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

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

1.Introduzione al problema

L'obiettivo del progetto è di profilare il comportamento di veicoli e rilevare anomalie nelle reti 5G-VANET, utilizzando delle tecniche di apprendimento non supervisionato e poi conseguentemente confrontarle a delle etichette predefinite che identificano comportamenti malevoli.

Le reti 5G-VANET (5G Vehicular Ad-hoc Networks) rappresentano un'evoluzione delle reti veicolari tradizionali grazie alla bassa latenza, all'elevata capacità di trasmissione e alla possibilità di supportare applicazioni critiche come guida assistita, cooperazione tra veicoli, gestione del traffico e servizi di sicurezza. I veicoli nelle VANET sono nodi di una rete in movimento, la cui configurazione cambia continuamente con lo spostamento delle posizioni. Tali nodi scambiano informazioni importanti, come la velocità del veicolo, la posizione o le condizioni del traffico. Consentono ai veicoli di comunicare con successo sia tra loro che con le infrastrutture stradali. Questa natura dinamica, però, introduce anche nuove criticità, in particolare per quanto riguarda la sicurezza e l'affidabilità delle comunicazioni. comportamenti anomali o malevoli possono compromettere l'integrità e la funzionalità della rete, rendendo necessarie strategie di rilevazione delle anomalie efficienti per mitigare tali minacce. Le anomalie nelle VANET si manifestano come pattern irregolari nelle comunicazioni, nel comportamento dei veicoli o nelle interazioni con le infrastrutture di rete. Per la loro rilevazione in tempo reale sono richieste metodologie avanzate, e l'applicazione di approcci di machine learning

Il progetto affronta quindi il problema della caratterizzazione del comportamento dei veicoli e dell'individuazione di anomalie all'interno di una rete 5G-VANET. L'obiettivo è identificare pattern di funzionamento "normali" e distinguere situazioni sospette (ad esempio legate ad attacchi o a comportamenti fuori distribuzione) senza dipendere necessariamente da etichette predefinite, che nella pratica possono essere incomplete, costose da ottenere o non disponibili.

1.1 File di input

L'analisi è basata su un dataset strutturato a livello di record temporali per veicolo (ogni riga rappresenta lo stato/attività di un veicolo in un determinato istante). Per la costruzione del dataset è stato effettuato uno scan ripetuto 10 volte, a intervalli regolari di 20 secondi, su una specifica area urbana, così da campionare l'evoluzione temporale dello stato dei veicoli. Il dataset include variabili di mobilità, contesto di rete e indicatori di performance, oltre a campi utilizzabili per validazione. Le feature considerate sono:

- Vehicle_ID: identificativo univoco del veicolo.

- Vehicle_Type: tipologia di veicolo (categoria o classe, utile per distinguere profili di mobilità/consumo differenti).
- Timestamp: riferimento temporale del record, usato per analisi dinamiche e trend nel tempo.
- Latitude, Longitude: coordinate geografiche del veicolo, utili per descrivere la mobilità e la distribuzione spaziale.
- Speed_kmh: velocità del veicolo (km/h), indicatore diretto del comportamento di guida e della dinamica di movimento.
- Direction_deg: direzione di marcia in gradi, utile per stimare traiettorie e cambiamenti improvvisi.
- Congestion_Level: livello di congestione (condizione del traffico o carico locale), che influenza sia mobilità sia performance di comunicazione.
- Vehicle_Density: densità veicolare nell'area, indicatore di contesto che può impattare contesa del canale e interferenze.
- Energy_Used: energia consumata nel periodo/record (o consumo associato alle attività di rete e movimento).
- Remaining_Energy: energia residua del veicolo, utile per valutare sostenibilità e comportamento energetico nel tempo.
- Delay_ms: ritardo di comunicazione (latenza) espresso in millisecondi.
- Throughput_Mbps: throughput della comunicazione (Mbps), misura della capacità effettiva di trasmissione.
- Packet_Delivery_Ratio: rapporto di consegna pacchetti (PDR), indicatore sintetico di affidabilità della comunicazione.

Sono inoltre presenti campi informativi collegati alla presenza di attività malevole:

- Is_Malicious: etichetta che indica se il record è associato a comportamento malevolo (binario).
- Attack_Type: tipologia di attacco, quando presente (categorico).

Anomalie legate alla rete: includono attacchi di tipo **DoS**, **Wormhole**, **Black Hole**, ciascuno progettato per violare l'integrità della rete attraverso la comunicazione.
In questo lavoro, tali campi vengono considerati principalmente per validazione e confronto ex-post, mentre l'identificazione di comportamenti anomali viene impostata come problema non supervisionato, più aderente a scenari reali in cui le etichette non sono sempre disponibili o affidabili.

2. Stato dell'arte

Nelle reti 5G-VANET i veicoli comunicano tra loro e con le infrastrutture mentre si muovono; quindi, la rete cambia continuamente e questo rende più difficile mantenere comunicazioni stabili e veloci. Per gestire meglio questa dinamicità, molti lavori in letteratura usano il clustering, cioè raggruppano i veicoli in "gruppi" guidati da un nodo centrale.I metodi di clustering si basano spesso su parametri di mobilità (distanza, velocità, direzione), ma sempre più spesso integrano anche tecniche "intelligenti" come machine learning e fuzzy logic per rendere i cluster più stabili e ridurre perdite di pacchetti e ritardi.
Un altro tema molto discusso è l'integrazione tra VANET e tecnologie moderne legate al 5G e strategie per migliorare throughput e latenza. In questo filone, si evidenzia che soluzioni cluster-based possono aiutare a gestire meglio risorse e qualità del servizio, soprattutto quando la densità di veicoli cambia rapidamente.
Sul fronte sicurezza, la letteratura mette in evidenza che le VANET sono esposte a diverse anomalie e attacchi. Queste si vedono come pattern irregolari nella comunicazione o nel comportamento del veicolo (es. messaggi ritardati, dati manipolati, traiettorie/velocità incoerenti). Per questo, negli ultimi anni si è diffuso l'uso di tecniche di anomaly detection basate su machine learning e deep learning, con approcci supervisionati, non supervisionati e ibridi, e con attenzione a soluzioni robuste e possibilmente utilizzabili in tempo reale.
In questo progetto ci si colloca proprio in questa direzione: si utilizza un dataset che combina variabili di mobilità e traffico, energia e prestazioni di comunicazione e si applicano metodi di apprendimento non supervisionato e anomaly detection per descrivere i comportamenti tipici dei veicoli e individuare deviazioni sospette, usando le etichette di sicurezza soprattutto per la fase di validazione dei risultati.

3. Feature Engineering

Per aiutare gli algoritmi nella profilazione dei veicoli, sono state aggiunte le seguenti feature:

- 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

- **Rimozione colonne irrilevanti:** Vengono eliminate colonne non necessarie;
- **Label encoding:** Variabili categoriche (come il tipo di veicolo e attacco) vengono convertite in valori numerici.

5. Risultati

5.1 Metodo non supervisionato: ISOLATION FOREST

In questa fase della sperimentazione, l'algoritmo **Isolation Forest** è stato applicato direttamente sui record dei singoli pacchetti di rete, senza l'aggregazione per veicolo. Questo approccio rappresenta un tentativo di rilevamento in **tempo reale**, dove il sistema deve decidere istantaneamente se un singolo pacchetto è malevolo o meno.

	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

Interpretazione dei Risultati

Il report ottenuto sui 50.000 record grezzi evidenzia l'incapacità dell'algoritmo di distinguere le anomalie in un flusso di dati ad alta frequenza senza un pre-processing comportamentale.

- **Precisione allo 0.02:** Questo valore indica che il 98% delle segnalazioni di allarme sono in realtà falsi positivi. Nelle reti VANET, un errore di questa portata porterebbe alla paralisi della rete, poiché verrebbe revocata l'autorizzazione a quasi tutti i veicoli legittimi.
- **Recall allo 0.09:** Il modello non solo sbaglia quasi sempre quando segnala un attacco, ma manca anche il 91% degli attacchi reali.
- **La Diagnosi:** Il problema non è l'algoritmo Isolation Forest, ma la sovrapposizione statistica. Sui singoli pacchetti, i valori di *Delay* o *Energy* di un attaccante sono indistinguibili da un picco di traffico o da un'interferenza radio naturale.

5.2. Secondo metodo con ISOLATION FOREST

Il metodo è un approccio a **due stadi** che trasforma un rilevamento istantaneo (punto per punto) in un rilevamento **comportamentale (basato sul veicolo)**. È una strategia fondamentale nelle reti VANET perché un singolo pacchetto "strano" potrebbe essere solo un problema di segnale, ma una serie di anomalie indica un attacco.

1. Il Primo Stadio: Isolation Forest

Inizialmente, l'algoritmo **Isolation Forest** analizza ogni singolo record e assegna un punteggio: 1 se è considerato normale, -1 se è considerato un'anomalia.

2. Il Secondo Stadio: Aggregazione per veicolo

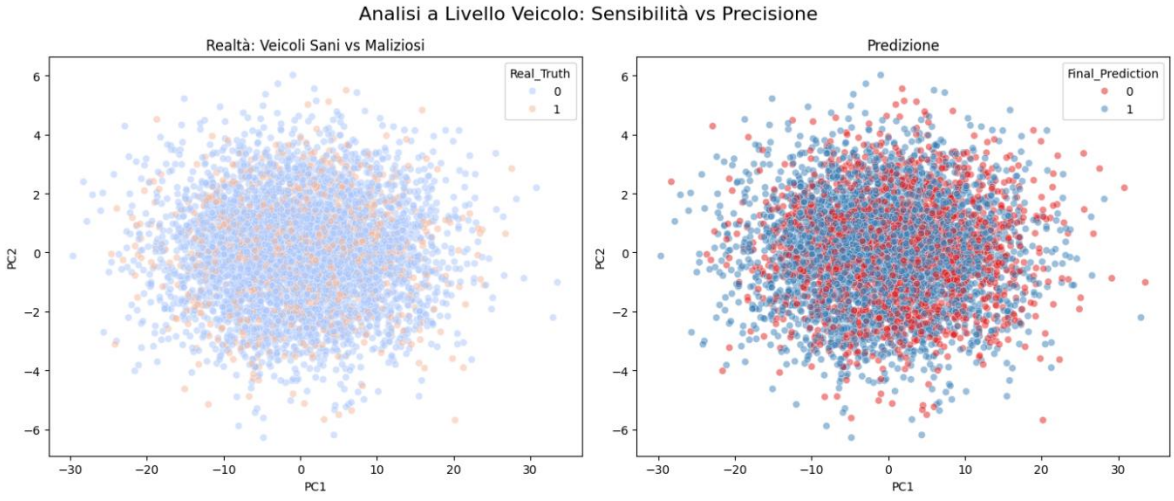
Viene analizzata la **finestra temporale** di ogni veicolo (composta da 10 record/rilevazioni):

- **Sano:** Se un veicolo invia 10 pacchetti e l'Isolation Forest dice che sono tutti normali, la somma sarà 10. Viene segnato come sano (0).
- **Sospetto:** Se anche solo uno di quei 10 pacchetti viene marcato come anomalo (punteggio -1), la somma sarà <10. Viene segnato come malizioso (1).
- **Confronto:** vengono confrontati i valori assegnati dopo l'aggregazione con quelli effettivi.

	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

Analisi dei risultati

- **Il miglioramento della Recall (0.63):** Rispetto al **0.09** dei record singoli, questo metodo è molto più efficace nel "beccare" i colpevoli. Guardando 10 pacchetti invece di uno, aumentano le probabilità che il modello isoli almeno un comportamento strano dell'attaccante.
- **Il problema della Precision (0.19):** Sebbene sia migliorata rispetto allo 0.02 iniziale, resta molto bassa. Significa che circa l'80% delle segnalazioni sono ancora errori. Questo accade perché basta un solo "falso allarme" su dieci pacchetti per condannare l'intero veicolo.
- **L'Accuratezza (0.42):** È molto bassa perché il modello è troppo "pessimista" e vede minacce ovunque, sbagliando la classificazione sulla maggior parte dei veicoli sani.



5.3. Terzo metodo con ISOLATION FOREST: distillazione selettiva

Il metodo implementato si basa su un approccio di **anomaly detection ibrida**, progettato per massimizzare il contrasto tra il comportamento "normale" e l'attività malevola. Invece di analizzare i dati in modo grezzo, la procedura opera una distinzione fondamentale nella rappresentazione delle classi:

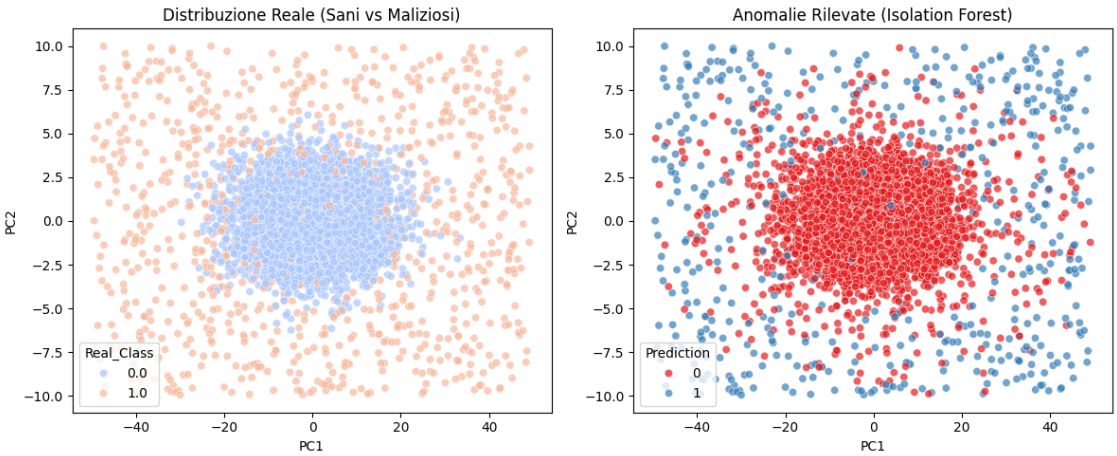
- **Rappresentazione della normalità:** I veicoli sani vengono aggregati calcolando la media statistica per ogni Vehicle_ID. Questa operazione serve a creare un "profilo comportamentale standard" stabile per ogni nodo onesto.
- **Identificazione dell'anomalia:** Per i veicoli maliziosi, vengono mantenuti i singoli record di attacco. Questo crea uno scenario in cui l'algoritmo deve distinguere tra un "comportamento medio consolidato" (i sani) e un "evento critico" (l'attacco).

Poiché gli attacchi presentano picchi energetici o di latenza che deviano nettamente dalla media dei profili sani, l'algoritmo riesce a isolarli con una precisione estremamente elevata.

	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

Analisi dei risultati

- **Precision 1.00 (Zero Falsi Positivi):** Questo è il successo principale. Avendo pulito i dati ed eliminato i record "sani" dai veicoli maliziosi, il modello ha imparato una distinzione netta. Quando l'Isolation Forest identifica un'anomalia, è **sicuro al 100%** che si tratti di un attaccante.
- **Recall 0.53:** Rispetto al test precedente, la Recall è diminuita. Questo accade perché, isolando solo i record maliziosi per fare la media, alcuni profili di attacco sono diventati così specifici che l'Isolation Forest (essendo un modello non supervisionato) non riesce a intercettarli tutti, considerandone alcuni come "non abbastanza isolati" rispetto alla massa dei sani.
- **F1-Score 0.70:** Un valore solido che bilancia la perfezione della precisione con la moderata capacità di rilevamento.



5.4. Metodo supervisionato: REGRESSIONE LOGISTICA

In ultima analisi, è stato implementato un modello di **Regressione Logistica** con l'unico scopo di osservare il comportamento di un approccio **supervisionato** a confronto con quelli non supervisionati. Questo test, effettuato bilanciando le classi, ha permesso di confermare che la scarsa precisione riscontrata inizialmente non era dovuta a un limite degli algoritmi scelti, ma alla natura intrinsecamente sbilanciata e mimetica dei dati grezzi nelle reti VANET.

1. **Gestione dello Sbilanciamento (class_weight='balanced'):** Poiché il dataset è fortemente sbilanciato (i nodi maliziosi sono una piccola minoranza), è stata utilizzata la tecnica del bilanciamento dei pesi. Questa istruisce il modello a dare "più importanza" agli errori commessi sui nodi maliziosi, evitando che l'algoritmo si limiti a classificare tutto come "sano" per ottenere un'accuratezza fittizia.
2. **Procedura di Training:** Il modello è stato addestrato su un set di dati etichettati, imparando la relazione lineare tra metriche come ritardo, throughput ed energia rispetto alla probabilità che un nodo sia malizioso.

	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

Analisi dei risultati

- **Precisione bassissima (0.018):** Nonostante la Regressione Logistica sia un metodo supervisionato, su dati non aggregati fallisce quasi quanto l'Isolation Forest. Una precisione dell'1.8% significa che il modello vede "attacchi ovunque".
- **Recall (0.43):** Identifica meno della metà dei record maliziosi.

Questo risultato è la prova definitiva che il problema non è l'algoritmo (visto che falliscono sia quello supervisionato che quello non supervisionato), ma è la natura dei dati. Senza aggregazione per veicolo, i record sono matematicamente sovrapposti.

6. Behavioral Profile

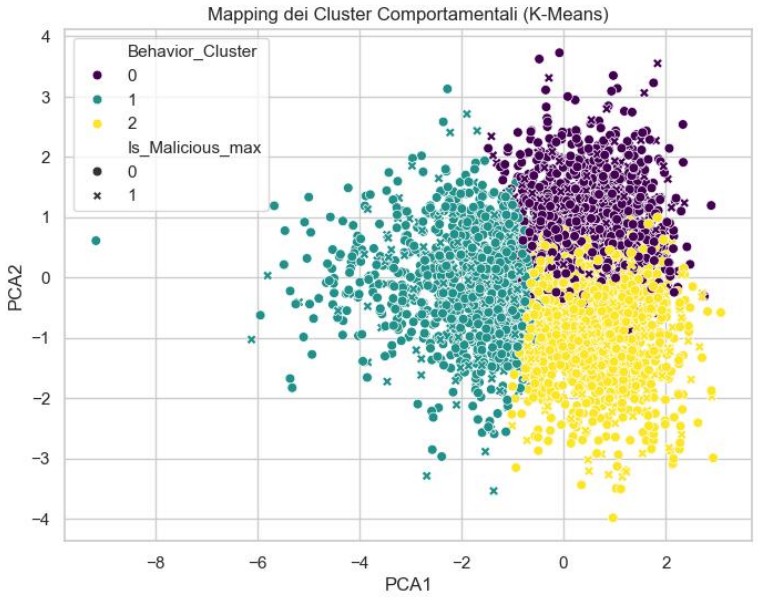
L'**Identikit dei Cluster Comportamentali** rappresenta la fase finale di "profilazione psicologica" dei veicoli all'interno della rete. Attraverso l'uso del clustering (K-Means), i veicoli non vengono solo divisi tra "sani" e "maliziosi", ma raggruppati in base a pattern di guida e di rete simili.

Analizzando i dati ottenuti:

- **Cluster 0 (Il profilo instabile):** Pur avendo un tasso di maliziosità simile (18.9%), mostra la massima **Mobility_Instability** (0.71). Questo cluster identifica probabilmente veicoli che operano in condizioni di traffico urbano intenso o che effettuano manovre brusche, dove il rumore di rete si confonde maggiormente con l'attacco.

- **Cluster 1 (Il profilo "Safe"):** È il cluster con il minor consumo energetico e il minor tasso di maliziosità, caratterizzato da velocità sostenute e flussi di dati (Throughput) stabili.
- **Cluster 2 (Il profilo ad alto rischio):** Presenta il **Malicious_Rate più elevato (19.39%)** e, coerentemente, i valori più alti di **Energy_Used** (2.71) ed **Energy_Intensity** (0.55). Questo conferma l'ipotesi della tesi: l'attività malevola non è invisibile, ma genera un "surriscaldamento" energetico del nodo. Interessante notare come questo cluster viaggi a velocità medie elevate (65 km/h) ma con la minor instabilità di mobilità, suggerendo attacchi portati da veicoli in movimento costante e rapido.

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. Conclusioni

L'analisi comparativa ha portato a una conclusione controintuitiva ma fondamentale: il metodo basato sul **Raggruppamento Post-Isolation Forest** si è dimostrato **il più efficace in termini di capacità di rilevamento (Recall)**, superando persino l'approccio di distillazione selettiva. Nonostante la distillazione selettiva sia un metodo semi-supervisionato che lavora su dati "puliti", tende a creare un profilo medio che può attenuare alcune caratteristiche estreme dell'attacco. Il modello diventa estremamente preciso (Precision 1.00) ma "miope", perdendo la capacità di intercettare quasi la metà degli attaccanti che non rientrano esattamente in quel profilo medio. La ricerca dimostra che nelle reti VANET il monitoraggio puntuale è inefficace, indipendentemente dall'algoritmo usato (sia esso la Regressione Logistica o l'Isolation Forest). La chiave della sicurezza risiede nel passaggio dal pacchetto al veicolo.