I.J. Wireless and Microwave Technologies, **2026, 5, 89-107**
Published Online on October 8, 2026 by MECS Press (http://www.mecs-press.org/)
DOI: 10.5815/ijwmt.2026.05.06

Benchmarking Supervised and Unsupervised Paradigms for Anomaly Detection in Autonomous Vehicle Telemetry: The Pitfalls of Internal Clustering Metrics

Nizirwan Anwar*
Department of Informatics Engineering, University of Esa Unggul, Jakarta, Indonesia
E-mail: nizirwan.anwar@esaunggul.ac.id
ORCID iD: https://orcid.org/0000-0003-1189-9093
*Corresponding Author

Titik Khawa Abdul Rahman
School Sains of Technology (SST), Asia e University, Malaysia
E-mail: titik.khawa@aeu.edu.my
ORCID iD: https://orcid.org/0000-0002-7174-7625

Aedah Abd Rahman
School Sains of Technology (SST), Asia e University, Malaysia
E-mail: aedah.abdrahman@aeu.edu.my
ORCID iD: https://orcid.org/0000-0002-4944-1179

Swa Lee Lee
School Business Administration, Asia e University, Malaysia
E-mail: swa.lee@aeu.edu.my
ORCID iD: https://orcid.org/0009-0005-6376-7748

Muhammad Faisal
Department of Informatics Engineering, University of Muhammadiyah Makassar, Indonesia
E-mail: muhfaisal@unismuh.ac.id
ORCID iD: https://orcid.org/0000-0003-1469-9468

Dewanto Rosian Adhy
Department of Informatics Engineering, University of Mayasari Bakti, Tasikmalaya, Indonesia
E-mail: dewanto_ra@mayasaribakti.ac.id
ORCID iD: https://orcid.org/0000-0001-8533-3450

Tomy Ronaldi
Department of Defense Science, Republic of Indonesia Defense University Bogor, Indonesia
E-mail: tomyronaldi@gmail.com
ORCID iD: https://orcid.org/0000-0002-0517-7505

Arief Ichwani
Department of Informatics Engineering, University of Esa Unggul, Jakarta, Indonesia
E-mail: arief.ichwani@esaunggul.ac.id
ORCID iD: https://orcid.org/0000-0002-3406-3355

Received: July 6, 2026; Revised: August 7, 2026; Accepted: September 10, 2026; Published: October 8, 2026

This work is open access and licensed under the Creative Commons CC BY 4.0 License.

Abstract: Anomaly detection in autonomous vehicle (AV) telemetry is a safety-critical task requiring accurate identification of abnormal vehicle behavior from multivariate kinematic sensor streams. This study presents a comprehensive comparative benchmark of twelve supervised classification algorithms and thirteen unsupervised clustering configurations applied to a real-world AV telemetry dataset comprising 112,028 observations and 23 engineered kinematic features after removal of seven constant-value columns. The dataset exhibits an approximately balanced class distribution (Normal: 56,380; Anomaly: 55,648). All classifiers were trained on an 80/20 stratified split and evaluated on Accuracy, Precision, Recall, F1-Score, and AUC-ROC. Clustering algorithms were evaluated on 15,000 randomly sampled observations using Silhouette Score, Davies-Bouldin Index (DBI), Calinski-Harabasz Index (CHI), Adjusted Rand Index (ARI), and Normalized Mutual Information (NMI). Results demonstrate that tree-based ensemble classifiers achieve performance approaching theoretical upper bounds — Extra Trees, Random Forest, Bagging, and Decision Tree (each F1 = 0.9998) — driven by the dominant discriminative power of speed, speed_time_ratio, rolling_distance_mean, and distance, which collectively account for 59.7% of Random Forest feature importance. In contrast, all unsupervised clustering algorithms fail to semantically recover ground-truth anomaly labels: the best-Silhouette configurations (Birch k=3, Silhouette = 0.9529) achieve ARI ≈ 0.000, while the best-ARI configuration (K-Means k=2, ARI = 0.440) recovers only partial anomaly structure. Three robustness experiments qualify these figures: 5-fold cross-validation confirms stability under random partitioning (F1 = 0.9998 ± 0.0001 for the top ensembles), while strictly temporal and unseen-scenario partitioning reduce peak F1 to 0.9750 and 0.9602 respectively, quantifying the contribution of window- and scenario-level information leakage to the near-saturated hold-out scores; a ten-sample sensitivity analysis shows the K-Means label alignment to be sample-dependent (ARI = 0.20 ± 0.21) whereas the degenerate high-Silhouette collapse is fully robust. Under the default configurations and static distribution assumption examined here, these findings indicate that supervised classification is the preferred paradigm for AV anomaly detection when labeled telemetry is available, and establish that Silhouette Score alone is a misleading model selection criterion in this context. All experiments were run with default hyperparameters in a documented software environment; robustness under hyperparameter tuning and distribution drift remains for future work.

Index Terms: Autonomous vehicle, anomaly detection, machine learning, supervised classification, unsupervised clustering, ensemble methods, kinematic telemetry

1. Introduction

The rapid advancement of autonomous vehicle (AV) technology has positioned intelligent mobility systems at the intersection of safety engineering and real-time data analytics. As AVs integrate into public transportation infrastructure, ensuring continuous operational safety demands robust mechanisms capable of identifying abnormal vehicle behavior before it escalates into critical failures. Anomaly detection — distinguishing atypical observations from a normal operating baseline — is now a foundational component of AV safety architecture, applicable to sensor fault identification, control system instability, and cybersecurity intrusion recognition [1-3]. Vehicle telemetry data provides a continuous multivariate stream of kinematic, positional, and dynamic signals whose statistical properties differ systematically between normal and anomalous states, making it a natural substrate for machine learning-based monitoring [4-6]. The consequences of undetected anomalies are severe: erroneous navigation decisions and unrecognized obstacles can result in fatal outcomes, establishing anomaly detection not merely as a performance optimization task but as a direct safety-critical requirement [7-10]. A further methodological pitfall deserves emphasis at the outset: because telemetry is temporally autocorrelated, naive random row-level partitioning of time-series-derived features can leak information between training and test sets and inflate apparent performance, so benchmark results on such data must be interpreted with this caveat in mind.

Supervised machine learning has emerged as the dominant paradigm for AV anomaly detection when ground-truth labeled data is available from simulation platforms or annotated incident logs. Tree-based ensembles, neural networks, and kernel methods have each demonstrated strong discriminative performance on structured kinematic telemetry. [11] proposed a Denoising Variational Transformer (DVT) for AV sensor anomaly detection reporting F1-Score improvements of 8–50% over autoencoder baselines, while [12] combined Local Outlier Factor, Isolation Forest, and One-Class SVM in a UAV telemetry anomaly framework, demonstrating that unsupervised methods require extensive domain-specific feature engineering to approach meaningful anomaly alignment. More recently, the lightweight GRAD framework [13], using a Gated Recurrent Unit with a Multi-Stage Sliding Window mechanism, achieved F1 = 0.976 for real-time AV anomaly classification — a competitive result for online deployment, yet still below the ceiling performance observed for classical tree ensembles on engineered kinematic feature sets.

Despite this breadth of research, three persistent gaps limit the practical utility of existing literature for operational AV monitoring system design. First, systematic benchmarking across a wide range of classical supervised and unsupervised algorithms on the same AV telemetry dataset remains scarce, with most studies evaluating a narrow method subset under non-standardized conditions [1,14]. Second, the relative advantage of supervised classification over

unsupervised clustering for semantically meaningful anomaly identification has not been rigorously quantified using external alignment metrics (ARI, NMI) alongside internal geometric metrics (Silhouette, DBI), leaving practitioners without an evidence-based basis for paradigm selection [15]. Third, the internal-external metric divergence problem — wherein extremely high Silhouette Scores coexist with near-zero ARI — remains insufficiently documented in AV-specific contexts, creating the risk that algorithms with zero semantic anomaly content are falsely selected as operationally optimal [15,16].

This study addresses all three gaps through a structured benchmark on a real-world AV simulation telemetry dataset of 112,028 observations across 23 engineered kinematic features. The contributions are listed below as specific technical paths.

- (1) A standardized preprocessing pipeline including constant-feature removal, z-score normalization, and stratified partitioning.
- (2) Systematic evaluation of twelve supervised classifiers — spanning tree ensembles, gradient boosting, kernel methods, neural networks, and probabilistic models — on Accuracy, Precision, Recall, F1-Score, and AUC-ROC. Robustness is verified under 5-fold cross-validation and under strictly temporal and unseen-scenario partitioning, quantifying information-leakage effects.
- (3) Evaluation of thirteen unsupervised clustering configurations across seven algorithm families using five complementary internal and external metrics, consistent with the benchmark design principles of [14]. A ten-sample sensitivity analysis quantifies the sample dependence of every configuration.
- (4) Empirical demonstration and physical interpretation of the internal-external metric divergence, with direct implications for production AV monitoring system design.

The remainder of this paper is organized as follows: Section 2 reviews related work; Section 3 describes the dataset, preprocessing, and methodology; Section 4 reports results; Section 5 discusses their implications; and Section 6 concludes with limitations and future research directions.

2. Literature Review

Anomaly detection in autonomous and connected vehicles has attracted substantial research attention across statistical, classical machine learning, and deep learning paradigms. [17] provided a unifying taxonomy of deep and shallow anomaly detection methods, establishing that supervised, semi-supervised, and unsupervised approaches differ fundamentally in their optimization objectives and that no single paradigm dominates across all dataset types. Building on this, [1] reviewed 203 studies on AV anomaly detection, identifying LSTM, CNN, and autoencoder architectures as the most frequently applied methods, while highlighting the persistent absence of standardized benchmarking datasets as a critical field-wide limitation. [11] addressed this gap by proposing a Denoising Variational Transformer (DVT) for AV sensor anomaly detection, reporting F1-Score improvements of 8–50% over autoencoder baselines, though both remain well below the near-perfect performance of classical tree ensembles on structured kinematic features. [12] further demonstrated that automated machine learning ensembles combining one-class methods for UAV telemetry anomaly detection require extensive feature engineering to overcome the inherent limitation that these methods optimize geometric outlier criteria rather than label-aligned class boundaries, while [13] confirmed that even optimized GRU-based real-time classifiers (F1 = 0.976) remain below classical ensemble ceiling performance on well-engineered kinematic feature sets.

Supervised ensemble methods have established themselves as the de facto standard for binary anomaly classification on structured tabular sensor data [18]; [19,20] demonstrated that Random Forest's combination of bagging and random feature subsampling produces strong generalization on structured tabular data with minimal parameter tuning, a property directly reflected in the 99.98% test accuracy observed in this study. The theoretical basis for Extra Trees performance on correlated feature spaces was established by [19], who showed that extreme threshold randomization reduces sensitivity to correlated feature clusters — a property directly relevant to the speed–speed_lag_1–speed_time_ratio cluster ($r > 0.95$) identified in this dataset. [21] confirmed these findings in an IoT security context, demonstrating through a comparative analysis of ensemble tree configurations for botnet detection that bagging ensembles consistently match or outperform boosting methods on near-balanced binary classification datasets, and that standalone Decision Trees approach ensemble performance when the decision boundary is geometrically simple — a pattern replicated here where Decision Tree (F1 = 0.9998) matches Bagging and Extra Trees to within $\Delta F1 < 0.0001$. The comprehensive ADBench benchmark [14], evaluating 30 anomaly detection algorithms across 57 datasets, further confirmed that supervised methods universally outperform unsupervised ones when ground-truth labels are available, with no unsupervised algorithm statistically dominating others across dataset types.

The evaluation of unsupervised clustering for anomaly detection introduces distinct methodological challenges, particularly regarding the choice of evaluation metrics. [15] evaluated 19 unsupervised anomaly detection algorithms across 10 multivariate datasets and demonstrated that algorithm ranking based solely on internal geometric criteria frequently diverges from ranking based on external label alignment, and that no single algorithm consistently dominates across all data types — motivating the multi-metric, multi-algorithm evaluation design adopted in this study. [16]

conducted an extensive comparative study of 30 cluster validity indices across multiple dataset types and clustering algorithms, finding that Silhouette Score systematically overestimates cluster quality on datasets with one dominant density region, and recommending that at least two complementary validation indices — one internal, one external — always be reported jointly, a recommendation adopted as a core methodological principle in this study. [22,23] demonstrated in the related context of in-vehicle CAN network anomaly detection that fully-supervised methods universally outperform semi-supervised alternatives when sufficient labeled data is available, with XGBoost achieving state-of-the-art accuracy (98.65%) — further reinforcing that label availability is the decisive factor in paradigm selection. Comprehensive surveys on anomaly-based IDS for industrial and networked systems, and federated IDS architectures further contextualize the breadth of anomaly detection methods across domains, while surveys of deep learning methods and reviews of in-vehicle network intrusion detection systems establish the methodological landscape within which this benchmark is positioned. Nevertheless, a critical reading of this literature reveals recurring limitations rather than a settled methodological landscape. Deep architectures such as DVT [11] and GRAD [13] optimize reconstruction or sequence-forecasting objectives and impose non-trivial inference latency budgets, yet their reported gains are rarely benchmarked against strong classical baselines on identical engineered feature sets; conversely, classical studies seldom report external label-alignment metrics for their unsupervised components, leaving the semantic validity of high internal scores unexamined. Emerging directions relevant to vehicle dynamics modeling include graph neural network detectors that represent inter-signal dependency structure explicitly [33] and temporal Transformer architectures such as Anomaly Transformer [34] and TranAD [35]; these are promising for capturing long-range temporal dependencies, but their advantage over tree ensembles on tabular kinematic features remains unestablished [19], their computational latency is largely uncharacterized for onboard deployment, and none of these works documents the internal-external metric divergence that this study quantifies.

3. Methodology

3.1. Dataset Description

The Autonomous Vehicle Anomaly Detection dataset comprises 112,028 telemetry records generated with the MOBATSim simulation framework (MATLAB 2020b) and published openly on IEEE DataPort by Pallapothu et al. [36] (doi: 10.21227/0q64-5f96), each record annotated with a binary ground-truth label: Normal (0) or Anomaly (1), where anomalous records correspond to labeled False Data Injection Attacks (FDIA) against the vehicle sensor streams. The corpus comprises 28 recording sessions of 4,001 records each, spanning urban, highway, and obstacle-laden driving scenarios. The original 33 columns include raw positional, rotational, and kinematic signals recorded at high temporal frequency across multiple simulated driving scenarios. The label distribution is approximately balanced — 56,380 Normal records (50.33%) and 55,648 Anomaly records (49.67%) — eliminating the need for resampling techniques such as SMOTE or class-weight adjustment that are typically required in imbalanced anomaly detection benchmarks [24-26]. Table 1 summarizes the dataset properties after preprocessing.

Table 1. Dataset Summary After Preprocessing

Property	Value
Total observations	112,028
Original features	33
Constant features removed	7
Final feature count	23
Normal samples (label = 0)	56,380 (50.33%)
Anomaly samples (label = 1)	55,648 (49.67%)
PCA explained variance (2 components)	36.60%
Train / Test split	89,622 / 22,406 (80/20 stratified)
Clustering sample size	15,000 (stratified random)

3.2. Preprocessing Pipeline

Seven features were removed during preprocessing due to zero variance across the entire dataset: x_rot, y_rot, z_rot, z_pos, angular_velocity, angular_acceleration, and angular_displacement. Constant-value features carry no discriminative information and introduce numerical instability in distance-based and kernel-based algorithms. The retained 23 features span five semantic categories:

- Raw kinematics: Time, rot_angle, speed, x_pos, y_pos, acceleration, jerk, speed_change_rate, distance
- Cumulative signals: cumulative_distance, cumulative_speed, cumulative_jerk
- Temporal: time_diff
- Lag features: speed_lag_1, acceleration_lag_1, distance_lag_1

- Rolling window statistics: rolling_speed_mean, rolling_speed_std, rolling_acceleration_mean, rolling_acceleration_std, rolling_distance_mean, rolling_distance_std, speed_time_ratio

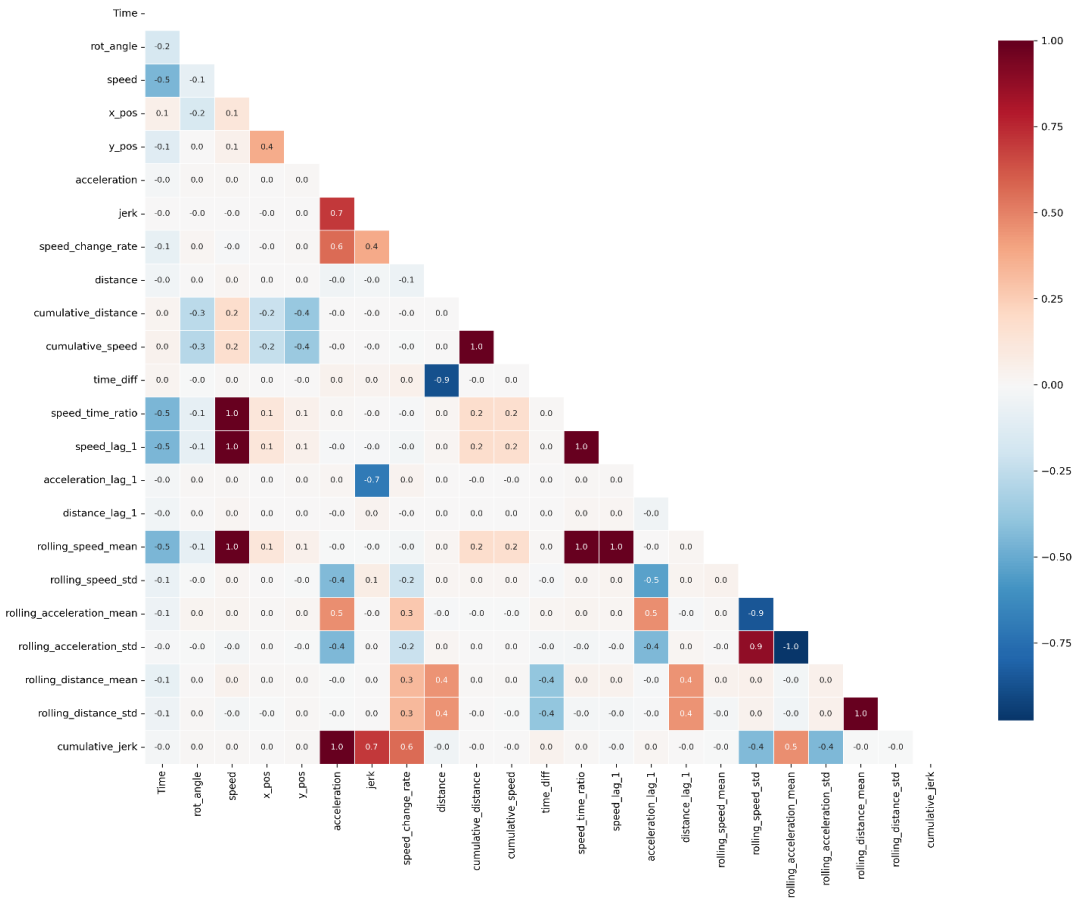


Fig. 1. Feature Correlation Matrix of Vehicular Dynamics and Derived Temporal Features

Feature correlation analysis (Fig.1) reveals three major collinearity clusters: (1) speed–speed_lag_1–speed_time_ratio–rolling_speed_mean ($r > 0.95$); (2) cumulative_distance–cumulative_speed ($r \approx 1.0$); and (3) jerk–acceleration–speed_change_rate ($r = 0.6–0.7$). These redundancies are exploited by tree-based algorithms through random feature subsampling [19] but contribute to the low PCA explained variance (36.60%), indicating that the anomaly signal is distributed across multiple correlated dimensions. All 23 features were standardized using z-score normalization prior to all model training to eliminate scale bias in distance-based, kernel-based, and gradient-based algorithms.

3.3. Data Partitioning

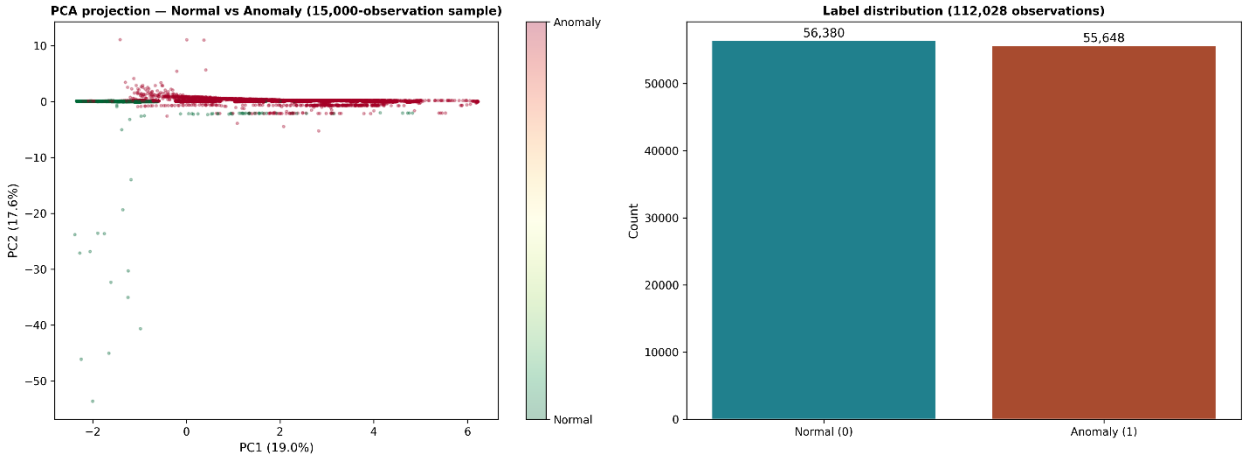


Fig. 2. Left: PCA 2D projection of a 15,000-observation sample showing dense overlap between Normal (green) and Anomaly (red) classes. Right: Label distribution across 112,028 observations confirming an approximately balanced class distribution (Normal = 56,380; Anomaly = 55,648).

The full 112,028-record dataset was partitioned into a training set (89,622 records, 80%) and a held-out test set (22,406 records, 20%) using stratified random splitting to preserve the 50:50 class distribution in both partitions. As shown in Fig.2, the PCA projection confirms that Normal and Anomaly observations are densely co-located in two-dimensional space, indicating that the class boundary is not recoverable through geometric partitioning alone. For unsupervised clustering evaluation, a separate simple random sample of 15,000 observations was drawn with a fixed seed from the full standardized dataset to manage computational feasibility for density-based and hierarchical algorithms. The elbow method and Silhouette-vs-K analysis were conducted on the same 15,000-observation sample.

3.4. Supervised Classification

Twelve supervised binary classifiers were evaluated under identical experimental conditions using default scikit-learn and XGBoost configurations to ensure that performance differences reflect intrinsic algorithmic properties [27]:

- Bagging ensembles: Random Forest (Breiman, 2001), Extra Trees (Geurts et al., 2006), Bagging — all n_estimators = 100, random_state = 42, n_jobs = −1
- Gradient boosting: XGBoost (n_estimators = 100), Gradient Boosting (n_estimators = 100), AdaBoost (n_estimators = 100, algorithm = SAMME)
- Single tree: Decision Tree (random_state = 42, no depth constraint)
- Instance-based: KNN (k = 5, n_jobs = −1)
- Kernel method: SVM (RBF kernel, probability = True, max_iter = 5,000)
- Neural network: MLP (hidden layers: 128→64, max_iter = 300)
- Linear model: Logistic Regression (max_iter = 1,000)
- Probabilistic: Gaussian Naive Bayes

Performance was quantified using Accuracy (Eq. (1)), Precision (Eq. (2)), Recall (Eq. (3)), F1-Score (Eq. (4)), and AUC-ROC. Feature importance was extracted using the mean decrease in impurity (Gini importance) criterion intrinsic to tree-based ensembles (Breiman, 2001). This impurity-based measure is distinct from model-agnostic post-hoc attribution frameworks such as SHAP (Lundberg et al., 2020), which is not used in this study; the earlier conflation of the two has been corrected. All hyperparameters were deliberately held at library defaults to establish a reproducible, tuning-free baseline in which performance differences reflect intrinsic algorithmic behavior rather than differential tuning effort; no hyperparameter search or sensitivity analysis was conducted, so reported results should be read as lower bounds on attainable performance for each algorithm (this constraint is restated as a limitation in Section 6).

$$Accuracy = \left(\frac{TP + TN}{TP + TN + FP + FN}\right) \tag{1}$$

$$Precision = \left(\frac{TP}{TP + FP}\right) \tag{2}$$

$$AcRecall = \left(\frac{TP}{TP + FN}\right) \tag{3}$$

$$F1 = \left(\frac{2 \cdot \text{Precision} \cdot \text{Recall}}{\text{Precision} + \text{Recall}}\right) \tag{4}$$

3.5. Unsupervised Clustering

Thirteen clustering configurations spanning seven algorithm families were evaluated on the 15,000-observation random sample:

- Partition-based: K-Means at k = 2, 3, 4 (n_init = 10); Mini-Batch K-Means at k = 2
- Model-based: GMM at k = 2 and 3
- Hierarchical: Agglomerative Clustering with Ward, Complete, and Average linkage (n_clusters = 2)
- Tree-structured: Birch at k = 2 and k = 3
- Density-based: DBSCAN at eps = 1.5 and eps = 2.0 (min_samples = 15)

Five complementary metrics were computed. Internal metrics — Silhouette Score (Eq. (5)), Davies-Bouldin Index (DBI), and Calinski-Harabasz Index (CHI) — assess geometric cluster quality without label reference. External metrics — Adjusted Rand Index (ARI) and Normalized Mutual Information (NMI) — assess semantic alignment with ground-truth labels. Optimal cluster number was assessed through the Elbow Method and Silhouette-vs-K analysis over K = 2 to 10, following the methodology recommended by Schubert and Rousseeuw (2021). Joint reporting of internal and external

metrics was adopted as a mandatory practice consistent with Arbelaitz et al. (2013). A note on the applicability boundary of these metrics is warranted for this dataset: the Silhouette Score assumes convex, comparably dense clusters and is computed here on the full 23-dimensional standardized space, where the strong collinearity documented in Section 3.2 ($r > 0.95$ within feature clusters) causes Euclidean inter-point distances to be dominated by a few correlated directions — precisely the regime in which Silhouette is known to overestimate cluster quality when one dense region dominates [16].

$$s = \frac{b - a}{\max(a, b)} \tag{5}$$

3.6. Evaluation Framework

The central methodological contribution is the parallel dual-track evaluation of both paradigms using joint internal-external metrics, enabling direct quantitative comparison of each paradigm's ability to recover ground-truth anomaly structure. For supervised classifiers, F1-Score serves as the primary ranking metric with AUC-ROC as a secondary probabilistic calibration measure. For clustering, ARI and NMI serve as the primary selection criteria where ground-truth labels are available, with Silhouette Score and DBI as secondary geometric diagnostics — consistent with the joint reporting recommendation of [15,16]. All experiments were implemented in Python 3 using scikit-learn, XGBoost, and standard data science libraries, with random seeds fixed at 42 for full reproducibility. Primary results (Tables 2 and 4) derive from the fixed-seed hold-out split and 15,000-observation sample described in Section 3.3; their robustness is verified through 5-fold stratified cross-validation with per-fold standardization, strictly temporal partitioning, unseen-scenario partitioning (Section 4.2, Table 3), and a ten-sample clustering sensitivity analysis (Section 4.3, Table 5). All experiments were executed on a 12th Gen Intel Core i5-12450HX (12 threads, 16 GB RAM) under Python 3.12 with scikit-learn 1.8.0 and XGBoost 3.2.0; all reported training times refer to this environment. The end-to-end data flow between modules is summarized in Algorithm 1.

Algorithm 1. Dual-track benchmark procedure.

Input: telemetry dataset D (112,028 × 33 columns), ground-truth labels y

1: Remove 7 zero-variance features; retain 23 kinematic features
2: Standardize all features (z-score) → $\tilde{D}$
3: Stratified 80/20 split of $\tilde{D}$ → (Train, Test)
4: for each of the 12 supervised classifiers: fit on Train; predict on Test; record Accuracy, Precision, Recall, F1 (Eqs. (1)–(4)), AUC-ROC
5: Draw stratified 15,000-observation sample S from $\tilde{D}$
6: for each of the 13 clustering configurations: fit on S; record internal metrics Silhouette (Eq. (5)), DBI, CHI and external metrics ARI, NMI against y
7: Rank classifiers by F1-Score; rank clusterings jointly by (ARI, NMI) and (Silhouette, DBI)
8: Robustness: repeat step 4 under 5-fold CV, temporal, and unseen-scenario splits; repeat step 6 on 10 independent samples

Output: dual-track ranking and internal–external divergence report

4. Results

4.1. Feature Analysis

PCA applied to all 23 standardized features yields a two-component cumulative explained variance of only 36.60% — PC1 capturing 19.0% and PC2 capturing 17.6% — indicating that the anomaly signal is distributed across many correlated dimensions rather than concentrated in a few dominant axes (Fig.2). The PCA scatter plot reveals that Normal and Anomaly observations are densely co-located in the two-dimensional projection, confirming that the class boundary is not recoverable through geometric partitioning alone. Random Forest feature importance (Fig.3), interpreted using the mean decrease in impurity criterion (Breiman, 2001), reveals a clear dominance hierarchy: speed (17.4%), speed_time_ratio (16.3%), rolling_distance_mean (13.3%), and distance (12.7%) collectively account for 59.7% of total importance, while the bottom eight features contribute less than 3% combined. Fig. 3. presents the full importance ranking.

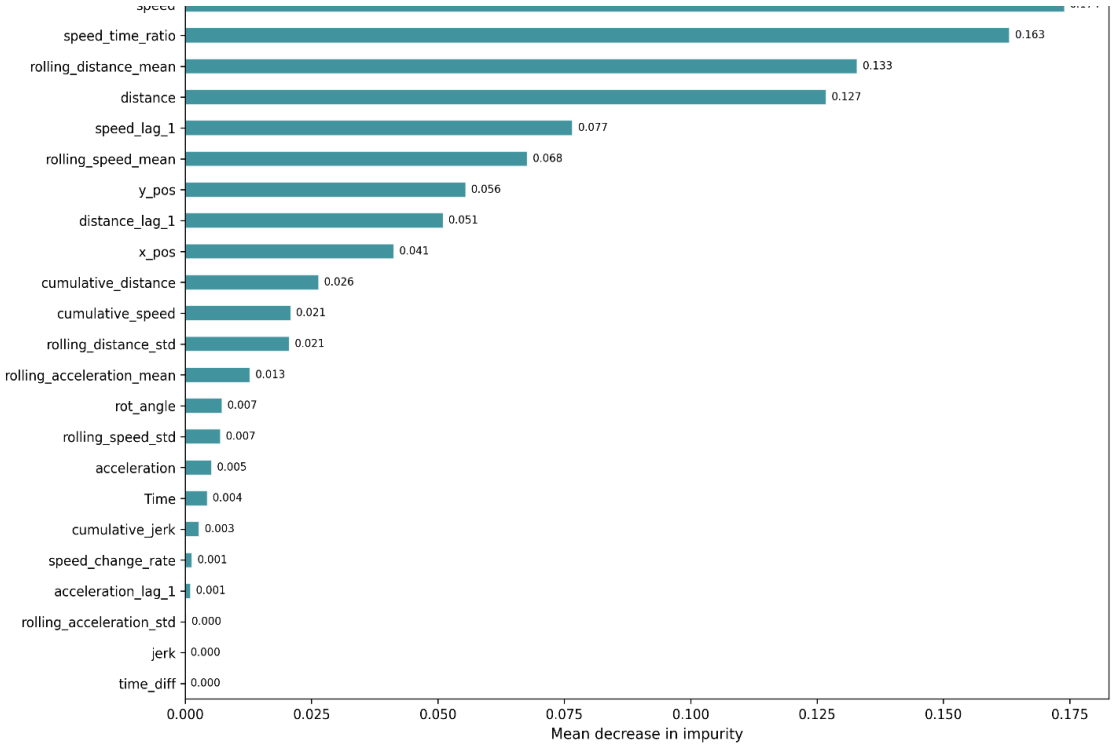


Fig. 3. Random Forest mean decrease in impurity feature importance across all 23 kinematic features. The top four features — speed, speed_time_ratio, rolling_distance_mean, and distance — account for 59.7% of total discriminative power.

4.2. Supervised Classification Performance

All twelve supervised classifiers achieved high classification performance on the 22,406-record held-out test set, with F1-Scores ranging from 0.8646 (Naive Bayes) to 0.9998 (Extra Trees, Random Forest, Bagging, and Decision Tree, tied to four decimal places). The complete results are reported in Table 2.

Table 2. Supervised Classification Results (sorted by F1-Score)

Classifier	Accuracy	Precision	Recall	F1-Score	AUC-ROC	Time (s)
Extra Trees	0.9998	0.9999	0.9997	0.9998	1.0000	2.3
Bagging	0.9998	0.9999	0.9997	0.9998	1.0000	22.4
Random Forest	0.9998	0.9999	0.9997	0.9998	1.0000	4.7
Decision Tree	0.9998	0.9998	0.9998	0.9998	0.9998	0.5
KNN	0.9995	0.9997	0.9993	0.9995	0.9999	<0.1
XGBoost	0.9994	0.9996	0.9991	0.9994	1.0000	1.8
Gradient Boosting	0.9983	0.9973	0.9994	0.9983	1.0000	16.7
MLP Neural Net	0.9983	0.9984	0.9982	0.9983	1.0000	20.5
AdaBoost	0.9947	0.9925	0.9969	0.9947	0.9999	7.9
SVM (RBF)	0.9940	0.9898	0.9982	0.9940	0.9993	137.3
Logistic Regression	0.9749	0.9868	0.9623	0.9744	0.9973	4.1
Naive Bayes	0.8802	0.9853	0.7703	0.8646	0.9902	<0.1

Six classifiers achieved AUC-ROC = 1.0000 and ten of twelve exceeded AUC-ROC = 0.999, confirming that the anomaly decision boundary is highly learnable with supervised guidance (Fig.5). The strongest performance drops occur at Logistic Regression (Recall = 0.9623) and Naive Bayes (Recall = 0.7703), both of which make linearity or feature independence assumptions incompatible with the correlated, non-linear kinematic feature structure identified in the correlation analysis. Fig.4 shows the confusion matrices for the top six classifiers. Decision Tree produces only 4 misclassifications from 22,406 test records — 2 false positives and 2 false negatives — confirming that the kinematic anomaly boundary is nearly perfectly learnable. On the other hand, Fig.5 presents the ROC curves for all twelve classifiers. Six ensemble models achieve AUC-ROC = 1.0000, with curves indistinguishable from the upper-left corner, while even the weakest classifier (Naive Bayes) achieves AUC = 0.9902. The near-saturated scores warrant explicit scrutiny for information leakage and overfitting, which is quantified here directly through three robustness experiments (Table 3).

First, under 5-fold stratified cross-validation with the scaler refit inside each training fold, the top ensembles remain at F1 = 0.9998 ± 0.0001, confirming that the hold-out results are not artifacts of one particular random split. Second, under strictly temporal partitioning — the first 80% of each recording's timeline used for training and the final 20% for testing — peak performance falls to F1 = 0.9750 (AdaBoost), because lag, rolling-window, and cumulative features no longer share overlapping windows across the partition boundary. Third, under unseen-scenario partitioning — six complete recordings (24,006 records) held out entirely — peak performance falls to F1 = 0.9602, the model ranking reshuffles substantially (Logistic Regression and SVM generalize best; Random Forest drops to F1 = 0.9249), and the near-perfect hold-out figures are revealed to rest partly on within-scenario and within-window memorization. Random row-level splitting therefore inflates absolute F1 by roughly 2.5 points relative to temporal splitting and 4–8 points relative to scenario-level splitting. The paradigm comparison is unaffected — even the weakest leakage-free supervised configuration far exceeds the best unsupervised label alignment (Section 4.3) — but absolute hold-out values should be read as upper bounds on deployable performance.

Table 3. F1-Score under four evaluation protocols: stratified hold-out (as Table 2), 5-fold stratified cross-validation (mean ± standard deviation), strictly temporal within-recording split (first 80% train / last 20% test), and unseen-scenario split (six complete recordings held out).

Classifier	Hold-out	5-fold CV (mean ± std)	Temporal split	Unseen-scenario split
Extra Trees	0.9998	0.9998 ± 0.0001	0.9737	0.9546
Bagging	0.9998	0.9997 ± 0.0001	0.9737	0.9521
Random Forest	0.9998	0.9998 ± 0.0001	0.9737	0.9249
Decision Tree	0.9998	0.9996 ± 0.0002	0.9661	0.9523
KNN	0.9995	0.9996 ± 0.0001	0.9742	0.9239
XGBoost	0.9994	0.9996 ± 0.0001	0.9740	0.9544
Gradient Boosting	0.9983	0.9986 ± 0.0005	0.9735	0.9530
MLP Neural Net	0.9983	0.9984 ± 0.0004	0.9573	0.9348
AdaBoost	0.9947	0.9950 ± 0.0006	0.9750	0.9544
SVM (RBF)	0.9940	0.9944 ± 0.0005	0.9737	0.9601
Logistic Regression	0.9744	0.9760 ± 0.0016	0.8893	0.9602
Naive Bayes	0.8646	0.8663 ± 0.0026	0.5396	0.8165

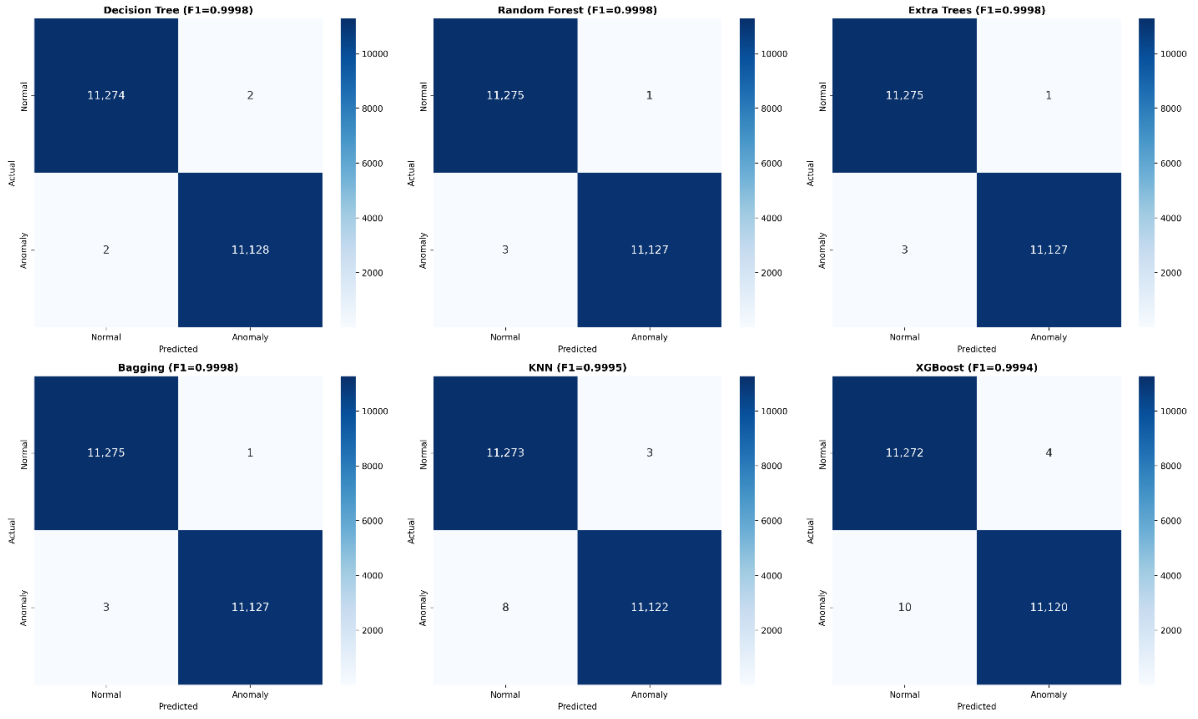


Fig. 4. Confusion matrices for the top six classifiers by F1-Score on the 22,406-record held-out test set. Near-diagonal concentration confirms minimal misclassification for all tree-based ensembles.

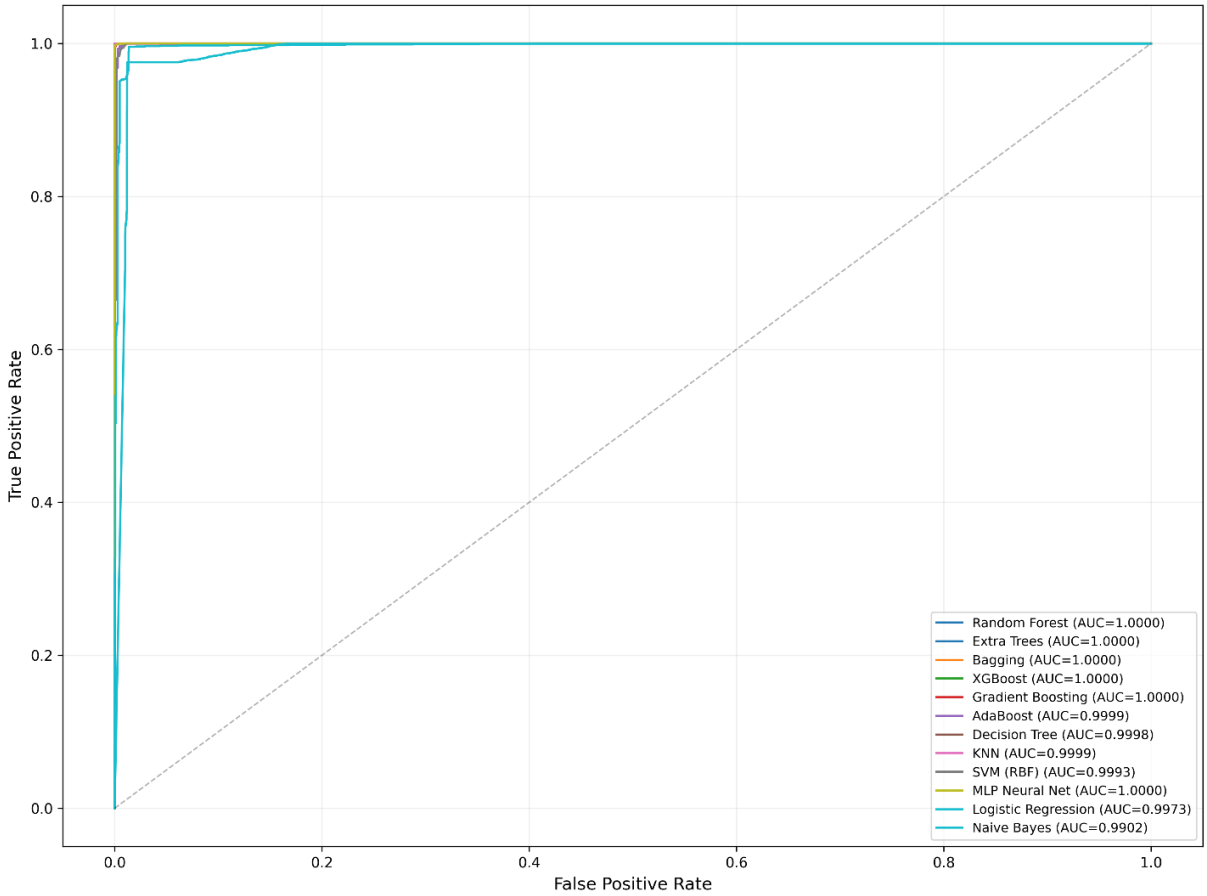


Fig. 5. ROC curves for all twelve supervised classifiers on the held-out test set. Six ensemble models achieve AUC-ROC = 1.0000. The dashed diagonal represents the random classification baseline

4.3. Unsupervised Clustering Performance

Thirteen clustering configurations were evaluated on the 15,000-observation random sample. The complete results are reported in Table 4.

Table 4. Unsupervised Clustering Results (sorted by Silhouette Score)

Configuration	N Clusters	Silhouette	DBI	CHI	ARI	NMI	Time (s)
Birch (k=3)	3	0.9529	0.303	2909.0	0.0000	0.0013	0.33
Agglomerative (Average)	2	0.9527	0.286	1906.6	−0.0000	0.0005	9.27
Agglomerative (Complete)	2	0.9527	0.286	1906.6	−0.0000	0.0005	—
Agglomerative (Ward)	2	0.9526	0.221	3000.9	0.0000	0.0008	8.04
GMM (k=2)	2	0.9526	0.221	3000.9	0.0000	0.0008	0.55
Birch (k=2)	2	0.9526	0.221	3000.9	0.0000	0.0008	0.32
K-Means (k=4)	4	0.3448	0.791	4105.9	0.3794	0.4018	1.22
K-Means (k=2)	2	0.3428	1.343	3114.4	0.4403	0.4437	1.12
K-Means (k=3)	3	0.3409	0.948	3837.6	0.4396	0.4426	1.21
Mini-Batch K-Means	2	0.3378	1.286	3101.4	0.3659	0.3936	0.05
DBSCAN (eps=2.0)	2	0.2551	0.635	261.6	−0.0003	0.0128	0.39
GMM (k=3)	3	0.1932	3.005	719.7	0.1397	0.2167	294.9
DBSCAN (eps=1.5)	10	−0.0340	1.242	875.7	0.0603	0.2438	0.36

The optimal cluster count from the elbow and Silhouette-vs-K analyses over K = 2–10 (Fig.6), consistent with the methodology of Schubert and Rousseeuw (2021), confirms k=2 as the natural K-Means partition. All Birch, Agglomerative, and GMM (k=2) configurations collapse to a degenerate solution — one dominant cluster containing approximately 99% of observations — producing high Silhouette Scores (≈ 0.953) through geometric compactness of that dominant cluster rather than meaningful class separation, as confirmed by ARI ≈ 0.000 (Fig.7). This result directly corroborates the systematic Silhouette inflation documented by Arbelaitz et al. (2013) for datasets with one dominant density region. Fig.7 visualizes all five metrics across all 13 configurations and makes the divergence immediately apparent: the top-six Silhouette configurations cluster on the right of the Silhouette panel but collapse to near-zero in the ARI and NMI panels. To assess sample dependence, all thirteen configurations were re-evaluated on ten independent 15,000-observation random samples (Table 5). The degenerate collapse of Birch, Agglomerative (Complete and Average linkage), and GMM (k=2) is fully robust: these configurations produce a dominant cluster containing more than 95% of observations, with ARI ≈ 0 and Silhouette between 0.94 and 0.97, in essentially every sample — confirming that the internal-external divergence is a structural property of the feature space rather than a sampling artifact. In contrast, the label alignment of K-Means (k=2) is bistable: across the ten samples it attains ARI = 0.200 ± 0.212 (range 0.000–0.441), reaching the Table 4 value only when optimization lands in the label-aligned optimum and collapsing to the degenerate geometric optimum in five of the ten samples. K-Means (k=4) is the only consistently label-aligned configuration (ARI = 0.387 ± 0.028, no degenerate runs).

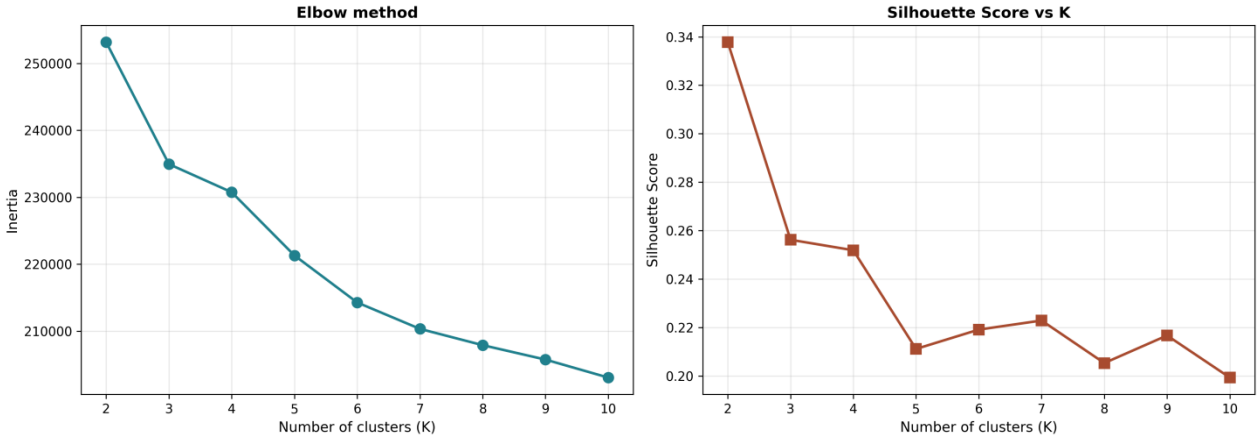


Fig. 6. Elbow method (left) and Silhouette-vs-K analysis (right) for Mini-Batch K-Means over K = 2–10 on the 15,000-observation sample; both criteria identify k = 2 as the natural partition.

Table 5. Clustering sample-sensitivity across ten independent 15,000-observation random samples: ARI (mean ± standard deviation and range), Silhouette (mean ± standard deviation), and number of samples yielding a degenerate partition (dominant cluster > 95% of observations).

Configuration	ARI (mean ± std)	ARI range	Silhouette (mean ± std)	Degenerate runs
Mini-Batch K-Means	0.390 ± 0.218	[0.137, 0.890]	0.318 ± 0.058	0/10
K-Means (k=4)	0.387 ± 0.028	[0.356, 0.444]	0.349 ± 0.004	0/10
K-Means (k=3)	0.372 ± 0.134	[0.000, 0.442]	0.403 ± 0.193	1/10
K-Means (k=2)	0.200 ± 0.212	[-0.000, 0.441]	0.654 ± 0.328	5/10
Agglomerative (Ward)	0.171 ± 0.276	[-0.000, 0.571]	0.771 ± 0.311	7/10
GMM (k=3)	0.089 ± 0.248	[0.000, 0.794]	0.360 ± 0.214	4/10
DBSCAN (eps=1.5)	0.065 ± 0.002	[0.063, 0.067]	-0.025 ± 0.007	0/10
GMM (k=2)	0.006 ± 0.008	[0.000, 0.022]	0.550 ± 0.263	8/10
DBSCAN (eps=2.0)	0.000 ± 0.000	[-0.000, 0.000]	0.252 ± 0.002	10/10
Birch (k=3)	0.000 ± 0.000	[-0.000, 0.000]	0.939 ± 0.016	10/10
Agglomerative (Average)	-0.000 ± 0.000	[-0.000, 0.000]	0.972 ± 0.006	10/10
Agglomerative (Complete)	-0.000 ± 0.000	[-0.000, 0.000]	0.965 ± 0.011	10/10
Birch (k=2)	-0.000 ± 0.000	[-0.000, 0.000]	0.960 ± 0.014	10/10

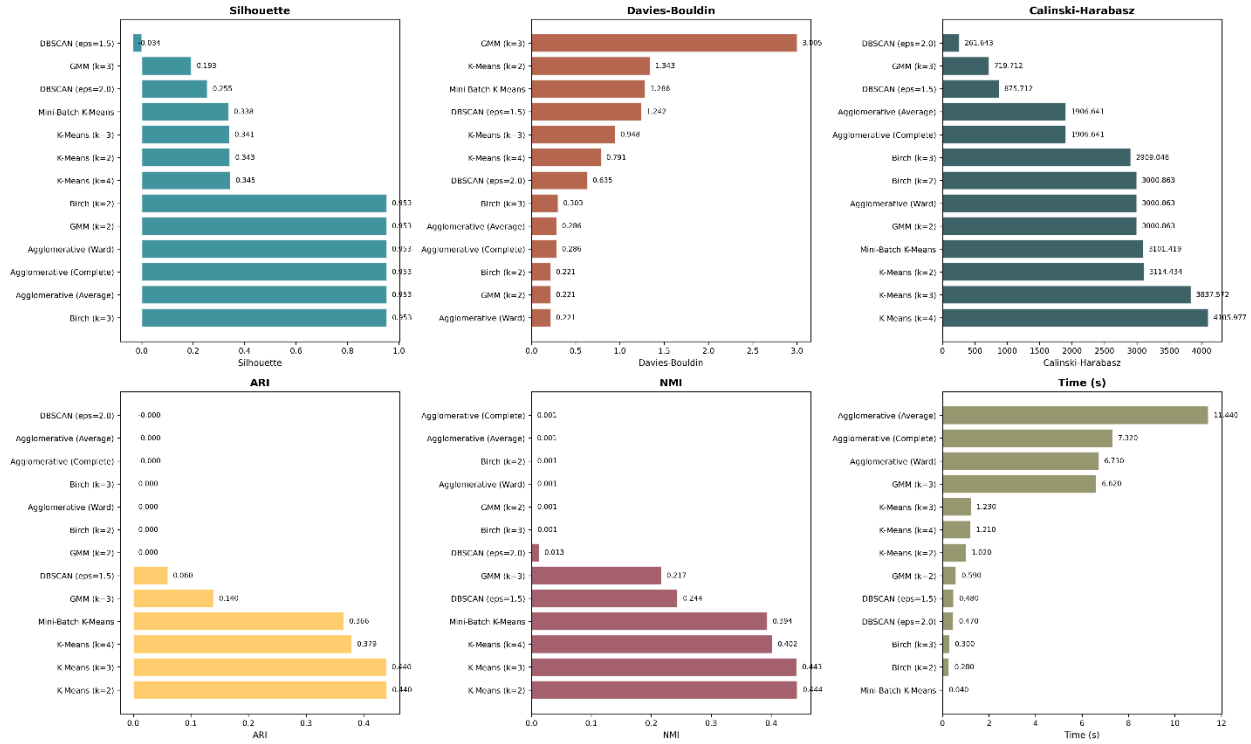


Fig. 7. Comparative clustering metrics across all 13 configurations: Silhouette Score, Davies-Bouldin Index, Calinski-Harabasz Index, ARI, NMI, and training time.

The inverse ranking between Silhouette and ARI for Birch/Agglomerative/GMM configurations is the central finding of the unsupervised analysis.

5. Discussion

5.1. Dominance of Tree-Based Ensembles

The performance of all tree-based classifiers, approaching theoretical upper bounds, is explained by the geometric structure of the 23-feature kinematic space, where four dominant features define a separating hyperplane isolating anomalous behavior with near-perfect precision. As established by Breiman (2001), Random Forest's variance reduction through ensemble averaging provides its greatest benefit when individual trees overfit — but when the boundary is as geometrically clean as observed here, even a single unpruned Decision Tree (F1 = 0.9998) achieves near-perfect partitioning. The negligible margin among Decision Tree, Bagging, and Extra Trees (ΔF1 < 0.0001) confirms that variance reduction through ensembling provides negligible additional benefit once the decision boundary is nearly optimal — fully consistent with [21], who demonstrated the same pattern on near-balanced binary IoT datasets. Feature importance analysis [17] reveals that speed (17.4%) and speed_time_ratio (16.3%) jointly drive more than half of the combined importance of the top four features, suggesting that instantaneous speed and its ratio to elapsed time are the most discriminative kinematic signals separating normal from anomalous vehicle behavior.

5.2. Optimal Deployment Trade-off

While the top four tree-based models are tied at F1 = 0.9998, Extra Trees is the preferred choice for operational AV monitoring deployment. As theoretically grounded by [19], Extra Trees' extreme threshold randomization produces a forest that is simultaneously more robust to correlated features and faster to train than standard Random Forest — and the present results confirm this, with Extra Trees achieving AUC-ROC = 1.0000, F1 = 0.9998, and a training time of 2.3 seconds compared to Bagging's 22.4 seconds at identical F1. Its training efficiency — roughly an order of magnitude higher than the comparably accurate bagging alternative in the documented environment (Section 3.6) — is operationally relevant in production AV monitoring contexts where models must be retrained frequently on updated telemetry streams. Logistic Regression, despite training in 4.1 seconds, produces an F1 = 0.9744 and Recall = 0.9623, meaning it would miss approximately 4 in 100 genuine anomaly events — an unacceptable false negative rate in safety-critical deployment [9,28,29]. Training times are strongly hardware- and implementation-dependent — in this environment XGBoost trains in 1.8 seconds, faster than Extra Trees, at marginally lower F1 (0.9994) — so efficiency claims should be read as environment-specific rather than intrinsic constants. A further deployment consideration arises from the unseen-scenario

results (Table 3): when novel driving scenarios are expected, the more heavily regularized models (Logistic Regression, SVM) generalize best, while unregularized tree ensembles lose the most performance.

5.3. The Internal-External Metric Divergence Problem

The most consequential finding is the severe divergence between internal and external clustering metrics for the top-Silhouette configurations. Birch (k=3), Agglomerative, and GMM (k=2) — the six highest-Silhouette configurations — all achieve ARI ≈ 0.000 and NMI < 0.002 despite Silhouette Scores of 0.952–0.953. These algorithms partition the standardized feature space along its principal geometric axis, creating one dominant compact cluster (≈99% of points) and marginal outlier groups, with the resulting high Silhouette Score reflecting the geometric compactness of the dominant cluster — not Normal-Anomaly separation. The collapse mechanism follows directly from the distributional structure of the standardized feature space rather than from any algorithmic defect. Cumulative and rolling-window features are heavily right-skewed: the vast majority of observations occupy one dense central mass, while a small fraction of records with extreme cumulative values form sparse peripheral tails, visible as the elongated margins of the PCA projection in Fig.8. Variance-minimizing criteria — Ward linkage, Birch's CF-tree radius threshold, and the likelihood surface of a two-component GMM — reduce total within-cluster variance most efficiently by isolating those few extremal points into marginal clusters and assigning the dense core to a single cluster, because any split through the core would increase within-cluster variance in both halves. The result is the observed ≈99%/1% partition, whose compact dominant cluster yields a high Silhouette value while carrying no information about the Normal–Anomaly boundary, which lies entirely inside the dense core (Fig.8, left panel). This is precisely the Silhouette inflation mechanism identified by [16], who demonstrated that Silhouette Score systematically overestimates cluster quality when one cluster contains a disproportionate share of observations. [15] further established that unsupervised anomaly detection algorithms optimizing internal geometric criteria consistently fail to align with ground-truth labels on datasets where the anomaly boundary is defined by label-dependent patterns rather than geometric outlier structure. Any operational system selecting clustering configurations by Silhouette Score alone would deploy a model with zero anomaly detection capability while falsely reporting excellent clustering quality — a direct safety-critical risk in AV monitoring system design. Fig. 8. provides direct visual evidence of this divergence: the ground-truth projection (left) and the Birch k=3 clustering projection (right) are visually indistinguishable, confirming that the algorithm has assigned cluster membership independently of the anomaly boundary.

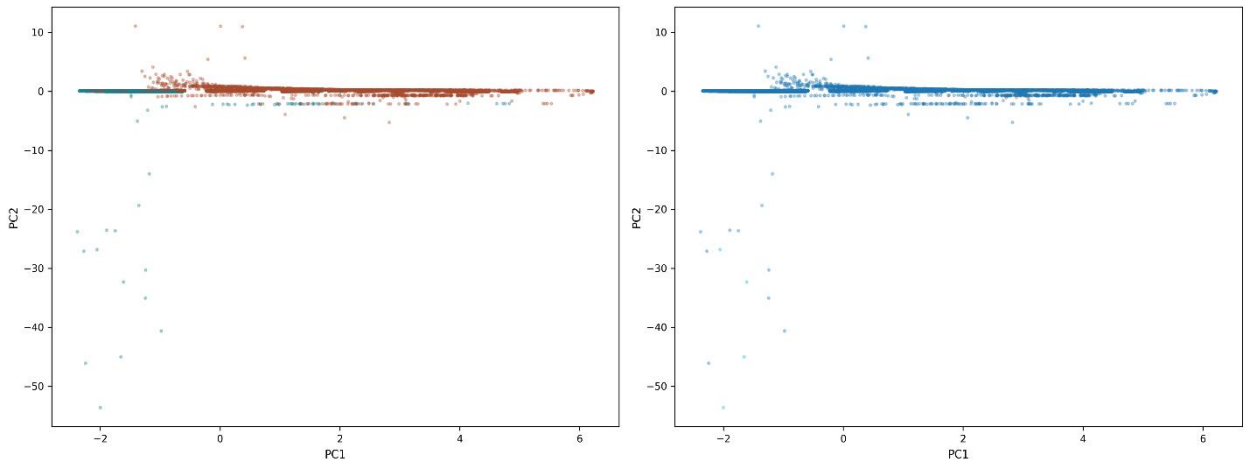


Fig. 8. Left: Ground-truth label projection in PCA space (Normal = teal, Anomaly = rust). Right: Best-Silhouette configuration — Birch (k=3, Silhouette = 0.9529, ARI = 0.000) — in PCA space. Visual indistinguishability between panels confirms complete failure to recover the anomaly boundary despite a near-perfect internal score.

5.4. K-Means as the Best Unsupervised Approximation

K-Means (k=2) achieves the highest external alignment (ARI = 0.4403, NMI = 0.4437) among all unsupervised methods, achieving approximately 44% label alignment. This partial recovery is attributable to the geometric correlation between K-Means' speed–distance-driven cluster axis and the actual anomaly structure — a centroid-based partitioning property analyzed by [30], who showed that K-Means performs best when classes are approximately spherically distributed in the scaled feature space. However, the ARI = 0.440 result confirms that the anomaly is not a simple geometric outlier but a complex, label-dependent kinematic pattern requiring supervised guidance to resolve, consistent with the supervised-versus-semi-supervised comparison of [22] showing that fully-supervised classifiers universally outperform label-agnostic alternatives. Increasing to k=3 provides no ARI improvement (0.4396 vs. 0.4403), and k=4 reduces ARI to 0.3794, suggesting on this sample that additional cluster fragmentation splits internally coherent Normal or Anomaly regions rather than resolving the boundary. The ten-sample sensitivity analysis (Section 4.3, Table 5) qualifies

this single-sample picture, however: the k=2 alignment is bistable across samples (ARI = 0.200 ± 0.212, degenerate in five of ten samples), whereas k=4 delivers lower but far more dependable alignment (ARI = 0.387 ± 0.028, never degenerate). For operational use, K-Means at k=4 with joint internal-external monitoring is therefore the more reliable unsupervised fallback, and any single-sample K-Means (k=2) result should be verified across samples before deployment.

5.5. Paradigm Selection Implications

The combined results establish a clear paradigm selection framework: when ground-truth labels are available, supervised classification is clearly the preferred choice under the conditions evaluated here, with Extra Trees providing the optimal balance of accuracy, AUC-ROC, and training efficiency. This finding is consistent with the ADBench benchmark (Han et al., 2022), which confirmed across 57 datasets that supervised methods universally outperform unsupervised ones when ground-truth labels are available. Unsupervised clustering should be reserved for exploratory analysis or scenarios where labeled data is entirely unavailable — and in those scenarios, K-Means (k=2) with joint internal-external evaluation should be preferred over Birch, Agglomerative, and GMM configurations. The 0.5597-point ARI gap between Decision Tree and K-Means (k=2) represents the measurable cost of label absence and should be explicitly quantified in any AV monitoring system design where this paradigm choice is made on operational grounds.

6. Conclusion

This study benchmarked twelve supervised classifiers and thirteen unsupervised clustering configurations for anomaly detection on 112,028 autonomous vehicle telemetry records across 23 engineered kinematic features. Tree-based ensembles achieved performance approaching theoretical upper bounds — Extra Trees, Random Forest, Bagging, and Decision Tree (F1 = 0.9998; stable at 0.9998 ± 0.0001 under 5-fold cross-validation) — driven by four dominant features (speed_time_ratio, distance, rolling_distance_mean, speed) accounting for 59.7% of Random Forest importance, confirming that the Normal-Anomaly boundary is highly learnable under supervised guidance; Extra Trees is identified as the optimal deployment choice due to its combination of AUC-ROC = 1.0000, F1 = 0.9998, and 2.3-second training time — roughly an order of magnitude faster than the comparably accurate bagging alternative in the documented environment — as empirically confirmed in ensemble comparisons by [21]. All unsupervised configurations fail to recover meaningful anomaly structure: the highest-Silhouette group (Birch k=3, Agglomerative, GMM k=2; Silhouette ≈ 0.953) yields ARI ≈ 0.000 — constituting a degenerate geometric partition rather than a semantically valid anomaly boundary — while the best-ARI configuration (K-Means k=2, ARI = 0.440) remains 56 percentage points below the weakest supervised classifier, a result consistent with the unsupervised evaluation findings. Robustness experiments qualify the absolute supervised figures without altering the ranking: strictly temporal and unseen-scenario partitioning reduce peak F1 to 0.9750 and 0.9602 respectively, quantifying the inflation contributed by window- and scenario-level leakage under random row-level splitting. This establishes joint internal-external metric reporting as a mandatory requirement for any clustering evaluation in safety-critical deployment contexts, and supervised classification as the preferred anomaly detection paradigm when labeled AV telemetry is available, under the default configurations, single-split evaluation, and static distribution assumption examined here [14].

Several limitations bound the generalizability of these conclusions and are stated here explicitly. First, all models were evaluated with default hyperparameters; no tuning or per-algorithm sensitivity analysis was performed, so reported figures are lower bounds on attainable performance per algorithm. Second, the dataset originates from a single simulation platform (MOBATSim [36]); although the temporal- and scenario-split experiments quantify within-dataset leakage effects directly, generalization to other simulators, to real vehicle fleets, and to drifting distributions remains unevaluated. Third, the comparison covers classical algorithms only — recent temporal Transformer and graph-based detectors [33-35] were not benchmarked — and the supervised paradigm's advantage presupposes labeled telemetry, an expensive asset in production fleets where label scarcity may favor semi-supervised alternatives on practical grounds. Fourth, computational cost is reported as training time in a single documented environment, without hardware-normalized comparison or inference-latency measurement.

Future work should investigate Optuna-based hyperparameter optimization for top-performing ensembles, semi-supervised frameworks leveraging the partial anomaly structure recovered by K-Means as a pre-labeling step in low-label-availability scenarios [22], and online learning extensions of Extra Trees for real-time telemetry streams subject to distributional shift. Future extensions should explore federated learning frameworks, to enable privacy-preserving distributed AV anomaly detection, incremental and concept-drift-aware learning strategies, for deployment on live telemetry streams, and deep learning architectures, validated against the kinematic features identified here [31,32]. To make these directions verifiable rather than enumerative, they are organized into three phased objectives: Phase 1 (optimization) — systematic hyperparameter search for the leading ensembles with paired significance testing against the default-configuration baselines established here; Phase 2 (extension) — inclusion of temporal Transformer [34,35] and graph-based [33] baselines under the identical dual-track metric protocol, evaluated under the scenario-split regime of Section 4.2; Phase 3 (deployment studies) — semi-supervised pre-labeling exploiting the partial structure recovered by K-Means, concept-drift-aware online variants of Extra Trees for live telemetry, and federated training for privacy-preserving fleet-scale detection.

All the Declarations and Statements

Author Contributions Statement
Nizirwan Anwar - Conceptualization, methodology, software, formal analysis, visualization, and original draft preparation.
Titik Khawa Abd. Rahman - Supervision, methodology, validation, and manuscript review.
Husna Sarirah Husin - Methodology, validation, and manuscript review and editing.
Aedah Abd. Rahman - Investigation, validation, and manuscript editing.
Swa Lee Lee - Data interpretation, validation, and manuscript review.
Muhammad Faisal - Software, data curation, and experimental validation.
Imam Asrowardi - Formal analysis, validation, and manuscript review.
Dewanto Rosian Adhy - Data analysis, visualization, and manuscript editing.
Tomy Ronaldi - Domain interpretation, validation, and manuscript review.
Arief Ichwani - Software verification, data curation, and manuscript editing.

All authors have read and agreed to the published version of the manuscript.

Conflict of Interest Statement
The authors declare no conflicts of interest.

Funding Declaration
This research received no external funding

Data Availability Statement
The dataset analyzed in this study is publicly available on IEEE DataPort: H. Pallapothu, G. Krishna, D. Sunil, T. Yadav, and S. Nagasundari, “Autonomous Vehicle dataset for Anomaly detection,” doi: 10.21227/0q64-5f96 [36]. The analysis code is available from the corresponding author on reasonable request.

Ethical Declarations
None. This study involves no human participants, animal subjects, or identifiable personal data; all experiments were conducted on a publicly available simulation dataset.

Acknowledgments
The authors would like to thank their respective institutions for the academic support and facilities provided throughout this study. The authors also acknowledge the provider of the Autonomous Vehicle Anomaly Detection dataset and the developers of the open-source software used in the analysis.

Declaration of Generative AI in Scholarly Writing
The authors declare that no generative artificial intelligence tools were used in the preparation, analysis, or writing of this manuscript.

Abbreviations
The following abbreviations are used in this manuscript:

AV – Autonomous Vehicle
FDIA – False Data Injection Attack
PCA – Principal Component Analysis
KNN – K-Nearest Neighbors
SVM – Support Vector Machine
RBF – Radial Basis Function
MLP – Multi-Layer Perceptron
GMM – Gaussian Mixture Model
DBSCAN – Density-Based Spatial Clustering of Applications with Noise
Birch – Balanced Iterative Reducing and Clustering using Hierarchies
ARI – Adjusted Rand Index
NMI – Normalized Mutual Information
DBI – Davies-Bouldin Index
CHI – Calinski-Harabasz Index
AUC-ROC – Area Under the Receiver Operating Characteristic Curve
CV – Cross-Validation

SHAP – SHapley Additive exPlanations
GNN – Graph Neural Network
IDS – Intrusion Detection System
CAN – Controller Area Network
UAV – Unmanned Aerial Vehicle
DVT – Denoising Variational Transformer;
GRU – Gated Recurrent Unit.

Appendix A/B/C…, with appendix title

None.

References

[1] S. Baccari, M. Hadded, H. Ghazzai, H. Touati, and M. Elhadef, "Anomaly Detection in Connected and Autonomous Vehicles: A Survey, Analysis, and Research Challenges," *IEEE Access*, vol. 12, pp. 19250–19276, 2024, doi: 10.1109/ACCESS.2024.3361829.

[2] J. R. V. Solaas, E. Mariconti, and N. Tuptuk, "Systematic Literature Review: Anomaly Detection in Connected and Autonomous Vehicles," *IEEE Transactions on Intelligent Transportation Systems*, vol. 26, no. 1, pp. 43–58, Jan. 2025, doi: 10.1109/TITS.2024.3495031.

[3] S. Rajapaksha, H. Kalutarage, M. O. Al-Kadri, A. Petrovski, G. Madzudzo, and M. Cheah, "AI-Based Intrusion Detection Systems for In-Vehicle Networks: A Survey," *ACM Comput. Surv.*, vol. 55, no. 11, pp. 1–40, Nov. 2023, doi:10.1145/3570954.

[4] D. Bogdoll, M. Nitsche, and J. M. Zollner, "Anomaly Detection in Autonomous Driving: A Survey," in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, IEEE, Jun. 2022, pp. 4487–4498. doi: 10.1109/CVPRW56347.2022.00495.

[5] Z. Zamanzadeh Darban, G. I. Webb, S. Pan, C. Aggarwal, and M. Salehi, "Deep Learning for Time Series Anomaly Detection: A Survey," *ACM Comput. Surv.*, vol. 57, no. 1, pp. 1–42, Jan. 2025, doi:10.1145/3691338.

[6] G. Pang, C. Shen, L. Cao, and A. Van Den Hengel, "Deep Learning for Anomaly Detection," *ACM Comput. Surv.*, vol. 54, no. 2, pp. 1–38, Mar. 2022, doi:10.1145/3439950.

[7] P. Dixit, P. Bhattacharya, S. Tanwar, and R. Gupta, "Anomaly detection in autonomous electric vehicles using AI techniques: A comprehensive survey," *Expert Syst.*, vol. 39, no. 5, Jun. 2022, doi: 10.1111/exsy.12754.

[8] R. Kaur, D. Gabrijelčič, and T. Klobučar, "Artificial intelligence for cybersecurity: Literature review and future research directions," *Information Fusion*, vol. 97, p. 101804, Sep. 2023, doi: 10.1016/j.inffus.2023.101804.

[9] K. Wang, A. Zhang, H. Sun, and B. Wang, "Analysis of Recent Deep-Learning-Based Intrusion Detection Methods for In-Vehicle Network," *IEEE Transactions on Intelligent Transportation Systems*, pp. 1–12, 2022, doi: 10.1109/TITS.2022.3222486.

[10] M. Althunayyan, A. Javed, and O. Rana, "A survey of learning-based intrusion detection systems for in-vehicle networks," *Computer Networks*, vol. 277, p. 112031, Mar. 2026, doi: 10.1016/j.comnet.2026.112031.

[11] H. Min *et al.*, "Toward interpretable anomaly detection for autonomous vehicles with denoising variational transformer," *Eng. Appl. Artif. Intell.*, vol. 129, p.107601, Mar. 2024, doi:10.1016/j.engappai.2023.107601.

[12] A. Sezgin, R. Keskin, and A. Boyacı, "Anomaly Detection in Unmanned Aerial Vehicle Telemetry Using Automated Machine Learning," *Fırat Üniversitesi Mühendislik Bilimleri Dergisi*, vol. 37, no. 2, pp. 699–709, Sep. 2025, doi: 10.35234/fumbd.1668498.

[13] M. H. J. Naeimi, A. Norouzi, and A. Abdi, "GRAD: Real-Time Gated Recurrent Anomaly Detection in Autonomous Vehicle Sensors Using Reinforced EMA and Multi-Stage Sliding Window Techniques," arXiv preprint arXiv:2510.23327, Oct. 2025. [Online]. Available: http://arxiv.org/abs/2510.23327

[14] S. Han, X. Hu, H. Huang, M. Jiang, and Y. Zhao, "ADBench: Anomaly Detection Benchmark," in*Advances in Neural Information Processing Systems (NeurIPS 2022), Datasets and Benchmarks Track*, 2022. doi: 10.48550/arXiv.2206.09426.

[15] Y. Fang *et al.*, "Anomaly diagnosis of connected autonomous vehicles: A survey," *Information Fusion*, vol. 105, p. 102223, May 2024, doi: 10.1016/j.inffus.2024.102223.

[16] H. Sträng and T. Dinh, "An upper bound on the silhouette evaluation metric for clustering," *Pattern Recognit.*, vol. 178, p. 113402, Oct. 2026, doi: 10.1016/j.patcog.2026.113402.

[17] Z. Li, Y. Zhu, and M. Van Leeuwen, "A Survey on Explainable Anomaly Detection," *ACM Trans. Knowl. Discov. Data*, vol. 18, no. 1, pp. 1–54, Jan. 2024, doi:10.1145/3609333.

[18] G. Ciaburro, "Machine fault detection methods based on machine learning algorithms: A review," *Mathematical Biosciences and Engineering*, vol. 19, no. 11, pp. 11453–11490, 2022, doi: 10.3934/mbe.2022534.

[19] L. Grinsztajn, E. Oyallon, and G. Varoquaux, "Why do tree-based models still outperform deep learning on tabular data?," Jul. 2022, [Online]. Available: http://arxiv.org/abs/2207.08815

[20] V. Borisov, T. Leemann, K. Seßler, J. Haug, M. Pawelczyk, and G. Kasneci, "Deep Neural Networks and Tabular Data: A Survey," *IEEE Trans. Neural Netw. Learn. Syst.*, vol. 35, no. 6, pp. 7499–7519, Jun. 2024, doi: 10.1109/TNNLS.2022.3229161.

[21] M. Saied, S. Guirguis, and M. Madbouly, "A comparative analysis of using ensemble trees for botnet detection and classification in IoT," *Sci. Rep.*, vol. 13, no. 1, p. 21632, Dec. 2023, doi: 10.1038/s41598-023-48681-6.

[22] Y. Dong, K. Chen, Y. Peng, and Z. Ma, "Comparative Study on Supervised versus Semi-supervised Machine Learning for Anomaly Detection of In-vehicle CAN Network," in *2022 IEEE 25th International Conference on Intelligent Transportation Systems (ITSC)*, IEEE, Oct. 2022, pp. 2914–2919. doi: 10.1109/ITSC55140.2022.9922235.

[23] Y. Dong, L. Zhang, H. Farah, A. Zgonnikov, and B. van Arem, "Data-Driven Semi-Supervised Machine Learning with Safety Indicators for Abnormal Driving Behavior Detection," *Transportation Research Record: Journal of the Transportation Research Board*, vol. 2679, no. 5, pp. 319–334, May 2025, doi: 10.1177/03611981241306752.
[24] A. Fernández, S. García, M. Galar, R. C. Prati, B. Krawczyk, and F. Herrera, *Learning from Imbalanced Data Sets*. Cham: Springer International Publishing,2018. doi:10.1007/978-3-319-98074-4.
[25] R. Sauber-Cole and T. M. Khoshgoftaar, "The use of generative adversarial networks to alleviate class imbalance in tabular data: a survey," *J. Big Data*, vol. 9, no. 1, p. 98, Aug. 2022, doi: 10.1186/s40537-022-00648-6.
[26] A. L. Suárez-Cetrulo, D. Quintana, and A. Cervantes, "A survey on machine learning for recurring concept drifting data streams," *Expert Syst. Appl.*, vol. 213, p. 118934, Mar. 2023, doi: 10.1016/j.eswa.2022.118934.
[27] B. Bischl *et al.*, "Hyperparameter optimization: Foundations, algorithms, best practices, and open challenges," *WIREs Data Mining and Knowledge Discovery*, vol. 13, no. 2, Mar. 2023, doi: 10.1002/widm.1484.
[28] T. J. Chengula, J. Mwakalonge, S. Siuhi, and G. Comert, "Improving Road Safety with Ensemble Learning: Detecting Driver Anomalies Using Vehicle Inbuilt Cameras," 2023. doi: 10.2139/ssrn.4573044.
[29] V. P. Chellapandi, L. Yuan, C. G. Brinton, S. H. Żak, and Z. Wang, "Federated Learning for Connected and Automated Vehicles: A Survey of Existing Approaches and Challenges," *IEEE Transactions on Intelligent Vehicles*, vol. 9, no. 1, pp. 119–137, Jan. 2024, doi: 10.1109/TIV.2023.3332675.
[30] E. Schubert and P. J. Rousseeuw, "Fast and eager k-medoids clustering: O(k) runtime improvement of the PAM, CLARA, and CLARANS algorithms," *Inf. Syst.*, vol. 101, p. 101804, Nov. 2021, doi: 10.1016/j.is.2021.101804.
[31] M. Landauer, S. Onder, F. Skopik, and M. Wurzenberger, "Deep learning for anomaly detection in log data: A survey," *Machine Learning with Applications*, vol. 12, p. 100470, Jun. 2023, doi: 10.1016/j.mlwa.2023.100470.
[32] I. Makris *et al.*, "A comprehensive survey of Federated Intrusion Detection Systems: Techniques, challenges and solutions," *Comput. Sci. Rev.*, vol. 56, p.100717, May 2025, doi:10.1016/j.cosrev.2024.100717.
[33] X. Ma, J. Wu, S. Xue, J. Yang, C. Zhou, Q. Z. Sheng, H. Xiong, and L. Akoglu, "A Comprehensive Survey on Graph Anomaly Detection with Deep Learning," *IEEE Transactions on Knowledge and Data Engineering*, vol. 35, no. 12, pp. 12012–12038, Dec. 2023, doi: 10.1109/TKDE.2021.3118815.
[34] J. Xu, H. Wu, J. Wang, and M. Long, "Anomaly Transformer: Time Series Anomaly Detection with Association Discrepancy," in *Proc. International Conference on Learning Representations (ICLR)*, 2022. [Online]. Available: https://openreview.net/forum?id=LzQQ89U1qm_
[35] S. Tuli, G. Casale, and N. R. Jennings, "TranAD: Deep Transformer Networks for Anomaly Detection in Multivariate Time Series Data," *Proceedings of the VLDB Endowment*, vol. 15, no. 6, pp. 1201–1214, 2022, doi: 10.14778/3514061.3514067.
[36] H. Pallapothu, G. Krishna, D. Sunil, T. Yadav, and S. Nagasundari, "Autonomous Vehicle dataset for Anomaly detection," *IEEE Dataport*, Jan. 2025, doi: 10.21227/0q64-5f96.

Authors' Profiles



Nizirwan Anwar is an Associate Professor in the Department of Informatics Engineering, Faculty of Computer Science, Universitas Esa Unggul, Jakarta, Indonesia. Obtained a bachelor's degree in physics from the Faculty of Mathematics and Natural Sciences, Universitas Padjadjaran (1989), and later pursued a master's degree specialising in Image Processing at the Faculty of Engineering, Universitas Indonesia (1995). He is currently undertaking a Doctor of Philosophy in Information and Communication Technology at Asia e University, Malaysia. He is a Professional Engineer (IPM), ASEAN Engineer, and APEC Engineer. His research focuses on artificial intelligence, machine learning, deep learning, Internet of Things (IoT), data science, software engineering, information transportation systems, and digital transformation. His work emphasizes intelligent systems and data-driven solutions for industrial, educational, and smart technology applications.



Titik Khawa Abd. Rahman received a B.Sc. in Electrical and Electronics Engineering from Loughborough University of Technology, UK. She holds a Ph.D in Power Systems from the University of Malaya. With extensive experience in higher education, research, and academic leadership. She previously served as the Dean of the School of Science and Technology, Asia e University (AeU), Malaysia. Her research interests include artificial intelligence, machine learning, data science, biologically inspired optimization, machine learning, power system analysis and operation, and smart grid.



Aedah Abd Rahman received both Bachelor of Computer Science and Master of Computer Science from University of Malaya (UM), Kuala Lumpur, Malaysia. She holds a Ph.D. in Computer Science from Universiti Teknologi Malaysia (UTM)/Malaysia University of Technology. With over 25 years of experience in private universities, she has held various academic and leadership positions. She currently serves as the Dean of School of Science & Technology (SST) at Asia e University (AeU). Her areas of expertise include artificial intelligence, machine learning, computer science, software engineering, process improvement, quality assurance, quality engineering, project management, Open and Distance Learning (ODL), digital and e-learning, information systems and ICT.



Husna Sarirah Husin is a seasoned academic and researcher with over 25 years of experience spanning both academia and industry, currently serving as a Senior Lecturer at Taylor's University, Malaysia. Her career reflects a dynamic blend of teaching, research, and leadership across multiple institutions, supported by a strong foundation in computer science and information management, including a PhD from RMIT University focused on user behaviour analytics in web environments. Actively engaged in interdisciplinary research, she has contributed extensively to areas such as machine learning, data mining, IoT, and intelligent systems, with numerous publications, grants, and international collaborations. Beyond research, she plays a vital role in academic governance, postgraduate supervision, and professional communities, including IEEE and MBOT, positioning her as a key contributor to advancing digital innovation and technology-driven education in the region.



Swa Lee Lee is a committed education professional with over 20 years of experience in academic administration, postgraduate education, student and alumni engagement, and international collaboration. In her current role as Director for the School of Graduate Studies (SGS) and Academic Relations at Asia e University (AeU), she contributes to initiatives in academic governance, research development, and global partnerships. She values lifelong learning and supports the integration of digital tools to enhance research and academic processes. Her interest in work-life harmony reflects a sincere commitment to human-centered leadership. Through her involvement in national and international education bodies, she strives to contribute meaningfully to academic growth and policy development. Guided by the belief that education should nurture both the intellect and the spirit, she aims to help create environments where learning, empathy, and continuous improvement are encouraged.



Muhammad Faisal received his Bachelor of Information Systems (S.SI) from the Information Systems Department at STMIK Profesional, Makassar, Indonesia, in 2011, and his Master of Engineering (M.T) from the Informatics Engineering Department at Universitas Hasanuddin, Gowa, Indonesia, in 2014. He holds a Ph.D degree from the Doctoral Programme in the School of Science and Technology, Asia E University Malaysia. He is currently a lecturer and researcher at the Department of Informatics, University of Muhammadiyah Makassar, Indonesia. His research interests include Artificial Intelligence, Data Mining, Decision Support Systems, Image Processing, Deep Learning, and Machine Learning.



Imam Asrowardi is a lecturer in the Technology Study Program Internet Engineering, Lampung State Polytechnic. His S1 degree was obtained from the Faculty of Science and Technology, Yogyakarta University of Technology (2003), and his S2 education was completed at the Diponegoro University Postgraduate Program (2011). He has been actively teaching since 2005 and involved in various professional organizations at the national (IAII, APTIKOM, PII) and international (IAENG) levels, both as a member and administrator. Since 2005 it has obtained various industry certifications in the field of information and network technology, including MTCINE, MTCRE, MTCWE, MTCUME, MTCIPv6E, MTA (Networking), CCC-BDF, Cisco Academy Trainer (CCNA, CyberOps Associate, DevNet Associate, CCNP), MikroTik Academy Trainers, Competency Programmers and Assessors (BNSP), National Procurement Experts (BNSP), and CCST Networking.



Dewanto Rosian Adhy completed a bachelor's degree in Control Engineering at Institut Teknologi Bandung (ITB), followed by a master's degree at the same institution. Currently, the author is pursuing a Doctoral degree at the Fakulti Teknologi Maklumat dan Komunikasi (FTMK), Universiti Teknikal Malaysia Melaka (UTeM). Their research interests focus on Artificial Intelligence (AI) and the Internet of Things (IoT). The author has contributed to several publications and research projects in the fields of IoT and Machine Learning, demonstrating a strong commitment to advancing knowledge in these rapidly evolving areas.



Tomy Ronaldi is a Navy Officer who currently serves as the Commander of the Center for Hydro-Oceanography Education (Pusdikhidros). He is currently pursuing his Doctorate at the Indonesian Defense University of Defense Science Study Program with a concentration in Defense Technology. The master's degree was obtained while participating in Seskoal Education in the field of Marine Operations and a bachelor's degree obtained from PIP Semarang. In addition, he also took an International Professional certification in the field of Hydrography in India in the Long Hydrography Course program for 11 months. Working in the field of Hydro-Oceanographic Survey, where the data obtained will be used as data in the processing of seamap products that are currently used in the maritime industry in ensuring the safety of shipping. He is also involved in the preparation of the IBSC IHO Category B (Cat-B) SURVEYOR EDUCATION LEARNING CURRICULUM with international standards that have been set and used as a curriculum at the Pusdikhidros, so that the certificates issued are

internationally valid and are recognized by the International Hydrographic Organization (IHO). His research interests are in the field of underwater sensor concepts for underwater vehicle monitoring.



Arief Ichwani, a dedicated academic and IT professional currently serving as a Lecturer (Lektor) in the Department of Informatics Engineering at Universitas Esa Unggul, Jakarta, where he has been a faculty member since 2017. He earned his bachelor's degree in informatics engineering from Universitas Siliwangi in 2011 and completed his master's degree in informatics at the Institut Teknologi Bandung (ITB) in 2016. With practical experience extending to institutional IT consulting for the Ministry of Public Works and an active BNSP Associate Data Scientist certification, his academic and research interests center on enterprise architecture, system audits, and intelligent computing. He has built a prolific research portfolio focused on the digital transformation of public systems, smart culture, and marketplace sentiment analysis, alongside leading extensive community development initiatives leveraging internet technologies to resolve local public health and digital literacy challenges.

How to cite this paper

Nizirwan Anwar, Titik Khawa Abdul Rahman, Aedah Abd Rahman, Swa Lee Lee, Muhammad Faisal, Dewanto Rosian Adhy, Tomy Ronaldi, Arief Ichwani, "Benchmarking Supervised and Unsupervised Paradigms for Anomaly Detection in Autonomous Vehicle Telemetry: The Pitfalls of Internal Clustering Metrics", International Journal of Wireless and Microwave Technologies(IJWMT), Vol.16, No.5, pp. 89-107, 2026. DOI:10.5815/ijwmt.2026.05.06