Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Detection of GNSS Spoofing and Jamming Using Signal Observables and Integrity Assessment

Sebastián Čikovský

sebastian.cikovsky@tuke.sk

Technical University of Košice

Patrik Šváb

Technical University of Košice

Peter Dzurovčin

Technical University of Košice

Peter Hanák

Technical University of Košice

Research Article

Keywords:

Posted Date: November 12th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-9626019/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

Detection of GNSS Spoofing and Jamming Using Signal Observables and Integrity Assessment

Sebastián Čikovský[1], Patrik Šváb[1] , Peter Dzurovčin[1], Peter Hanák[1]

[1]Faculty of Aeronautics, Technical University of Kosice, Department of Air Transport Management, 04001, Kosice

Corresponding author: Sebastián Čikovský, sebastian.cikovsky@tuke.sk

Abstract Global Navigation Satellite Systems (GNSS) are essential for positioning, navigation, and timing, yet their signals remain highly vulnerable to interference. Jamming and spoofing can significantly degrade receiver performance and compromise integrity, particularly under degraded signal conditions where benign distortions may resemble malicious behavior.This study proposes an integrity-aware, two-stage framework for GNSS interference detection based on signal-level observables. In the first stage, an integrity-gating mechanism evaluates whether the received signal provides sufficient evidence for reliable interpretation. Only accepted observations are forwarded to a classification stage that distinguishes between nominal, spoofing, and jamming conditions. By separating evidence validation from interference discrimination, the approach reduces the risk of overconfident decisions under uncertain signal conditions. Experimental evaluation on representative GNSS datasets demonstrates improved robustness and more reliable decision behavior while maintaining competitive detection performance. The proposed framework provides a practical basis for receiver-level integrity monitoring in operational GNSS environments.

Keywords GNSS; jamming; spoofing; meaconing; interference detection; signal-level observables; integrity; monitoring; navigation signals

1. Introduction

Global Navigation Satellite Systems (GNSS) constitute a fundamental technological backbone for modern positioning, navigation, and timing (PNT) services. Originally designed for navigation purposes, GNSS technologies have gradually become deeply integrated into a wide spectrum of critical infrastructures, including transportation systems, aviation, maritime operations, telecommunications networks, financial time synchronization, and energy grid management [1-2]. In addition to these traditional applications, GNSS-derived positioning and timing information now plays a crucial role in emerging cyber-physical systems such as autonomous vehicles, unmanned aerial systems, intelligent transportation infrastructures, and large-scale sensor networks [3]. The reliability and integrity of GNSS signals therefore directly affect the safe operation of many modern technological systems and services [4]. The widespread reliance on GNSS has also exposed an inherent vulnerability: satellite navigation signals received on Earth are extremely weak, typically several orders of magnitude below the noise floor. [5] This characteristic makes GNSS particularly susceptible to intentional or unintentional radio-frequency interference [6-7]. Among the most significant threats are jamming, spoofing, and meaconing.

Jamming attacks attempt to deny service by transmitting high-power signals that overwhelm legitimate satellite transmissions, effectively preventing receivers from acquiring or tracking GNSS signals. Spoofing attacks, on the other hand, aim to mislead receivers by generating counterfeit GNSS signals that mimic authentic satellite transmissions while manipulating navigation parameters [7-8]. A

successful spoofing attack can cause a receiver to compute incorrect position, velocity, or timing information without necessarily raising immediate alarms [9]. Meaconing represents a related threat in which authentic GNSS signals are captured and re-broadcast with delay or modification [4], thereby inducing position or timing errors while preserving certain signal characteristics that may appear legitimate to the receiver.

These interference mechanisms pose serious risks not only for civilian navigation but also for safety-critical and autonomous systems that rely on accurate and continuous PNT information. For instance, interference affecting aviation navigation systems, maritime traffic monitoring, or autonomous vehicle localization could lead to severe operational disruptions or safety hazards. Furthermore, the increasing integration of GNSS with automated decision-making systems amplifies the consequences of corrupted positioning data. As a result, the development of reliable methods for detecting and classifying GNSS interference has become a major research topic within navigation security and resilience [10]. Traditional approaches to GNSS interference detection have relied primarily on signal processing techniques and rule-based monitoring methods implemented at the receiver level. Examples include monitoring signal power levels, analyzing correlation distortions, evaluating carrier-to-noise ratios, or detecting inconsistencies in navigation solutions [11]. While these methods can be effective in certain scenarios, they often rely on manually designed thresholds or heuristics and may struggle to generalize across diverse interference conditions. Moreover, sophisticated spoofing or composite interference attacks can exploit weaknesses in rule-based detection schemes, making it increasingly difficult to rely solely on classical signal-processing indicators.

In recent years, machine learning (ML) has emerged as a promising approach for GNSS interference detection and classification [12]. By leveraging statistical learning techniques, ML models can automatically identify complex patterns in receiver observables or signal features that may indicate the presence of interference. Numerous studies have explored supervised learning algorithms such as support vector machines, random forests, neural networks, and deep learning architectures for detecting spoofing or jamming events [9][13-14]. These approaches typically rely on feature vectors derived from GNSS receiver measurements, including navigation-solution outputs, signal quality metrics, carrier-to-noise ratios, pseudorange statistics, or spectral characteristics of received signals. Within the ML-based literature, two principal methodological directions can be identified. The first group of approaches focuses on receiver-level or tabular data representations, where engineered features derived from GNSS observables are used as inputs to conventional machine-learning models [15]. These approaches are often computationally efficient, interpretable, and compatible with existing receiver monitoring outputs. The second direction involves signal-level or image-based methods, where raw intermediate-frequency samples, spectrograms, or other signal representations are processed using deep neural networks, including convolutional architectures. While signal-level methods can capture detailed temporal or spectral structures associated with interference signals, they typically require significantly larger datasets, higher computational resources, and direct access to raw receiver signal streams.

Despite the rapid growth of machine-learning research in GNSS security, several important limitations remain within the existing literature. First, a substantial portion of published work focuses on binary classification tasks, such as distinguishing between authentic signals and spoofing or identifying the presence or absence of interference. While these formulations provide valuable insights, they do not fully reflect the complexity of real-world interference environments, where multiple attack types may occur simultaneously or exhibit overlapping characteristics [14]. In operational contexts, distinguishing between

different interference mechanisms, such as jamming, spoofing, or meaconing, can be crucial for selecting appropriate mitigation strategies. Second, evaluation methodologies used in many studies may lead to overly optimistic estimates of model performance. GNSS datasets are often constructed from long measurement recordings that are subsequently segmented into shorter windows for machine-learning processing. If these segments are randomly distributed between training and test sets, highly correlated samples originating from the same measurement scenario may appear in both partitions [14-15]. Such scenario leakage can artificially inflate classification accuracy because the model effectively learns scenario-specific patterns rather than generalizable interference characteristics.

Third, many studies emphasize the development of increasingly complex models without systematically investigating which groups of observables carry the discriminative information required for reliable attack recognition. GNSS receivers provide a large variety of potential measurements, including navigation-solution outputs, signal quality indicators, monitoring metrics, and raw observation data [16]. However, the relative importance of these observable families for interference recognition remains insufficiently explored in the literature. Without controlled ablation studies, it is difficult to determine whether model performance stems from meaningful physical indicators of interference or from incidental correlations present in specific datasets [17]. Finally, the research community still lacks open, reproducible multiclass benchmarks that would enable systematic comparison of machine-learning approaches under strict evaluation conditions. Although several datasets containing GNSS interference recordings have been released in recent years, many studies rely on proprietary data or employ evaluation protocols that are difficult to reproduce. The absence of standardized multiclass benchmarks limits the ability to rigorously compare models and hinders progress toward robust, operationally deployable detection frameworks.

To address these challenges, this paper introduces a reproducible multiclass GNSS attack recognition benchmark based on the open JammerTest 2024 dataset. The benchmark focuses on four classes representing distinct interference scenarios: jamming, spoofing, meaconing, and combined jamming–spoofing attacks. Instead of relying on random sample-level splitting, the evaluation framework adopts strict scenario-level validation protocols, including grouped repeated splits and leave-one-scenario-out testing. This approach ensures that the training and test sets contain data originating from different acquisition scenarios, thereby reducing the risk of leakage and providing a more realistic estimate of model generalization. The study further investigates the role of different GNSS observable families in attack recognition by constructing feature representations from three principal data sources available in the dataset: navigation-solution outputs, monitoring outputs, and RINEX-derived observation data. Compact statistical summaries are extracted from these measurements to form interpretable tabular representations suitable for conventional machine-learning algorithms. To ensure fair evaluation, potential shortcut variables or metadata that could encode scenario information are excluded from the feature construction process.

The resulting benchmark is evaluated using several interpretable machine-learning baselines under strict scenario-level validation. In addition to reporting overall classification performance, the study performs family-wise ablation experiments that systematically evaluate the predictive contribution of different observable groups. This analysis allows the identification of which types of GNSS receiver outputs provide the most informative signals for distinguishing between interference mechanisms.

The main contributions of this work can be summarized as follows. First, the paper formulates GNSS interference recognition as a multiclass machine-learning task covering jamming, spoofing,

meaconing, and combined attack scenarios. Second, it introduces a strict anti-leakage evaluation framework tailored to scenario-structured GNSS datasets. Third, it provides a reproducible benchmark pipeline based on the open JammerTest 2024 dataset that facilitates systematic evaluation of machine-learning approaches. Fourth, the study compares interpretable tabular baselines under conservative validation conditions. Fifth, it conducts observable-family ablation analyses to determine which GNSS measurements carry the most discriminative information for attack recognition. Finally, the results demonstrate that robust multiclass classification performance can be achieved using lightweight receiver-level representations and reveal that navigation-solution observables dominate predictive performance under strict scenario-level evaluation. By providing a transparent evaluation framework and a detailed analysis of observable contributions, the proposed benchmark contributes to a deeper understanding of how machine learning can be effectively applied to GNSS interference recognition while minimizing methodological pitfalls related to data leakage and evaluation bias

2. Methodology

This section presents the methodological framework adopted in the study, including the dataset, signal sources, windowing strategy, statistical feature extraction, anti-leakage filtering, feature-set ablation, classification models, and evaluation protocols. The methodology was designed to support a transparent and reproducible analysis of multiclass GNSS attack recognition under strict scenario-level validation.

2.1. Dataset and Problem Formulation

This study uses the open Jammertest 2024 dataset released on Zenodo under the title *GNSS Dataset Under Jamming, Spoofing, and Meaconing Conditions (JammerTest 2024)* [18]. The Zenodo record states that the archive contains controlled jamming, spoofing, meaconing, and combined attacks, collected under both stationary and dynamic receiver conditions, and includes raw UBX files, RINEX observations, RF monitoring, and navigation solutions. In the present work, the primary feature sources were the scenario-level files mon_{rf}, nav_{pvt}, and rinex [18].

Let the scenario set be:

$$\mathcal{S} = \{S_1, S_2, \dots, S_M\}, \tag{1}$$

where each scenario S_m is associated with a class label:

$$y_m \in \mathcal{Y} = \{\text{ jamming, spoofing, meaconing, jamming} \setminus \text{spoofing }\}. \tag{2}$$

Each scenario is partitioned into a set of fixed-size windows,

$$S_m \to \{W_{m,1}, W_{m,2}, \dots, W_{m,N_m}\} \tag{3}$$

and each window is represented by a feature vector:

$$\boldsymbol{x_{m,n}} \in R^d. \tag{4}$$

The primary task is multiclass attack recognition, i.e. the estimation of a mapping:

$$f: R^d \rightarrow \mathcal{Y}. \tag{5}$$

that predicts the attack type from window-level observables.

2.2. Signal and Observation Sources

The dataset provides multiple GNSS-related file families, including RINEX observations, receiver monitoring information, and navigation-solution outputs [18]. RINEX 3.05 is the current IGS/RTCM exchange specification for GNSS observation and navigation files and defines the general structure of observation and navigation message files. The RINEX 3.05 specification explicitly describes the format as consisting of ASCII file types with header and data sections, including observation and navigation message files [19].

The nav_{pvt} and mon_{rf} sources used in this work are consistent with message families documented in the u-blox ZED-F9P Interface Description, which serves as the technical reference for the UBX protocol and receiver interface of the high-precision GNSS module used in many open GNSS workflows. The interface description is the appropriate source for the semantics of UBX navigation and monitoring messages, including navigation solution reporting and RF-monitoring-related message structures [20].

Accordingly, three feature families were defined:

- NAV family: descriptors extracted from nav_{pvt},
- MON family: descriptors extracted from mon_{rf},
- RINEX family: descriptors extracted from numeric variables in rinex.

2.3. Window Construction and Scenario-Level Grouping

Because windows extracted from a single scenario are not statistically independent, the scenario was treated as the fundamental grouping unit throughout the evaluation. For each scenario, a numeric anchor table was selected in the order of nav_{pvt}, monrf, and rinex, and subsequently segmented into non-overlapping windows of fixed length L=120L = 120L=120 rows.

For a sequence of length T_m, the n-th window was defined as:

$$W_{m,n} = \{\, t \mid t \in [(n-1)L,\ min(nL, T_m)) \,\}. \tag{6}$$

A minimum valid window length constraint was enforced, and scenarios with excessive numbers of candidate windows were uniformly subsampled to reduce dominance by a small subset of longer recordings.

All downstream train/test partitions were performed at scenario level:

$$\mathcal{S}_{\text{tan}} \cap \mathcal{S}_{\text{ts}} = \emptyset. \tag{7}$$

This design was introduced explicitly to prevent window-level leakage, i.e. overly optimistic evaluation caused by placing temporally or structurally related windows from the same scenario in both training and testing subsets.

2.4. Statistical Feature Extraction

For any numeric variable z(t) within a given window, let the valid numeric sample vector be:

$$\boldsymbol{z} = [z_1, z_2, \dots, z_N]. \tag{8}$$

For each such variable, a compact descriptor set was computed.

The mean and standard deviation were defined as:

$$\mu(\boldsymbol{z}) = \frac{1}{N}\sum_{i=1}^{N} z_i, \tag{10}$$

$$\sigma(z) = \sqrt{\frac{1}{N}\sum_{i=1}^{N}(z_i - \mu)^2}. \tag{11}$$

The empirical quantile-based descriptors were:

$$p_{10} = Q_{0.10}(z), \qquad p_{50} = Q_{0.50}(z), \qquad p_{90} = Q_{0.90}(z), \tag{12}$$

together with the interquartile range:

$$\mathrm{IQR}(z) = Q_{0.75}(z) - Q_{0.25}(z). \tag{13}$$

The extrema were:

$$z_{\min} = \min(\boldsymbol{z}), \quad z_{\max} = \max(\boldsymbol{z}). \tag{14}$$

To encode local temporal drift, a least-squares slope was estimated. Let $u_i = i - \bar{\imath}$ be the centered sample index. Then:

$$\beta(\boldsymbol{z}) = \frac{\sum_{i=1}^{N} u_i(z_i - \bar{z})}{\sum_{i=1}^{N} u_i^2}. \tag{15}$$

Missingness was summarized by:

$$r_{\text{miss}} = 1 - \frac{N}{N_{\text{orig}}}. \tag{16}$$

where N_{orig} is the original window size before invalid numeric values were removed.

Thus, each raw variable was mapped to:

$$\phi(\boldsymbol{z}) = [\mu, \sigma, z_{\min}, z_{\max}, p_{50}, \text{IQR}, p_{10}, p_{90}, \beta, r_{\text{miss}}]. \tag{17}$$

The final feature vector for one window was constructed by concatenating the descriptors of all retained variables:

$$\boldsymbol{x_{m,n}} = \left[\phi\left(\boldsymbol{z_{m,n}^{(1)}}\right), \phi\left(\boldsymbol{z_{m,n}^{(2)}}\right), \ldots, \phi\left(\boldsymbol{z_{m,n}^{(K)}}\right)\right]. \tag{18}$$

This representation was chosen because it is transparent, reproducible, and directly compatible with heterogeneous numeric sources derived from RINEX and UBX-level outputs [18 - 20].

2.5. Anti-Leakage Filtering

A central methodological objective was to eliminate variables that could encode scenario identity or dataset structure rather than interference behavior. Therefore, the following were removed before model fitting:

1. metadata-derived variables,
2. source-presence flags,
3. file-shape descriptors such as row or column counts,
4. identifier-like fields.

Let $\mathcal{F}$ denote the set of all initially extracted variables. After anti-leakage filtering, only content-derived numeric summaries were retained:

$$\mathcal{F}_{\text{sf}} \subset \mathcal{F}. \tag{19}$$

Constant features were then removed according to:

$$\mathcal{F}_{final} = \{f_j \in \mathcal{F}_{\text{sf}} \mid \text{Var}(f_j) > 0\}. \tag{20}$$

This filtering step is a methodological design choice of the present study and is not copied from a single source. It was introduced to ensure that the learned mapping reflects GNSS observables rather than acquisition artifacts.

2.6. Feature-Set Design and Ablation Strategy

To quantify the contribution of each observable family, several feature subsets were evaluated. Let:

$$\mathcal{F}_{\text{NAV}}, \mathcal{F}_{\text{MON}}, \mathcal{F}_{\text{RINEX}} \tag{21}$$

denote the feature subsets derived from navigation, monitoring, and RINEX variables, respectively.

The main sets were:

$$\begin{aligned} \mathcal{F}_{all} &= \mathcal{F}_{NAV} \cup \mathcal{F}_{MON} \cup \mathcal{F}_{RINEX}, \\ \mathcal{F}_{nav} &= \mathcal{F}_{NAV} \\ \mathcal{F}_{-nav} &= \mathcal{F}_{MON} \cup \mathcal{F}_{RINEX}, \\ \mathcal{F}_{-mon} &= \mathcal{F}_{NAV} \cup \mathcal{F}_{RINEX}. \end{aligned} \tag{22}$$

This ablation design was created to answer two different questions:

- whether one feature family alone is already sufficient for robust recognition.
- whether a family contributes unique information beyond the remaining families.

In the final analysis, these ablations become scientifically central because they allow attribution of predictive value to physically interpretable observable families.

2.7. Classification Models

Three supervised classifiers were evaluated:

2.7.1. Multinomial logistic regression

A linear baseline was implemented using multinomial logistic regression with median imputation and feature standardization. Its class-posterior form follows the standard multinomial logistic model [21]:

$$P(y = c \mid \boldsymbol{x}) = \frac{\exp(\boldsymbol{w_c^\top x} + b_c)}{|\mathcal{Y}|} \sum_{k=1} \exp\left(\boldsymbol{w_k^\top x} + b_k\right). \tag{23}$$

This model was included as a low-capacity baseline to test whether the multiclass structure is approximately linearly separable in the engineered feature space [21].

2.7.2. Random forest

Random forest was used as the primary nonlinear ensemble baseline. In Breiman's formulation, a forest is an ensemble of randomized trees whose outputs are aggregated by voting [22]:

$$\hat{y} = \text{mode}\{h_1(\boldsymbol{x}), h_2(\boldsymbol{x}), \dots, h_T(\boldsymbol{x})\}. \quad (23)$$

Random forests were selected because they are well suited to nonlinear interactions in structured tabular data and provide a strong baseline without requiring heavy feature scaling assumptions [22].

2.7.3. Histogram-based gradient boosting

A second nonlinear baseline was implemented using histogram-based gradient boosting. This follows Friedman's stage-wise additive boosting formulation [23]:

$$F_M(\boldsymbol{x}) = \sum_{m=1}^{M} \gamma_m h_m(\boldsymbol{x}), \quad (24)$$

where h_m denotes the weak learner fitted at stage m, and γ_m its step size [23]. Gradient boosting was included as a strong competing tree-based method for nonlinear tabular classification [23-24].

2.8. Evaluation Protocols

Two complementary evaluation protocols were used.

2.8.1. Grouped repeated split

In the first protocol, repeated train/test partitions were generated at the scenario level:

$$\mathcal{S}_{train} \cap \mathcal{S}_{\text{test}} = \emptyset. \quad (24)$$

All windows extracted from the same scenario remained on the same side of the partition. This protocol estimates average generalization under repeated scenario-disjoint train/test splits.

2.8.2. Leave-one-scenario-out (LOSO)

The primary robustness benchmark was leave-one-scenario-out evaluation:

$$\mathcal{S}_{\text{test}} = \{S_m\}, \qquad \mathcal{S}_{train} = \mathcal{S} \setminus \{S_m\}. \quad (25)$$

LOSO is the strictest protocol used here because the test subset is always an entirely unseen scenario. This protocol was therefore treated as the primary benchmark for robustness to scenario shift.These split definitions are methodological choices of the present work. Their purpose is to evaluate

generalization under explicit scenario-level non-independence rather than under random window shuffling.

2.9. Performance Metrics

The primary metric was balanced accuracy. Following [25], balanced accuracy is the average recall across classes:

$$\mathrm{BAcc} = \frac{1}{C}\sum_{c=1}^{C}\frac{TP_c}{TP_c + FN_c}. \tag{26}$$

In addition, macro-F1 was computed. Following the standard classification-measure formulation summarized by [26], class-wise precision and recall are:

$$P_c = \frac{TP_c}{TP_c + FP_c}, \qquad R_c = \frac{TP_c}{TP_c + FN_c}, \tag{27}$$

and class-wise F1 is:

$$F1_c = \frac{2P_cR_c}{P_c + R_c}. \tag{28}$$

The macro-average is then:

$$\mathrm{MacroF1} = \frac{1}{C}\sum_{c=1}^{C}F1_c. \tag{29}$$

Overall accuracy was also reported as:

$$\mathrm{Acc} = \frac{1}{N}\sum_{i=1}^{N}\mathrm{b1}(\hat{y_i} = y_i). \tag{30}$$

Balanced accuracy was treated as the primary decision metric because it gives equal weight to all classes regardless of their support, while macro-F1 provides additional sensitivity to class-wise precision–recall balance [26 - 27].

2.10. Scenario-level error analysis

To go beyond aggregate performance, scenario-level diagnostics were explicitly constructed. For LOSO evaluation, a model-by-scenario performance matrix was defined as:

$$\boldsymbol{A} \in R^{M\times K}, \tag{31}$$

where A_{mk} denotes the balanced accuracy of model k on held-out scenario S_m.

The mean scenario difficulty was defined as:

$$d_m = \frac{1}{K}\sum_{k=1}^{K} A_{mk}. \tag{32}$$

which ranks scenarios from hardest to easiest.

The protocol-dependent robustness gap for model k was defined as:

$$\Delta_k = \mathrm{BAcc}_{\mathrm{k}}^{\mathrm{grouped}} - \mathrm{BAcc}_{\mathrm{k}}^{\mathrm{LOSO}}. \tag{31}$$

These quantities are not taken from an external source. They are analytical summaries introduced in this study to localize scenario-dependent error concentration and quantify degradation under stricter evaluation.

2.11. Principal component analysis for visualization

For exploratory visualization of the final feature space, principal component analysis (PCA) was applied after median imputation and standardization. PCA originated with Pearson [26] and was subsequently developed into a general multivariate method by Hotelling. Modern statistical treatment is summarized by [27]. The transformation projects the standardized feature matrix onto orthogonal directions of maximal variance [26 - 27].

If $X \in R^{N\times d}$ denotes the standardized feature matrix, PCA seeks orthogonal directions $v_1, v_2, \ldots$ such that the projected variances are maximized under orthogonality constraints. The first two principal components were used only for visualization and not for model training [26 - 27].

3. Results

The empirical findings are organized from dataset composition and benchmark structure to classification performance, feature-family ablation, class-wise error analysis, and scenario-level robustness. Particular emphasis is placed on the comparison between grouped repeated evaluation and leave-one-scenario-out validation, as well as on identifying which observable family contributes most strongly to robust GNSS attack recognition.

3.1. Overview of the Dataset Composition and Experimental Setup

Analyzed benchmark was constructed from the open JammerTest 2024 scenarios [18]. The final dataset composition after scenario discovery, preprocessing, and fixed-size window construction is summarized in Table 1 and visualized in Fig. 1 and 2. In total, the modeling set contained 24 scenarios and 378 window-level samples, distributed across four attack classes: spoofing (138 windows), jamming (116 windows), meaconing (64 windows), and combined jamming–spoofing (60 windows). The main numerical characteristics of the benchmark are reported in Table 1.

Table 1. Dataset summary after preprocessing and window construction.

Metric	Value
Number of scenarios	24
Number of window samples	378
Number of all-safe features	469
Number of NAV features	186
Number of MON features	202
Number of RINEX features	80
Jamming windows	116
Spoofing windows	138
Meaconing windows	64
Combined jamming–spoofing windows	60

The numerical summary reported in Table 1 shows that the benchmark combines a moderate number of scenarios with a heterogeneous anti-leakage feature space derived from multiple observable families. In addition to the total number of scenarios and windows, the table highlights the dimensionality of the final feature representation and the distribution of samples across attack classes. The scenario-level structure of the benchmark is illustrated in Fig. 1.

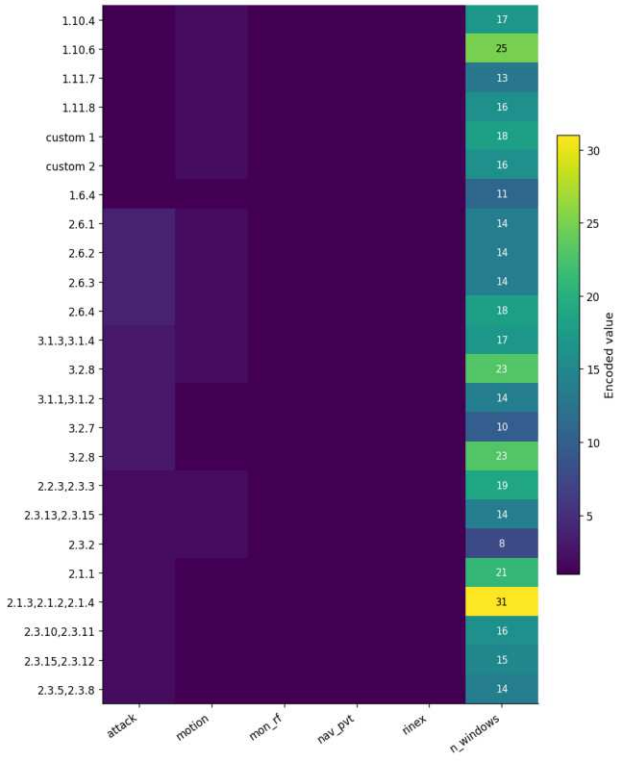


Fig. 1 Scenario-level composition of the JammerTest 2024 benchmark.

The scenario-level composition presented in Fig. 1 shows that the benchmark varies not only in attack class, but also in motion condition and source-level structure. Variation in the number of extracted windows further indicates that the benchmark is heterogeneous both in

content and in scenario-level sample support. This is methodologically important because the dataset cannot be treated as a single homogeneous acquisition. Rather, it should be interpreted as a collection of distinct scenario-level realizations, which justifies the use of grouped and scenario-held-out evaluation protocols. The resulting class distribution after preprocessing and window construction is presented in Fig. 2.

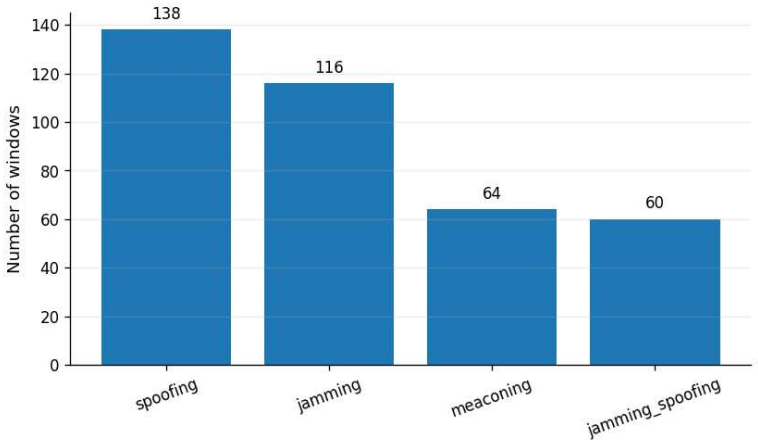


Fig. 2 Distribution of Window-Level Samples Across Attack Classes After Preprocessing and Fixed-Size Window Construction.

The class distribution in Fig. 2 indicates a moderate imbalance across the four attack categories. Spoofing and jamming contribute more window-level samples than meaconing and combined jamming–spoofing, although the imbalance is not severe. Even so, this distribution makes plain accuracy less informative than class-balanced measures. For that reason, balanced accuracy and macro-F1 were treated as the primary evaluation metrics in the subsequent experiments. Taken together, the observed class proportions and scenario variability define a compact but nontrivial multiclass benchmark.

3.2. Main Multiclass Classification Performance

The primary task considered in this study was multiclass recognition of jamming, spoofing, meaconing, and combined jamming–spoofing scenarios. The main comparative results obtained with the anti-leakage all-safe feature set are summarized in Tables 2 and 3 and illustrated in Fig. 3 and 4. In particular, Fig. 3 reports performance under grouped repeated scenario-level evaluation, whereas Fig. 4 presents the stricter leave-one-scenario-out (LOSO) results.

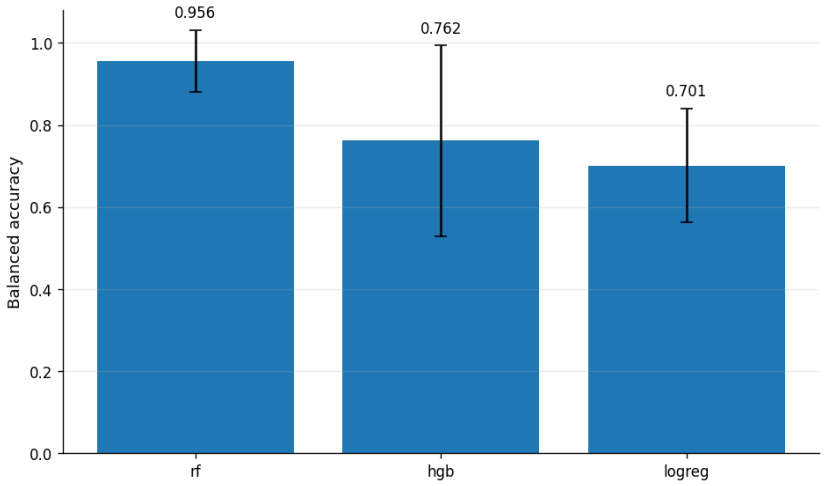


Fig. 3. Multiclass Attack Recognition Under Grouped Repeated Scenario-Level Evaluation Using All-Safe Features.

Table 2. Multiclass attack recognition under grouped repeated scenario-level evaluation using all-safe features.

Model	Accuracy (mean ± std)	Balanced accuracy (mean ± std)	Macro-F1 (mean ± std)
Random Forest	0.960 ± 0.070	0.956 ± 0.076	0.956 ± 0.076
HistGradientBoosting	0.810 ± 0.241	0.762 ± 0.232	0.701 ± 0.308
Logistic Regression	0.672 ± 0.110	0.701 ± 0.139	0.576 ± 0.183

Under grouped repeated evaluation, the random forest classifier clearly outperformed the other models, reaching balanced accuracy = 0.956 and macro-F1 = 0.956 on the all-safe feature set. The next-best model, histogram gradient boosting, achieved a substantially lower balanced accuracy of 0.762, while logistic regression reached 0.701. These results indicate that the final feature space is strongly informative, but also that the discrimination problem is not purely linear. The large gap between logistic regression and the tree-based ensembles suggests that nonlinear interactions among GNSS observables play an important role in multiclass attack recognition. The stricter LOSO benchmark is shown in Fig. 4 and summarized numerically in Table 3.

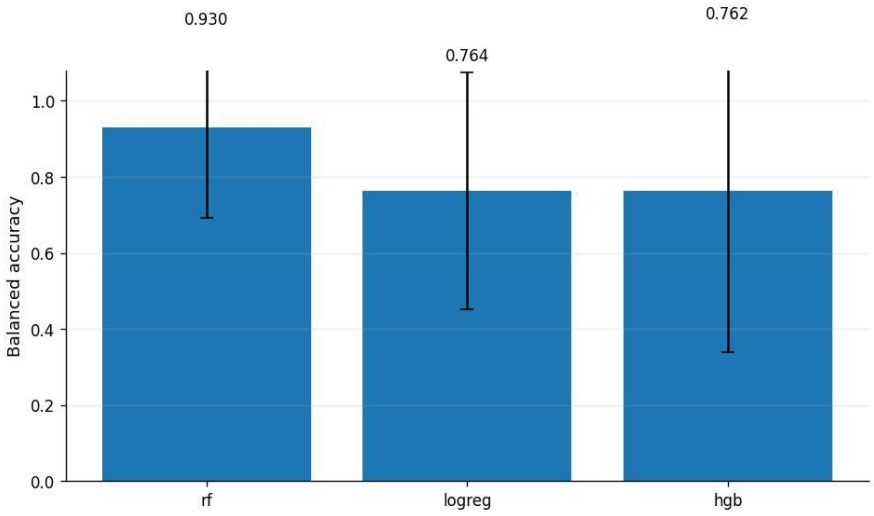


Fig. 4 Multiclass Attack Recognition Under Leave-One-Scenario-Out Evaluation Using All-Safe Features.

Table 3. Multiclass attack recognition under leave-one-scenario-out evaluation using all-safe features.

Model	Accuracy (mean ± std)	Balanced accuracy (mean ± std)	Macro-F1 (mean ± std)
Random Forest	0.930 ± 0.238	0.930 ± 0.238	0.925 ± 0.251
Logistic Regression	0.764 ± 0.311	0.764 ± 0.311	0.557 ± 0.356
HistGradientBoosting	0.762 ± 0.422	0.762 ± 0.422	0.754 ± 0.429

The LOSO protocol confirms the same overall ranking. On unseen scenarios, random forest again remained the strongest model, achieving balanced accuracy = 0.930 and macro-F1 = 0.925. In comparison, logistic regression and histogram gradient boosting were substantially weaker, both remaining close to 0.76 balanced accuracy. This is important because it shows that the best-performing model remains highly competitive even when each test fold is formed by a completely unseen scenario. The performance drop from grouped repeated to LOSO evaluation for random forest is limited, suggesting that the model captures structure that transfers reasonably well across unseen scenarios. Taken together, Fig. 4 and Table 3 provide the strictest estimate of multiclass generalization performance in this study.

3.3. Robustness of the Navigation Feature Family

A central result of this study is that the navigation-derived feature family provides the strongest contribution to robust predictive performance. This can be seen most directly in the source-only ablation results summarized in Table 4 and illustrated in Fig. 5. The best model trained on the nav-only feature set matched the grouped repeated performance of the best model trained on the full all-safe representation, namely balanced accuracy = 0.956 and macro-F1 = 0.956. In practical terms, this indicates that adding monitoring- and RINEX-derived descriptors on top of the navigation observables did not improve the strongest model.

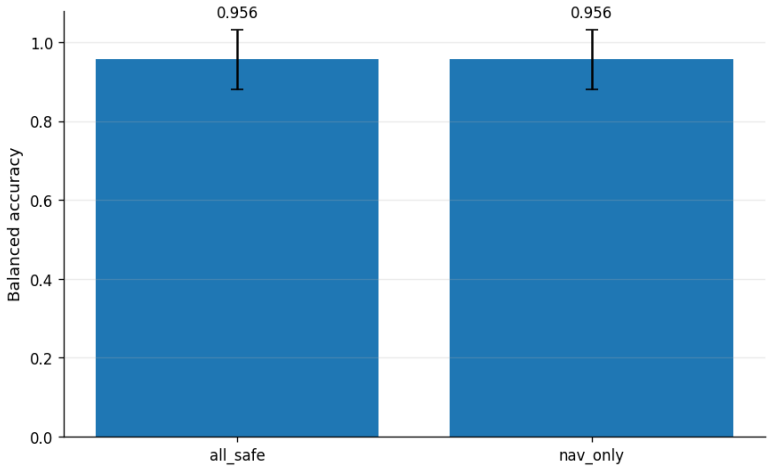


Fig. 5 Source-Only Ablation for Multiclass GNSS Attack Recognition Under Grouped Repeated Evaluation.

Table 4. Source-only ablation for multiclass attack recognition using grouped repeated evaluation.

Feature set	Best model	Accuracy (mean ± std)	Balanced accuracy (mean ± std)	Macro-F1 (mean ± std)
all-safe	Random Forest	0.960 ± 0.070	0.956 ± 0.076	0.956 ± 0.076
nav-only	Random Forest	0.960 ± 0.070	0.956 ± 0.076	0.956 ± 0.076

The agreement between the all-safe and nav-only conditions shown in Fig. 5 and Table 4 shifts the interpretation of the benchmark away from a generic multimodal fusion narrative and toward a more specific conclusion: the dominant discriminative information is concentrated in the navigation-derived observables. In other words, the main class-separating structure appears to be captured by the navigation-derived feature family alone. This conclusion becomes even stronger when combined with the leave-one-family-out analysis presented in Fig. 6 and Table 5.

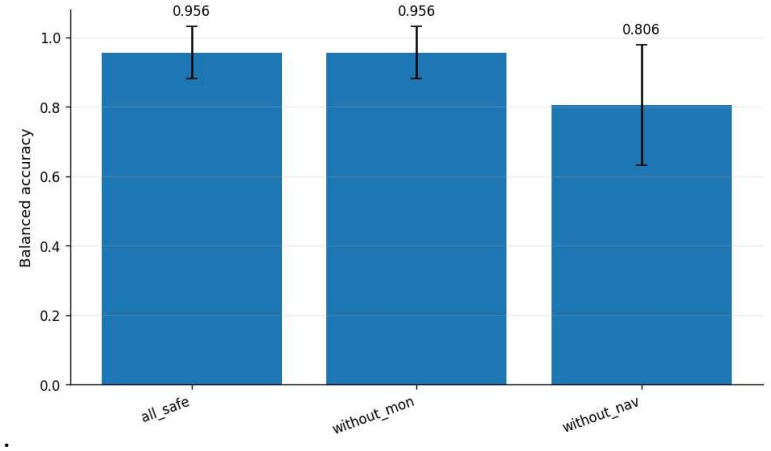


Fig. 6 Leave-One-Family-Out Ablation for Multiclass GNSS Attack Recognition Under Grouped Repeated Evaluation.

Table 5. Leave-one-family-out ablation for multiclass attack recognition using grouped repeated evaluation.

Feature set	Best model	Accuracy (mean ± std)	Balanced accuracy (mean ± std)	Macro-F1 (mean ± std)
all-safe	Random Forest	0.960 ± 0.070	0.956 ± 0.076	0.956 ± 0.076
without-mon	Random Forest	0.960 ± 0.070	0.956 ± 0.076	0.956 ± 0.076
without-nav	HistGradientBoosting	0.851 ± 0.174	0.806 ± 0.173	0.755 ± 0.228

The comparison in Fig. 6 and Table 5 shows that removing MON features did not reduce performance relative to the all-safe condition. The without-mon configuration achieved the same grouped repeated balanced accuracy (0.956) and macro-F1 (0.956) as the full representation. In contrast, removing NAV features led to a substantial degradation. The best without-nav configuration reached only balanced accuracy = 0.806 and macro-F1 = 0.755, despite being evaluated with a nonlinear tree-based model. This reduction is therefore not a minor fluctuation, but a marked loss of multiclass recognition capability. Overall, the ablation results indicate that robust multiclass discrimination in this benchmark is driven primarily by navigation-derived features, whereas monitoring-derived features add little additional predictive value in the tested setting.

3.4. Error Structure Under Strict LOSO Evaluation

The class-wise error structure of the best LOSO model is visualized in Fig. 7, which shows the confusion matrix for the top-performing all-safe LOSO configuration. This analysis is important because it complements the aggregate performance metrics by identifying where the remaining classification errors are concentrated.

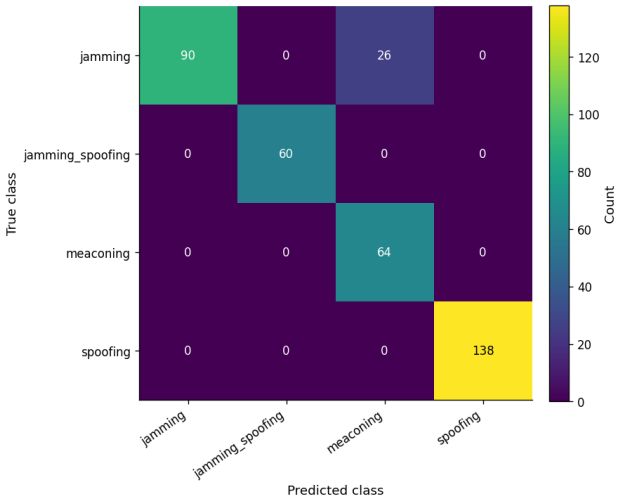


Fig.7 LOSO Confusion Matrix for the Best Multiclass Model.

The confusion matrix in Fig. 7 indicates that the remaining errors are concentrated mainly in the jamming class, whereas spoofing, meaconing, and especially combined jamming–spoofing remain more cleanly separated. The most visible error pattern corresponds to jamming samples being assigned to the meaconing class, while the other categories retain high separability. This suggests that the residual

weakness of the best model is not distributed uniformly across the label space, but is instead localized at a more specific class boundary. Under strict scenario-held-out evaluation, the main remaining challenge is therefore not general multiclass instability, but a more localized ambiguity affecting selected attack categories.

3.5. Scenario-Wise Robustness and Difficulty Variation

Aggregate performance measures do not reveal whether a model behaves uniformly across all held-out scenarios. This issue is addressed in Fig. 8, which reports LOSO balanced accuracy for each held-out scenario and each model, and in Fig. 11, which ranks scenarios according to mean LOSO balanced accuracy across models. The scenario-wise LOSO performance map is shown first in Fig. 8.

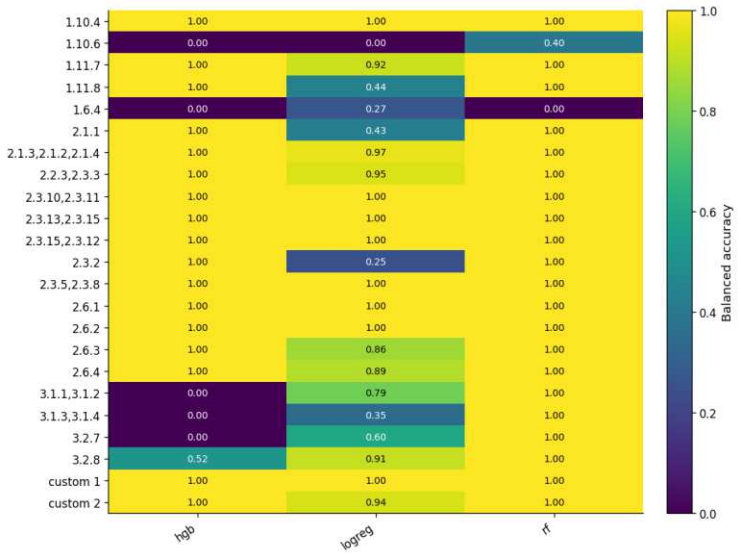


Fig. 8 LOSO Balanced Accuracy Across Held-Out Scenarios and Models.

As shown in Fig. 8, random forest is not only the strongest model on average, but also the most stable model across individual held-out scenarios. Its scenario-wise performance is dominated by values close to perfect accuracy, whereas the alternative models exhibit more frequent low-performing cases. This is an important finding because a model can achieve a good average score while still failing severely on a subset of scenarios. The heatmap therefore shows that the advantage of random forest is not limited to one or two favorable folds, but is observed across most of the benchmark. A complementary view is provided by the scenario-difficulty ranking in Fig. 9.

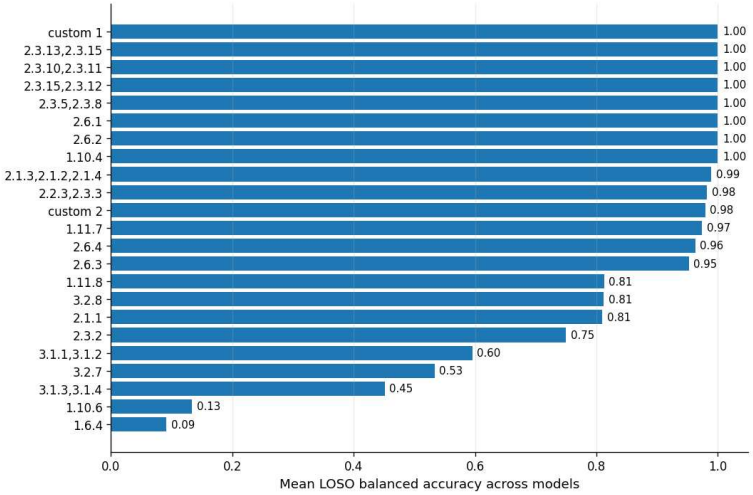


Fig. 9. Scenario Difficulty Ranking Based on Mean LOSO Balanced Accuracy Across Models.

The ranking reported in Fig. 9 shows that the benchmark is not uniformly easy. Some held-out scenarios are solved almost perfectly by all models, whereas others remain substantially harder. This is methodologically useful because it shows that the benchmark is not saturated in a trivial sense. Instead, it retains scenario-dependent variation that is informative for robustness analysis, model comparison, and future error-focused work. Together, Fig. 8 and 9 show that strong average LOSO performance does not imply uniformly easy scenario-level behavior, and that a subset of held-out scenarios remains distinctly more difficult than the rest.

3.6. Stability Across Evaluation Protocols

The relationship between grouped repeated evaluation and the stricter LOSO protocol is summarized in Fig. 10 and 11. These figures provide a direct view of how strongly each model degrades when the evaluation is shifted from repeated scenario-level splits to a fully unseen-scenario benchmark. The paired comparison across protocols is shown in Fig.10.

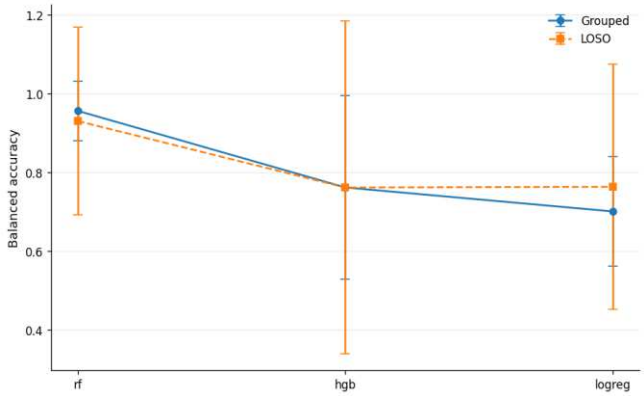


Fig. 10 Stability of Multiclass Performance Across Grouped Repeated and LOSO Evaluation.

As shown in Fig. 10, random forest remains the strongest model under both protocols, while histogram gradient boosting and logistic regression remain clearly weaker. More importantly, the relative ordering of the models is broadly preserved when moving from grouped repeated splitting to LOSO

evaluation. This suggests that the main ranking observed in the benchmark is not an artifact of one specific validation protocol. The corresponding protocol-dependent performance reduction is visualized in Fig. 11.

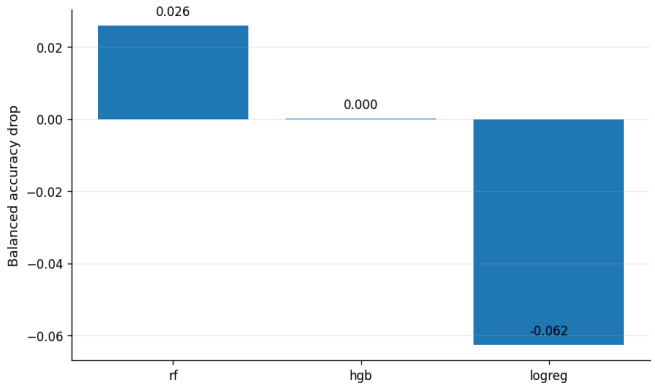


Fig. 11. Generalization Drop From Grouped Repeated to LOSO Evaluation.

The comparison in Fig. 10 shows that random forest remains the strongest model under both evaluation protocols, whereas histogram gradient boosting and logistic regression remain clearly weaker. More importantly, the relative ordering of the models is broadly preserved when moving from grouped repeated splitting to LOSO evaluation. This indicates that the ranking observed in the benchmark is not merely an artifact of a single validation protocol. The corresponding protocol-dependent reduction in performance is shown in Fig. 11.

3.7. Summary of the Main Empirical Findings

Across the full experimental analysis, four empirical conclusions emerged consistently. First, random forest was the strongest multiclass model under both grouped repeated and leave-one-scenario-out evaluation, achieving the best performance in the full all-safe configuration. Second, the ablation study showed that the dominant predictive structure is concentrated in the navigation-derived feature family. The nav-only and all-safe conditions were identical, and the without-mon configuration also matched the full representation. Third, removing navigation-derived features caused a substantial drop in multiclass performance, confirming that NAV observables are not merely correlated with success but are the primary driver of discrimination in this benchmark. Fourth, scenario-wise analyses showed that strong average LOSO performance was not achieved by masking widespread catastrophic failure, although a subset of held-out scenarios remained clearly harder than others. Taken together, the results support a consistent interpretation - under strict anti-leakage filtering and scenario-held-out evaluation, robust GNSS attack recognition in this benchmark is driven primarily by navigation-derived observables, with random forest providing the strongest and most stable overall performance.

4. Discussion

The results support a coherent and practically relevant interpretation of the benchmark. The most important finding is not merely that one classifier achieved the highest score, but that the ablation results in Tables 4 and 5 consistently show that the dominant predictive information is concentrated in the navigation-derived feature family. In particular, the best nav-only configuration matched the best all-safe configuration, while the without-mon setting also matched the full feature set. By contrast, removing navigation features caused a substantial degradation of multiclass performance. This pattern strongly suggests that, in the present benchmark, the navigation-solution observables carry most of the attack-

discriminative information, whereas the monitoring-derived summaries do not add measurable gains under the tested protocols.

This conclusion is particularly important because it refines the narrative that often appears in GNSS interference literature. A common expectation is that combining more observable families should automatically improve detection or classification performance. The present results do not support such a generalized claim. Instead, they indicate that additional families can be redundant, or at least not beneficial, once a sufficiently informative navigation-level representation is available. In this study, the relevant contrast is especially clear in Figs. 5 and 6, where the source-only and leave-one-family-out ablations produce nearly identical performance for all-safe, nav-only, and without-mon, whereas the without-nav configuration exhibits a marked drop. Therefore, the contribution of the paper is better framed as an observable-family analysis rather than as a generic multimodal fusion study.

The second major result is the stability of the random forest model across evaluation protocols. As shown in Tables 2 and 3, random forest was the strongest model under both grouped repeated evaluation and leave-one-scenario-out evaluation. More importantly, Figs. 10 and 11 show that its generalization drop from grouped repeated to LOSO was limited relative to its absolute performance level. This matters because a model can appear strong under repeated grouped splits but degrade substantially when tested on a completely unseen scenario. That pattern was not observed here for the best random forest configuration. The model remained highly competitive even when the test fold consisted of a previously unseen scenario, which gives credibility to the claim that the model learned scenario-transferable structure rather than split-specific artifacts.

The scenario-level analysis strengthens this interpretation. The LOSO heatmap in Fig. 8 demonstrates that the best-performing model did not win only by averaging over a few favorable folds. Rather, it remained strong across most held-out scenarios. At the same time, the scenario-difficulty ranking in Fig.9 shows that the benchmark is not saturated in a trivial sense. Some scenarios are solved nearly perfectly by all models, but others remain considerably harder. This is methodologically valuable. If the benchmark had consisted only of easy scenarios, the strong aggregate scores reported in Table 3 would have had limited analytical meaning. Instead, the variation across scenarios confirms that the benchmark still contains informative structure for robustness analysis, model comparison, and future error-focused work.

The confusion matrix in Fig. 7 adds a more localized view of the residual error structure. The remaining errors are not evenly distributed across classes but instead appear to be concentrated mainly around the jamming class. In contrast, spoofing, meaconing, and especially the combined jamming–spoofing class appear much more cleanly separated. This asymmetry is important. It indicates that the remaining weakness of the model is not general instability, but rather a more specific boundary problem involving one part of the label space. In practical terms, this suggests that the benchmark is already highly separable in some attack categories but still leaves room for improvement near the jamming-related boundary.

From the perspective of prior work, the present findings fit within a broader trend in GNSS interference research. A recent review [2] noted that machine-learning-based methods have become increasingly prominent in jamming and spoofing detection, and that recent work has moved toward more specialized classification and identification frameworks rather than only binary detection . The present study is aligned with that direction, but it differs from much of the existing literature by focusing on strict scenario-level generalization, anti-leakage filtering, and feature-family attribution, rather than on raw peak accuracy alone.

The comparison with prior machine-learning spoofing studies is particularly instructive. In [9 - 10] validated supervised learning for spoofing and meaconing detection on real-world data in a two-part study, emphasizing the practical potential of supervised models at receiver level and showing that real-world validation is essential for moving beyond purely synthetic experiments. The present work shares that emphasis on real data, but extends the problem formulation in two important ways. First, it treats the task as a four-class recognition problem rather than a narrower manipulated-versus-authentic setting.

Second, it uses scenario-level grouping and LOSO evaluation to make the estimate of generalization more conservative. Because the datasets and task formulations are different, the numerical results are not directly comparable. However, the conceptual similarity is important: both studies indicate that receiver-level observables can support machine-learning-based interference recognition, while also showing that real-world validation is indispensable [9-10].

The findings also resonate with earlier work on interference monitoring from engineered tabular features [15] proposed a twin support vector machine method for real-time GNSS interference monitoring, showing that compact receiver-level descriptors can support machine-learning-based interference detection . Likewise, [12] demonstrated that a multi-layer neural network can be used for spoofing detection in single-frequency GPS receivers . These earlier studies support the general premise that GNSS interference leaves structured signatures in receiver-side observables. What the present study adds is a more explicit demonstration that, within a heterogeneous open benchmark, not all observable families contribute equally, the navigation-level descriptors appear to dominate robust multiclass recognition.

Recent work on image-based and deep-learning approaches provides another useful point of comparison. Authors of [17] classified seven GNSS jamming signal categories from spectral images and reported precision above 90% for the best-performing ResNet configuration. Composite investigated jamming detection and recognition in [15] using deep learning on preprocessed time-frequency representations and reported that time-frequency diagrams gave the best performance, with 90% detection at −20 dB JNR and 90% identification at −13 dB JNR. These studies show that deep architectures can be highly effective when working directly with signal-domain or image-domain representations. However, they address a somewhat different problem from the one solved here. The present work does not rely on spectrograms, raw IQ tensors, or heavy image preprocessing. Instead, it shows that compact tabular summaries extracted from open scenario files can already achieve strong unseen-scenario multiclass performance. In that sense, the present study occupies a different and complementary point in the design space, it trades raw signal richness for interpretability, ease of reproducibility, and lightweight deployment.

A particularly relevant comparison is the recent papers, which used JammerTest 2024 data for machine-learning-based detection and mitigation of jamming, meaconing, and spoofing with a CNN-LSTM architecture and multi-sensor inputs. That work is close in spirit because it also leverages JammerTest 2024 and treats multiple interference types jointly [29]. The present study complements that direction by showing that even without a deep sequential architecture, and even after explicit anti-leakage filtering, strong scenario-level multiclass recognition can be obtained from compact tabular summaries. Therefore, the current paper is not best positioned as a competitor to deep sequence models, but rather as a transparent and reproducible baseline-plus-analysis study that clarifies which observable families matter most in this open benchmark.

These comparisons also help clarify what should and should not be claimed. The current results do not prove that navigation-derived features are universally sufficient for all GNSS interference settings. They only show that, in the tested benchmark and under the present feature construction strategy, navigation-derived observables dominate generalization. It is entirely possible that richer RF-level descriptors, alternative monitoring statistics, or signal-domain representations would contribute more strongly in other settings. Likewise, the current paper should not claim direct state-of-the-art superiority over prior works, because many of those studies solve different tasks, use different datasets, or evaluate under less comparable protocols [9,10][26-29]. The defensible claim is narrower and stronger - within this open benchmark and under strict scenario-held-out evaluation, the most robust predictive structure is concentrated in navigation-solution observables.

The study also has several limitations that should be stated explicitly. First, the benchmark remains relatively small in terms of the number of scenarios. Although LOSO is methodologically stronger than random splitting, performance estimates on 24 scenarios still carry variance, especially for scenario-wise analyses. Second, the features used here are intentionally compact statistical summaries. This improves interpretability and reproducibility, but it may discard temporal and spectral structure that

more advanced sequential or signal-domain models could exploit. Third, the current analysis is focused on the released scenario files and therefore inherits the acquisition design of the JammerTest 2024 campaign [17-18]. This is appropriate for reproducibility, but it also means that the conclusions should not be generalized too aggressively beyond this benchmark without additional external validation. Fourth, the present work is centered on multiclass attack recognition within attacked scenarios and does not yet include a full authentic-versus-attacked benchmark merged with independent clean-field datasets. Such a merged benchmark would be highly valuable for operational deployment studies. Finally, although the anti-leakage filtering is strong, future work should still consider broader cross-dataset transfer tests to evaluate how well the learned structure survives changes in receiver hardware, environment, and collection protocol.

Despite these limitations, the overall picture is encouraging. The benchmark results are strong under strict LOSO evaluation, the best model remains stable across protocols, and the ablation study yields a clear and interpretable scientific conclusion. From a practical perspective, this means that robust attack recognition may be feasible without heavy raw-signal processing, provided that sufficiently informative navigation-level observables are available. From a scientific perspective, the main contribution is that the paper does not stop at reporting a high score. It identifies which feature family matters, where the residual errors are concentrated, how robust the best model is to stricter scenario holdout, and which parts of the benchmark remain challenging. That combination of benchmark performance, anti-leakage design, and observable-family analysis is what makes the study more than a routine classification exercise.

In summary, the discussion leads to four final conclusions. First, random forest is the strongest and most stable model in the present benchmark, especially under strict leave-one-scenario-out evaluation. Second, navigation-derived observables dominate multiclass recognition performance, while monitoring-derived summaries do not provide measurable gains in the tested setup. Third, the scenario-level analyses show that high average performance is not achieved by hiding widespread catastrophic failures, although a subset of scenarios remains clearly harder than others. Fourth, the main value of the study lies not only in the achieved accuracy, but in the fact that it provides a reproducible, anti-leakage, scenario-held-out analysis of which observable families drive robust GNSS attack recognition.

5. Limitations and Threats to Validity

Several limitations and threats to validity should be acknowledged when interpreting the results of this study.

First, the number of scenarios remains limited. Although the benchmark includes multiple attack classes under both stationary and dynamic conditions, the final evaluation still relies on 24 scenarios and 378 window-level samples. This is sufficient for a meaningful benchmark study, but it is still a relatively small scenario set for making strong claims about universal GNSS generalization. For that reason, the results should be interpreted as scenario-held-out evidence within this benchmark rather than as proof of broad deployment robustness. The use of grouped repeated splitting and leave-one-scenario-out evaluation was intended to make the analysis more conservative under these conditions, but it does not remove the fundamental limitation imposed by the finite number of scenarios.

Second, the conclusions are tied to a single open benchmark campaign, namely JammerTest 2024 [18]. This is a strength from the perspective of reproducibility, but it is also a limitation from the perspective of external validity. The acquisition design, receiver configuration, environmental conditions, and attack realizations are all specific to this campaign. As a result, the present study cannot claim that the same feature-family ranking or the same model ordering will necessarily hold across different hardware platforms, datasets, or operational environments. In particular, the claim that navigation-derived observables dominate robust recognition should be understood as a benchmark-grounded conclusion rather than as a universal statement about all GNSS interference datasets or receiver architectures.

Third, the feature representation is intentionally compact and statistical. This improves interpretability, computational efficiency, and reproducibility, but it also imposes a representational limit.

The current pipeline summarizes each window using descriptive statistics, quantiles, slope estimates, and missing-value ratios. Such features are well suited to transparent tabular modeling, but they may not capture finer temporal, spectral, or signal-domain structure that could be exploited by sequential or deep-learning approaches [26-29]. Therefore, the present work should not be interpreted as evidence that compact statistical summaries are always sufficient, but rather that they are sufficient to achieve strong multiclass performance in this benchmark under the tested protocol.

Fourth, the study does not propose a new classifier architecture, a new signal-processing front-end, or a new deep feature extractor. The novelty of the paper is therefore primarily methodological and analytical rather than architectural. More specifically, the contribution lies in combining anti-leakage filtering, strict scenario-level evaluation, feature-family attribution, and scenario-wise robustness analysis within one reproducible benchmark framework. In review settings that prioritize strong algorithmic novelty, this may be perceived as a weaker form of contribution. The intended positioning of the paper is therefore not as a new model paper, but as a rigorous benchmark-and-analysis study that clarifies which observable families support robust multiclass GNSS attack recognition.

Fifth, the current evaluation remains centered on attacked scenarios and multiclass attack recognition. While this is an important and nontrivial task, the present work does not yet provide a full operational benchmark that includes large-scale authentic clean data from independent campaigns. In real-world deployment, an attack-recognition module would usually be preceded by a clean-versus-attacked decision stage, and performance would depend not only on multiclass separation among attack types but also on robustness against benign environmental variation. Consequently, the present benchmark should be viewed as an important intermediate step rather than as a complete end-to-end operational validation.

Sixth, although the anti-leakage protocol removes metadata-derived variables, source-presence flags, and file-structure descriptors, no leakage analysis can be considered absolutely final in a single benchmark study. It remains possible that some benchmark-specific regularities persist in the remaining observables. The fact that LOSO performance remains strong after anti-leakage filtering is encouraging, but validation on independent datasets would still be necessary to exclude benchmark-specific shortcuts more confidently.

Finally, the comparative analysis is limited to established tabular baselines, namely multinomial logistic regression, random forest, and histogram-based gradient boosting [21-23]. These models were deliberately chosen because they are transparent, robust on structured data, and easy to reproduce. However, the resulting conclusions should not be interpreted as implying that more advanced sequence models, signal-domain networks, or hybrid architectures would not perform better under richer input representations [17-18][29]. Rather, the current study shows that even within a lightweight and interpretable modeling framework, strong scenario-held-out multiclass performance is achievable, and that the navigation-derived feature family is the dominant source of predictive information in the tested setting.

The study supports a strong claim about robust multiclass recognition within an open, scenario-held-out benchmark, and it provides evidence that navigation-derived observables dominate performance in that benchmark. At the same time, it does not claim universal external validity, receiver independence, or methodological superiority over all possible signal-domain or deep-learning approaches. Future work should therefore focus on cross-dataset transfer, richer external validation, alternative receiver platforms, and hybrid feature representations that preserve interpretability while incorporating more detailed temporal and spectral structure.

6. Conclusion

This paper presented a reproducible and anti-leakage benchmark for multiclass GNSS attack recognition using the open JammerTest 2024 dataset. The study focused on four interference classes—jamming, spoofing, meaconing, and combined jamming–spoofing—and evaluated compact statistical

representations derived from navigation, monitoring, and RINEX-related observables under grouped repeated and leave-one-scenario-out evaluation protocols.

The results support three main conclusions. First, random forest was the strongest and most stable classifier across the evaluated settings. It achieved the best multiclass performance under both grouped repeated evaluation and the stricter LOSO protocol, indicating that the model captures structure that generalizes beyond individual scenario partitions. Second, the ablation analysis showed that the dominant discriminative information is concentrated in navigation-derived observables. In particular, the navigation-only representation matched the best full-feature performance, while removing monitoring-derived features did not degrade the strongest model. Third, removing navigation-derived features caused a clear and substantial drop in performance, showing that navigation-level observables are the primary driver of robust multiclass recognition in the present benchmark.

The scenario-level analyses further reinforced this interpretation. The LOSO heatmap and scenario-difficulty ranking showed that the benchmark is not trivially solved, even though the best model performs strongly on most unseen scenarios. At the same time, the confusion analysis indicated that the remaining errors are concentrated mainly around the jamming class boundary rather than being uniformly distributed across all classes. This suggests that future improvements should focus less on further inflating average scores and more on resolving specific hard scenarios and ambiguous class boundaries.

The study also has important limitations. The number of scenarios remains limited, and broader claims about universal GNSS generalization should therefore be made with caution. The current feature representation is intentionally compact and interpretable, but it does not exploit richer temporal, spectral, or raw-signal structure that could be beneficial under more difficult conditions. In addition, the conclusions are tied to the acquisition design of JammerTest 2024. and should ideally be validated on independent datasets, receiver platforms, and environments.

Despite these limitations, the study makes a useful contribution in two respects. Methodologically, it provides a transparent scenario-level evaluation framework with explicit anti-leakage filtering and family-wise ablation. Empirically, it shows that strong multiclass GNSS attack recognition can be achieved from lightweight tabular summaries and, more importantly, that the most robust predictive structure in this benchmark resides in navigation-derived observables. For that reason, the value of the paper lies not only in the achieved classification accuracy, but also in the clearer identification of which GNSS observable family supports robust unseen-scenario attack recognition.

Acknowledgments

The authors would like to thank the organizers and contributors of the JammerTest 2024 campaign for making the dataset publicly available. The authors also acknowledge the open dissemination of the benchmark materials through Zenodo, which made the present reproducible analysis possible.

Funding

Not applicable

Competing Interests

The authors have no competing interests to declare.

Author Contributions

Conceptualization: S.Č., P.Š.; Methodology: S.Č., P.Š.; Formal analysis: S.Č., P.Dz.; Investigation: S.Č., P.Dz., P.H.; Visualization: S.Č.; Writing – original draft: S.Č.; Writing – review & editing: S.Č., P.Š., P.Dz., P.H.; Supervision: P.Š.

Data Availability

The dataset analyzed in this study is publicly available on Zenodo under the title *GNSS Dataset Under Jamming, Spoofing, and Meaconing Conditions (JammerTest 2024)* at: https://doi.org/10.5281/zenodo.15911589. The generated tables, figures, and processed outputs supporting the findings of this study are included in the manuscript and its associated submission files.

Code Availability

The code used in this study is available from the corresponding author upon reasonable request.

Ethics Approval

Not applicable.

Consent to Participate

Not applicable.

Consent for Publication

Not applicable.

References

[1] Psiaki, M. L., & Humphreys, T. E. (2016). GNSS Spoofing and Detection. Proceedings of the IEEE, 104(6), 1258–1270. https://doi.org/10.1109/jproc.2016.2526658

[2] Radoš, K., Brkić, M., & Begušić, D. (2024). Recent advances on jamming and spoofing detection in GNSS. Sensors, 24(13), 4210. https://doi.org/10.3390/s24134210

[3] Wesson, K. D., Gross, J. N., Humphreys, T. E., & Evans, B. L. (2017). GNSS signal authentication via power and distortion monitoring. IEEE Transactions on Aerospace and Electronic Systems, 54(2), 739–754. https://doi.org/10.1109/taes.2017.2765258

[4] Borhani-Darian, P., Li, H., Wu, P., & Closas, P. (2024). Detecting GNSS spoofing using deep learning. EURASIP Journal on Advances in Signal Processing, 2024(1). https://doi.org/10.1186/s13634-023-01103-1

[5] Schmidt, D., Radke, K., Camtepe, S., Foo, E., & Ren, M. (2016). A survey and analysis of the GNSS spoofing threat and countermeasures. ACM Computing Surveys, 48(4), 1–31. https://doi.org/10.1145/2897166

[6] Zhu, X., Hua, T., Yang, F., Tu, G., & Chen, X. (2021). Global positioning system spoofing detection based on Support Vector Machines. IET Radar Sonar & Navigation, 16(2), 224–237. https://doi.org/10.1049/rsn2.12178

[7] Xu, R., Ding, M., Qi, Y., Yue, S., & Liu, J. (2018). Performance Analysis of GNSS/INS Loosely Coupled Integration Systems under Spoofing Attacks. Sensors, 18(12), 4108. https://doi.org/10.3390/s18124108

[8] Sun, C., Cheong, J. W., Dempster, A. G., Demicheli, L., Cetin, E., Zhao, H., & Feng, W. (2018). Moving variance-based signal quality monitoring method for spoofing detection. GPS Solutions, 22(3). https://doi.org/10.1007/s10291-018-0745-7

[9] Semanjski, S., Semanjski, I., De Wilde, W., & Muls, A. (2020). Use of Supervised Machine Learning for GNSS Signal Spoofing Detection with Validation on Real-World Meaconing and Spoofing Data—Part I. Sensors, 20(4), 1171. https://doi.org/10.3390/s20041171

[10] Semanjski, S., Semanjski, I., De Wilde, W., & Gautama, S. (2020). Use of Supervised Machine Learning for GNSS Signal Spoofing Detection with Validation on Real-World Meaconing and Spoofing Data—Part II. Sensors, 20(7), 1806. https://doi.org/10.3390/s20071806

[11] Ferre, R. M., De La Fuente, A., & Lohan, E. S. (2019). Jammer classification in GNSS bands via machine learning algorithms. Sensors, 19(22), 4841. https://doi.org/10.3390/s19224841

[12] Shafiee, E., Mosavi, M. R., & Moazedi, M. (2017). Detection of Spoofing Attack using Machine Learning based on Multi-Layer Neural Network in Single-Frequency GPS Receivers. Journal of Navigation, 71(1), 169–188. https://doi.org/10.1017/s0373463317000558

[13] Ivanov, V., Scaramuzza, M., & Wilson, R. C. (2024). Deep temporal semi-supervised one-class classification for GNSS radio frequency interference detection. Journal of Navigation, 77(1), 59–81. https://doi.org/10.1017/s0373463324000134

[14] Mehr, I. E., & Dovis, F. (2024). A deep neural network approach for classification of GNSS interference and jamming. IEEE Transactions on Aerospace and Electronic Systems, 61(2), 1660–1676. https://doi.org/10.1109/taes.2024.3462662

[15] Liu, C., Ren, B., Xie, Y., & Chen, F. (2025). Deep learning-based GNSS composite jamming detection and recognition technology. Frontiers in Signal Processing, 5. https://doi.org/10.3389/frsip.2025.1567926

[16] Sung, Y., Park, S., Kim, D., & Kim, S. (2022). GPS spoofing detection method for small UAVs using 1D convolution Neural Network. Sensors, 22(23), 9412. https://doi.org/10.3390/s22239412

[17] Steiner, J., & Pešík, J. (2024). Machine Learning image recognition for GNSS jamming signals categorization. Neural Network World, 34(6), 341–360. https://doi.org/10.14311/nnw.2024.34.019

[18] Sayyaf, M. I., Ortiz, M., & Renaudin, V. (2025). GNSS Dataset under jamming, spoofing, and meaconing conditions (JammerTest 2024) [Dataset]. In Zenodo (CERN European Organization for Nuclear Research). https://doi.org/10.5281/zenodo.15911589

[19] IGS/RTCM RINEX Working Group, RINEX: The Receiver Independent Exchange Format Version 3.05, 2020. DOI: unavailable. Official format specification.

[20] u-blox AG, ZED-F9P Interface Description, document UBX-18010854, revision R08, 28-May-2020. DOI: unavailable. Official receiver interface documentation for UBX message families including navigation and monitoring messages.

[21] Böhning, D. (1992). Multinomial logistic regression algorithm. Annals of the Institute of Statistical Mathematics, 44(1), 197–200. https://doi.org/10.1007/bf00048682

[22] Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/a:1010933404324

[23] Friedman, J. H. (2001). Greedy function approximation: A gradient boosting machine. The Annals of Statistics, 29(5). https://doi.org/10.1214/aos/1013203451

[24] Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. Proceedings of the 20th International Conference on Pattern Recognition (ICPR), 3121–3124. https://doi.org/10.1109/icpr.2010.764

[25] Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. Information Processing & Management, 45(4), 427–437. https://doi.org/10.1016/j.ipm.2009.03.002

[26] Pearson, K. (1901). LIII. On lines and planes of closest fit to systems of points in space. The London Edinburgh and Dublin Philosophical Magazine and Journal of Science, 2(11), 559–572. https://doi.org/10.1080/14786440109462720

[27] Jolliffe, I. T. (1986). Principal component analysis. In Springer series in statistics. https://doi.org/10.1007/978-1-4757-1904-8

[28] Li, W., Huang, Z., Lang, R., Qin, H., Zhou, K., & Cao, Y. (2016). A Real-Time interference monitoring technique for GNSS based on a twin support vector machine method. Sensors, 16(3), 329. https://doi.org/10.3390/s16030329

[29] Bohlig, P., García, J. M., Lalgudi, R., & Fischer, J. (2025). Detection and Mitigation of Jamming, Meaconing, and Spoofing based on Machine Learning and Multi-Sensor Data. Proceedings of the Satellite Division's International Technical Meeting (Online)/Proceedings of the Satellite Division's International Technical Meeting (CD-ROM), 765–777. https://doi.org/10.33012/2025.20368