainly considered for validation and ex-post comparison, while the identification of anomalous behavior is set up as an unsupervised problem, more adherent to real scenarios in which labels are not always available or reliable.

2. State of the art

In 5G-VANET networks, vehicles communicate with each other and with infrastructure as they move; therefore, the network is constantly changing, making it more difficult to maintain stable and fast communications. To better manage this dynamism, many works in the literature use clustering, i.e. grouping vehicles into "groups" guided by a central node. Clustering methods are often based on mobility parameters (distance, speed, direction), but increasingly they also integrate "smart" techniques such as machine learning and fuzzy logic to make clusters more stable and reduce packet loss and delays.

Another much-discussed topic is the integration between VANET and modern technologies related to 5G and strategies to improve throughput and latency. In this vein, it is highlighted that cluster-based solutions can help to better manage resources and quality of service, especially when the density of vehicles changes rapidly.
On the security front, the literature highlights that VANETs are exposed to various anomalies and attacks. These are seen as irregular patterns in vehicle communication or behavior (e.g., delayed messages, manipulated data, inconsistent trajectories/speeds). For this reason, in recent years the use of anomaly detection techniques based on machine learning and deep learning has become widespread, with supervised, unsupervised and hybrid approaches, and with a focus on robust solutions that can be used in real time.
In this project, we are moving precisely in this direction: we use a dataset that combines mobility and traffic variables, energy and communication performance and apply unsupervised learning methods and anomaly detection to describe typical vehicle behaviors and identify suspicious deviations, using safety labels especially for the validation phase of the results.

3. Feature Engineering

To help algorithms in vehicle profiling, the following features have been added:

- Speed_Density_Ratio = Speed_kmh / Vehicle_Density
- Congestion_Adjusted_Speed= Speed_kmh / (Congestion_Level + 1)
- Energy_Communication_Index = Energy_Used / Throughput_Mbps
- Behavioral_Stress_Index = (Energy_Used * Delay_ms) / Packet_Delivery_Ratio
- Energy_per_Speed = Energy_Used / Speed_kmh
- Energy_Ratio = Energy_Used / Remaining_Energy
- Throughput_per_Delay = Throughput_Mbps / Delay_ms
- Communication_Stability = Packet_Delivery_Ratio / Delay_ms

4. Data Pre-processing

- **Removing irrelevant columns:** Deleting unnecessary columns;
- **Label encoding:** Categorical variables (such as vehicle type and attachment) are converted into numeric values.

5. Results

5.1 Unsupervised Method: ISOLATION FOREST

In this phase of the trial, the **Isolation Forest** algorithm was applied directly to the records of individual network packets, without per-vehicle aggregation. This approach represents an attempt at **real-time** detection, where the system must instantly decide whether a single packet is malicious or not.

Tabella 1 Isolation Forest Classification Report

	precision	*recall*	*f1-score*	*support*
0	0.98	0.90	0.94	48966
1	0.02	**0.09**	0.03	1034

accuracy			0.88	50000
macro avg	0.50	0.50	0.48	50000
weighted avg	0.96	0.88	0.92	50000

Interpretation of Results

The report obtained on the 50,000 raw records highlights the algorithm's inability to distinguish anomalies in a high-frequency data stream without behavioral pre-processing.

- **Accuracy to 0.02:** This value indicates that 98% of alarm reports are actually false positives. In VANET networks, an error of this magnitude would lead to grid paralysis, as almost all legitimate vehicles would be revoked.
- **Recall at 0.09:** Not only does the model almost always get it wrong when reporting an attack, but it also misses 91% of real attacks.
- **The diagnosis:** The problem is not the Isolation Forest algorithm, but the statistical overlap. On individual packets, an attacker's *Delay* or *Energy* values are indistinguishable from a traffic spike or natural radio interference.

5.2. Second method with ISOLATION FOREST

The method is a two-stage approach that transforms an instantaneous (point-by-point) detection into a **behavioral (vehicle-based)** detection. It's a critical strategy in VANET networks because a single "strange" packet might just be a signal problem, but a series of anomalies indicates an attack.

1. The First Stage: Isolation Forest

Initially, the **Isolation Forest algorithm** analyzes each individual record and assigns a score: 1 if it is considered normal, -1 if it is considered an anomaly.

2. The Second Stage: Aggregation by vehicle

The time window **of each vehicle (consisting of 10 records/detections)** is analyzed:

- **Healthy:** If a vehicle sends 10 packets and the Isolation Forest says they are all normal, the sum will be 10. It is marked as healthy (0).
- **Suspicion:** If even one of those 10 packets is marked as abnormal (score -1), the sum will be <10. He is marked as mischievous (1).
- **Comparison:** Compare the values assigned after aggregation with the actual values.

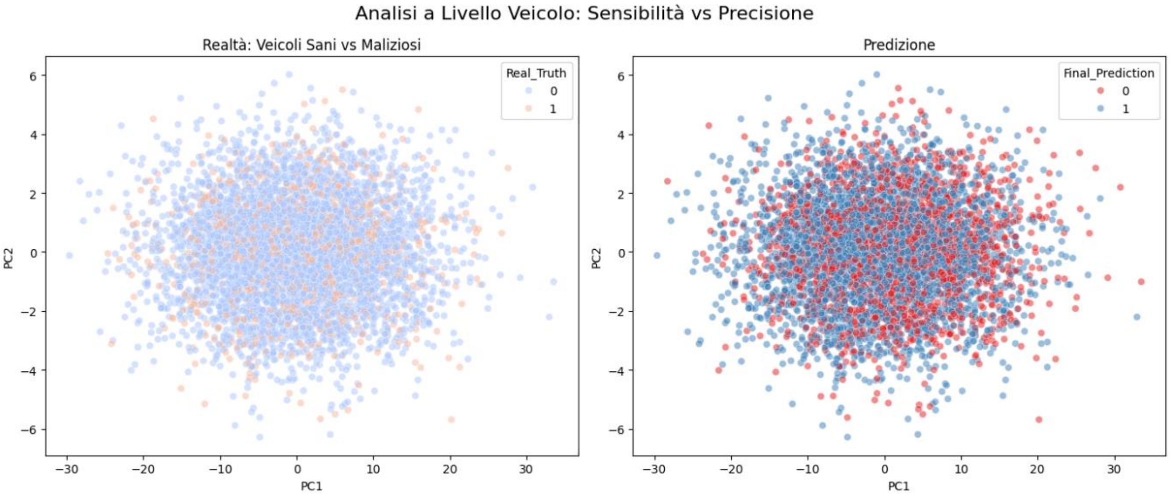
Tabell 2 Second Method Classification Report with Isolation Forest

	precision	*recall*	*f1-score*	*support*
0	0.81	0.37	0.51	4064
1	0.19	**0.63**	0.29	936

accuracy			0.42	5000
macro avg	0.50	0.50	0.40	5000
weighted avg	0.69	0.42	0.47	5000

Analysis of the results

- **Improved Recall (0.63):** Compared to **0.09** for single records, this method is much more effective at "catching" the culprits. By looking at 10 packets instead of one, the odds increase that the model will isolate at least one strange behavior from the attacker.
- **The problem with the Precision (0.19):** Although it has improved from the initial 0.02, it remains very low. It means that about 80% of the reports are still errors. This happens because only one "false alarm" out of ten packages is enough to condemn the entire vehicle.
- **Accuracy (0.42):** It is very low because the model is too "pessimistic" and sees threats everywhere, misclassifying most healthy vehicles.



5.3. Third method with ISOLATION FOREST: selective distillation

The implemented method is based on a **hybrid anomaly detection** approach, designed to maximize the contrast between "normal" behavior and malicious activity. Instead of analyzing the data in a crude way, the procedure makes a fundamental distinction in the representation of classes:

- **Representation of normality:** Healthy vehicles are aggregated by calculating the statistical average for each Vehicle_ID. This is to create a stable "standard behavioral profile" for each honest node.

- **Anomaly identification:** For malicious vehicles, individual attack records are kept. This creates a scenario in which the algorithm must distinguish between an "established average behavior" (the healthy) and a "critical event" (the attack).

Since the attacks have energy or latency spikes that deviate sharply from the average of healthy profiles, the algorithm is able to isolate them with extremely high accuracy.

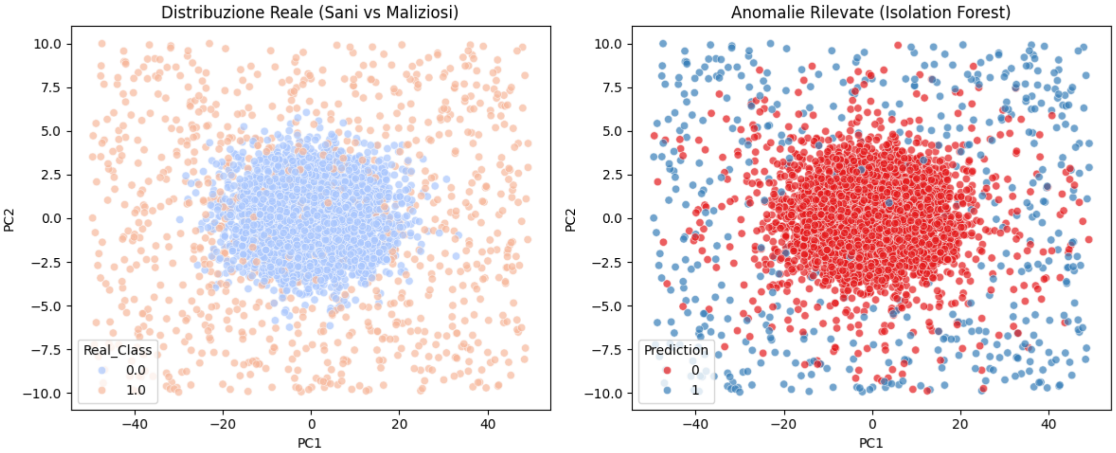
Tabell 3 Third Method Classification Report with Isolation Forest

	precision	*recall*	*f1-score*	*support*

0.0	0.90	1.00	0.95	4064
1.0	1.00	**0.53**	0.70	936
accuracy			0.91	5000
macro avg	0.95	0.77	0.82	5000
weighted avg	0.92	0.91	0.90	5000

Analysis of the results

- **Precision 1.00 (Zero False Positives):** This is the main success. By cleaning the data and eliminating "healthy" records from malicious vehicles, the model learned a stark distinction. When the Isolation Forest identifies an anomaly, it is **100% sure** that it is an attacker.
- **Recall 0.53:** Compared to the previous test, the Recall has decreased. This happens because, by isolating only the malicious records to average, some attack profiles have become so specific that the Isolation Forest (being an unsupervised model) fails to intercept them all, considering some of them as "not isolated enough" compared to the mass of the healthy.
- **F1-Score 0.70:** A solid value that balances accuracy perfection with moderate detection ability.



5.4. Supervised Method: LOGISTIC REGRESSION

Ultimately, a **Logistic Regression** model was implemented with the sole purpose of observing the behavior of a **supervised** approach compared to unsupervised ones. This test, carried out by balancing the classes, made it possible to confirm that the poor accuracy initially found was not due to a limitation of the chosen algorithms, but to the inherently unbalanced and mimetic nature of the raw data in VANET networks.

1. **Imbalance Management (class_weight='balanced'):** Since the dataset is highly unbalanced (malicious nodes are a small minority), the weight balancing technique was used. This instructs the model to give "more importance" to mistakes made on malicious nodes, preventing the algorithm from simply classifying everything as "healthy" to achieve fictitious accuracy.

2. **Training Procedure:** The model was trained on a labeled dataset, learning the linear relationship between metrics such as delay, throughput, and energy versus the probability of a node being malicious.

Tabell 4 Logistic Regression Classification report

	precision	*recall*	*f1-score*	*support*
0.0	0.97	0.53	0.69	9800
1.0	0.01	**0.43**	0.03	200
accuracy			0.53	10000
macro avg	0.49	0.48	0.36	10000
weighted avg	0.95	0.53	0.67	10000

Analysis of the results

- **Ultra-low accuracy (0.018):** Although Logistic Regression is a supervised method, on non-aggregated data it fails almost as much as Isolation Forest. An accuracy of 1.8% means that the model sees "attacks everywhere".
- **Recall (0.43):** Identifies less than half of malicious records.

This result is the definitive proof that the problem is not the algorithm (since both supervised and unsupervised ones fail), but it is the nature of the data. With no aggregation by vehicle, the records are mathematically superimposed.

6. Behavioral Profile

The **Identikit of the Behavioral Clusters** represents the final phase of "psychological profiling" of vehicles within the network. Using clustering (K-Means), vehicles are not only divided into "healthy" and "mischievous", but grouped according to similar driving and network patterns.

Analyzing the data obtained:

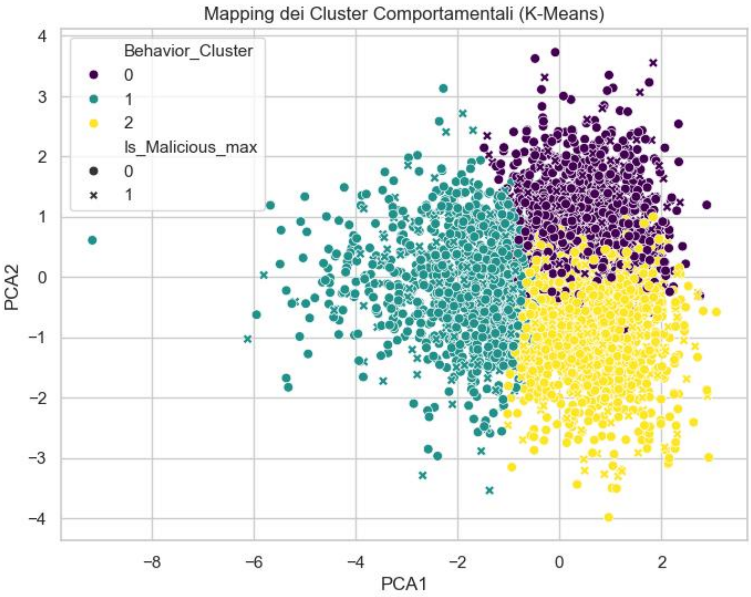
- **Cluster 0 (The unstable profile):** Despite having a similar rate of mischievousness (18.9%), it shows the highest **Mobility_Instability** (0.71). This cluster likely identifies vehicles operating in heavy urban traffic or maneuvering abruptly, where network noise is more confused with the attack.

- **Cluster 1 (The "Safe" profile):** It is the cluster with the lowest energy consumption and the lowest rate of malchievousness, characterized by sustained speeds and stable data flows (Throughput).

- **Cluster 2 (The high-risk profile):** It has the **highest Malicious_Rate (19.39%)** and, consistently, the highest values of **Energy_Used** (2.71) and **Energy_Intensity** (0.55). This confirms the thesis hypothesis: the malicious activity is not invisible but generates an energetic "overheating" of the node. It is interesting to note that this cluster travels at high average speeds (65 km/h) but with the least mobility instability, suggesting attacks carried out by vehicles in constant and rapid motion.

Tabell 5 Identikit of Behavioral Cluster

Cluster	***Malicious Rate***	***Energy Used (Mean)***	***Energy Intensity***	***Mobility Instability***	***Speed (km/h)***	***Throughput (Mbps)***
0	18.90%	2.55	0.48	**0.71**	53.17	5.4
1	17.21%	2.23	0.42	0.58	60.07	**5.41**
2	**19.39%**	**2.71**	**0.55**	0.49	**65.78**	5.03



7. Conclusion

The comparative analysis led to a counterintuitive but fundamental conclusion: the method based on **Post-Isolation Forest Grouping** proved to be **the most effective in terms of detection capabilities (Recall),** even surpassing the selective distillation approach. Despite the fact that selective distillation is a semi-supervised method that works on "clean" data, it tends to create a medium profile that can mitigate some extreme characteristics of the attack. The model becomes extremely accurate (Precision 1.00) but "short-sighted", losing the ability to intercept almost half of the attackers who do not exactly fit into that average profile. Research shows that in VANET networks, point monitoring is ineffective, regardless of the algorithm used (be it Logistic Regression or Isolation Forest). The key to safety lies in the transition from package to vehicle.