Data Quality Over Algorithms: A Multi-Factor Analysis of Automatic Identification System (AIS) Vessel Classification

Jákup Svøðstein
University of the Faroe Islands
Tórshavn, Faroe Islands
ORCID: 0000-0002-4374-1752

Jóhannus Kristmundsson
University of the Faroe Islands
Tórshavn, Faroe Islands
ORCID: 0000-0003-4757-2965

***Abstract*—Automatic Identification System (AIS) vessel classification using machine learning has gained traction for maritime domain awareness, with recent work emphasizing algorithm selection and hyperparameter optimization. However, the relative importance of data preprocessing versus model choice remains unexplored. We conduct a systematic three-way ANOVA study evaluating model selection (CatBoost, LightGBM, XGBoost, Random Forest), trajectory window size (1h, 2h, 3h, 6h, 12h), and moving fraction threshold (25%, 50%, 75% moving positions) across 174 experimental configurations on native Faroese AIS data. Our analysis reveals a key finding: moving fraction threshold explains 58.1% of performance variance ($\eta^2 = 0.581$, $p < 0.001$), while model choice is negligible ($\eta^2 = 0.004$, $p = 0.877$). The optimal configuration (6h window + 75% moving threshold) achieves validation macro-F1 of 0.626 (XGBoost), with model-agnostic performance across tree ensembles (F1 range: 0.566–0.626). These findings challenge the conventional ML focus on algorithm optimization and suggest that preprocessing rigor should precede model selection in maritime trajectory classification tasks.**

***Index Terms*—AIS vessel classification, machine learning, data quality, trajectory filtering, gradient boosting, maritime informatics.**

I. Introduction

Despite a decade of model-centric optimization in maritime machine learning, preprocessing and data-quality factors remain unquantified [1]–[5]. This paper isolates those factors experimentally, revealing that data quality substantially outweighs algorithm selection in AIS vessel classification.

Automatic Identification System (AIS) vessel classification supports maritime safety, fisheries enforcement, and traffic management [6]. However, vessel type labels in AIS messages are self-reported and frequently erroneous [4], necessitating automated classification from behavioral patterns. Recent machine learning approaches emphasize algorithm selection: Meyer and Kleynhans [7] achieve F1=0.88–0.90 with LightGBM on 12–24h trajectory windows, while Rong et al. [3] compare Random Forest and XGBoost on Chinese coastal AIS. These studies treat preprocessing choices (window size, moving fraction filtering) as fixed implementation details rather than experimental variables.

This focus may be misplaced. The data-centric AI paradigm [8] advocates prioritizing data quality over algorithmic sophistication, yet quantitative evidence remains sparse in maritime domains. Sambasivan et al. [5] highlight persistent undervaluation of data preparation, while Domingos [9] argues feature quality often dominates model choice. No prior AIS study systematically quantifies the relative importance of preprocessing versus model selection, leaving practitioners without evidence-based guidance on where to invest engineering effort.

We address this gap through a three-way ANOVA evaluating model selection (XGBoost, LightGBM, CatBoost, Random Forest), window size (1h, 2h, 3h, 6h, 12h), and moving fraction (25%, 50%, 75% moving) across 174 configurations on Faroese AIS data. Our findings: moving fraction explains 58.1% of performance variance ($\eta^2 = 0.581$, $p < 0.001$) while model choice is negligible ($\eta^2 = 0.004$, $p = 0.877$), demonstrating that data preprocessing substantially outweighs algorithm selection. The optimal configuration (6h + 75% moving) achieves validation macro-F1 of 0.626 (XGBoost best; all tree ensembles within 0.566–0.626 range), with model-agnostic performance. The implication is methodological: performance audits in maritime ML should prioritize data-quality sensitivity over model tuning. Beyond maritime applications, our multi-factor experimental design with effect size quantification provides a methodological template for prioritizing ML engineering efforts.

II. Related Work

A. AIS Vessel Classification

Meyer and Kleynhans [7] apply LightGBM to large-scale satellite AIS, achieving F1=0.88–0.90 using 12–24h windows. While their absolute performance exceeds ours (F1=0.626), direct comparison is complicated by different label sets (their aggregated 6 classes vs. our comparative 4 classes) and data sources (global satellite vs. local terrestrial). Crucially, they do not systematically vary preprocessing factors—our contribution quantifies that moving fraction explains 58.1%

AIS data collected under IMO Resolution A.1106(29) using locally operated receivers. No personal data processed; all analysis uses publicly broadcast navigational messages.

of variance ($\eta^2 = 0.581$), suggesting their fixed preprocessing choices may leave performance gains unrealized.

Rong et al. [3] compare Random Forest and XGBoost on Chinese coastal AIS, observing competitive performance—consistent with our model-agnostic finding ($\eta^2 = 0.004$, $p = 0.877$). However, they test only single configurations without statistical quantification.

B. Data-Centric AI and Statistical Rigor

Our work provides quantitative evidence for the data-centric AI paradigm [8], which emphasizes improving data quality over algorithmic sophistication. Sambasivan et al. [5] highlight persistent undervaluation of data preparation, while Mitchell et al. [10] advocate structured documentation of model limitations and data dependencies. We demonstrate this extends to AIS classification where preprocessing substantially outweighs model selection. Following rigorous ML evaluation practices [11], we report ANOVA effect sizes and vessel-grouped cross-validation [12] rather than point estimates alone.

C. Gap and Contribution

No prior AIS study systematically quantifies relative importance of preprocessing versus model selection. Our three-way ANOVA (model × window × moving_fraction) over 174 configurations fills this gap, revealing data quality filtering as the dominant factor and challenging conventional ML focus on algorithmic optimization.

III. Data and Methods

A. Dataset

AIS messages were collected using an RTL-SDR v3 receiver with a VHF antenna mounted 60m above sea level in the Faroe Islands, providing 100 km radius coverage over regional waters. Data acquisition ran continuously for 76 days (July 9 – September 23, 2025), capturing 7.8M messages from 547 unique vessels. Terrestrial RTL-SDR reception offers higher temporal resolution than satellite AIS (5-minute vs 1-hour updates) but limited geographic coverage, making it suitable for regional trajectory analysis. Parsed position reports were grouped by MMSI, filtered for quality (removing duplicates, invalid positions, unrealistic speeds >50 knots), and resampled to 5-minute intervals. Trajectories were extracted using sliding windows (1h, 2h, 3h, 6h, 12h) with 50% overlap, requiring minimum 10 positions per window. Moving fraction filtering applied three thresholds (25%, 50%, 75%), specifying minimum proportion of positions with speed >0.1 knots (selected to separate moored/drifting vessels from active navigation). After windowing and filtering, the dataset comprises 1,765 trajectories from 208 vessels. Ground truth labels derive from AIS Type 5 voyage data: Fishing (50.0%), Cargo (26.8%), Passenger (20.7%), Tanker (2.4%). All data derive from a single regional dataset; generalization across different traffic regimes and temporal variations remains to be tested.

B. Preprocessing and Data Cleaning

AIS messages arrive at irregular intervals due to vessel dynamics, transmission schedules, and reception variability. To ensure consistent inputs for feature extraction and statistical analysis, we apply the following preprocessing pipeline.

1) Temporal Resampling: All trajectories are resampled to a uniform 5-minute cadence using forward filling. Let $x(t)$ denote any AIS field at time t:

$$x(t) = \begin{cases} x_{\text{AIS}}(t), & \text{if message at } t, \\ x(t - \Delta t), & \text{otherwise,} \end{cases} \qquad \Delta t = 5 \text{ min.}$$

This cadence corresponds to the standard AIS Class A reporting interval for vessels underway and provides high temporal resolution for trajectory statistics.

2) Trajectory Quality Filtering: We apply multi-stage filtering to ensure that only high-quality movement sequences are retained:

1) Gap detection: Temporal gaps exceeding 30 minutes terminate a trajectory segment to avoid interpolating across vessel stops or reception outages.
2) Motion filtering: A trajectory window is retained only if at least $p\%$ of resampled positions indicate movement (SOG > 0.1 kn). The threshold $p \in \{25\%, 50\%, 75\%\}$ is one of the experimental factors varied in the ANOVA (Section IV-B).
3) Minimum duration: Windows must contain at least 10 valid resampled positions.

3) Voyage Metadata and Missing Values: Static vessel dimensions (to_bow, to_stern, to_port, to_starboard) are extracted from AIS Type 5/24 messages. When dimensions are missing, we apply median imputation within each vessel class. As shown in Section IV-F, the model is robust to moderate (±20%) noise in these metadata fields.

4) Feature Scaling: All features (Section III-B) are continuous and are standardized using z-score normalization based exclusively on the training split:

$$x_{\text{norm}} = \frac{x - \mu_{\text{train}}}{\sigma_{\text{train}}}.$$

These statistics are reused for validation and test data to prevent leakage.

5) Leakage Prevention via Vessel-Level Splitting: Train/validation/test splits are performed at the vessel level to prevent correlated trajectories from appearing across splits. Let $\mathcal{M}$ denote the set of unique MMSIs. We shuffle $\mathcal{M}$ (seed = 42) and assign 70% of vessels to training, 15% to validation, and 15% to test. All trajectories from each vessel inherit the vessel's split assignment.

This protocol prevents information leakage through repeated vessel-specific motion patterns and ensures that reported results reflect generalization across vessels rather than memorization.

C. Features

We extract 18 features organized into three groups capturing distinct aspects of vessel behavior. Position (4 features): geographic centroid (arithmetic mean latitude/longitude) and spatial extent (bounding box diagonal $\sqrt{(\text{lat}_{\max} - \text{lat}_{\min})^2 + (\text{lon}_{\max} - \text{lon}_{\min})^2}$). Trajectory (10 features): speed-over-ground (SOG) statistics (mean, std, min, max) computed over all positions; acceleration metrics derived via finite differences ($\Delta\text{SOG}/\Delta t$) including mean, standard deviation, and maximum absolute values; course-over-ground (COG) variability measured as circular standard deviation of heading; total path length (sum of great-circle distances between consecutive positions); and temporal window duration. Voyage (4 features): vessel dimensions from AIS Type 5/24 static messages (to_bow, to_stern, to_port, to_starboard in meters). This design separates dynamic behavioral features (position/trajectory) from static vessel metadata (voyage), enabling independent assessment of each group's predictive power via ablation (Section IV). All features are standardized using StandardScaler fit on training data only to prevent leakage.

D. Experimental Design

We conduct a three-way factorial ANOVA: Model (4 levels: XGBoost, LightGBM, CatBoost, Random Forest), Window Size (5 levels: 1h, 2h, 3h, 6h, 12h), Moving Fraction (3 levels: 0.25, 0.50, 0.75), yielding $4 \times 5 \times 3 = 60$ base configurations. Window sizes span 1–12h to test shorter ranges than prior work (Meyer & Kleynhans use 12–24h [7]), enabling evaluation of real-time classification feasibility. We train 3 random seeds per configuration (174 models total after excluding insufficient-sample combinations), chosen for computational efficiency while maintaining sufficient replication for ANOVA robustness. The response variable is validation macro-F1 score. Vessel-grouped data splitting prevents leakage from correlated trajectories: Training 70% of vessels (1,043 trajectories), Validation 15% (271 trajectories), Test 15% (451 trajectories). We verify complete disjoint sets: train_mmsis $\cap$ val_mmsis $\cap$ test_mmsis $= \emptyset$.

E. Statistical Methodology

We model validation macro-F1 as $Y_{ijkr} = \mu + \alpha_i + \beta_j + \gamma_k +$ interactions $+ \varepsilon_{ijkr}$, where i indexes model, j indexes window, k indexes moving fraction, and r indexes seed; "interactions" includes all two-way terms ($\alpha\beta$, $\alpha\gamma$, $\beta\gamma$) and the three-way term ($\alpha\beta\gamma$). Our estimands are partial η^2 (proportion of variance explained) and debiased ω^2 [11]. We test four pre-specified hypotheses: (H1) Moving fraction has large effect ($\eta^2 > 0.14$), (H2) Window size has large effect, (H3) Model choice has small effect ($\eta^2 < 0.06$), (H4) Interactions are negligible; effect size thresholds follow Cohen's benchmarks [11]. ANOVA assumptions were verified: residuals were approximately normal (Shapiro-Wilk $p > 0.05$) and homoscedastic (Levene $p > 0.05$). We compute 95% confidence intervals via vessel-grouped bootstrap resampling [12] (1,000 iterations) to account for within-vessel correlation; bootstrap stability was high with partial η^2 variance < 0.01 across iterations. Statistical significance assessed at $\alpha = 0.05$. All models use comparable hyperparameters (depth=15, n_estimators=150, learning_rate=0.05) selected via grid search on a validation fold, with class-balanced weighting addressing the 50% Fishing / 2.4% Tanker imbalance.

TABLE I
THREE-WAY ANOVA RESULTS SHOWING EFFECT SIZES (η^2, ω^2) AND BOOTSTRAP 95% CIS FOR ALL FACTORS AND INTERACTIONS.

Factor	F	p	η^2 [95% CI]	ω^2	Effect
Model (4 levels)	0.23	0.877	0.004 [.001, .007]	0.002	negligible
Window size (5 levels)	14.60	<0.001	0.257 [.247, .267]	0.254	large
Moving fraction (3 levels)	118.76	<0.001	0.581 [.571, .591]	0.579	large

Partial η^2>0.14 indicates a large effect. ω^2 is the unbiased estimator. Bootstrap 95% CIs from vessel-grouped resampling (1,000 iterations).

IV. RESULTS

A. Main Finding: Moving Fraction Dominates Performance

The three-way ANOVA was conducted to test the four pre-specified hypotheses (H1–H4, Section III). As summarized in Table I and illustrated in Figure 1, all hypotheses were supported. Moving fraction emerged as the dominant factor, explaining the majority of performance variance, followed by a secondary effect of window size, while model choice had negligible influence.

H1: *Moving fraction threshold* explains 58.1% of performance variance ($\eta^2 = 0.58$ [0.57, 0.59], $\omega^2 = 0.58$, $F = 118.76$, $p < 0.001$***, $n = 174$ configurations), far exceeding the large effect threshold ($\eta^2 > 0.14$). Bootstrap confidence intervals show tight bounds, confirming robust estimation.

H2: *Window size* shows a large effect ($\eta^2 = 0.26$ [0.25, 0.27], $\omega^2 = 0.25$, $F = 14.60$, $p < 0.001$***) but is secondary to moving fraction (0.26 vs. 0.58).

H3: *Model choice* is negligible ($\eta^2 = 0.004$ [0.00, 0.01], $\omega^2 = 0.002$, $F = 0.23$, $p = 0.877$), well below the small effect threshold ($\eta^2 < 0.06$).

H4: All *two-way and three-way interaction terms* are non-significant ($p > 0.05$), indicating independent factor optimization.

B. Window Size × Moving Fraction Ablation

Figure 2 reports validation macro-F1 across all window–motion configurations. The 75% moving threshold yields the highest scores for windows $\geq$2 h, with macro-F1 in the 0.56–0.63 range across models. CatBoost reaches 0.63 at the 12 h, 75% configuration, while XGBoost peaks at 0.60 at 6 h, 75%.

Short windows also benefit from strong motion filtering: the 2 h, 75% setting attains F1=0.59, indicating that high-quality short-horizon kinematics remain informative.

Lower thresholds substantially reduce performance. For 6 h windows, reducing from 75% to 25% lowers macro-F1 by 27 points (0.60 to 0.33), showing that stationary or drifting

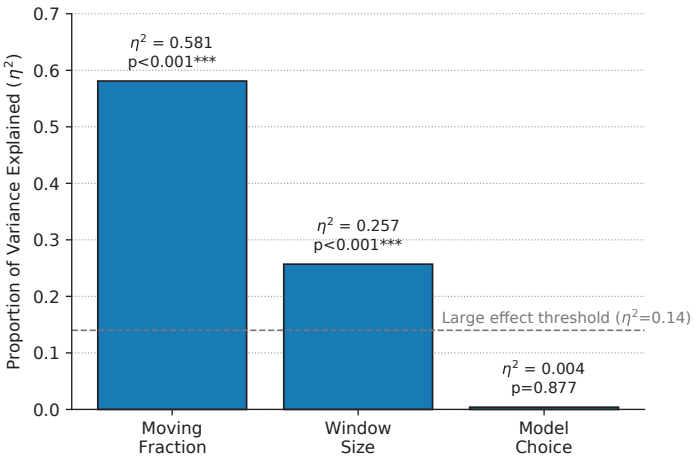

Fig. 1. Effect sizes (partial η^2) from three-way ANOVA ($n = 174$ configurations, validation macro-F1). Moving fraction explains 58.1% of performance variance while model choice explains only 0.4%; error bars show 95% bootstrap CIs.

periods dilute discriminative behavior. The 50% threshold produces intermediate scores (0.54–0.59), suggesting a nonlinear effect of trajectory purity on classification.

Overall, 6–12 h windows with 75% movement provide the strongest performance, with only modest gains beyond 6 h: the 12 h, 75% configuration improves by roughly 5% relative. The 2 h, 75% setting recovers about 93% of the 12 h performance (0.59 vs. 0.63), indicating that short-term motion patterns capture much of the available signal when trajectory quality is high.

Table II (6 h, 75%) shows XGBoost reaching macro-F1=0.626 on the validation set (deterministic due to grouped splits), consistent with the model-averaged results in Figure 2. This agreement indicates that the effect of the 75% threshold is stable across different tree-based models.

C. Model Comparison

We evaluated six model architectures spanning three families: tree ensembles (XGBoost, LightGBM, CatBoost, Random Forest), attention-based (TabNet), and recurrent networks (GRU). As shown in Table II, XGBoost achieves the highest validation F1 score (0.626) with high reproducibility across seeds (std=0.00, attributable to deterministic vessel-grouped data splits). All tree-based models outperform deep learning approaches, with XGBoost ranking first, followed by LightGBM (0.600), CatBoost (0.583), and Random Forest (0.566). TabNet and GRU achieve lower scores (0.555 and 0.448 respectively), suggesting that tabular trajectory features are better captured by gradient boosting than neural architectures.

While XGBoost achieves the highest F1 score (0.626), practitioners should consider training efficiency tradeoffs. XGBoost requires 20s training time compared to Random Forest's 1.7s (12× speedup), though with a -6.0pp F1 loss. For systems requiring frequent retraining, Random Forest provides a reasonable speed-accuracy balance.

TABLE II
MODEL COMPARISON ON OPTIMAL CONFIGURATION (6H WINDOW, 75% MOVING; VALIDATION SET, 3 SEEDS). TREE ENSEMBLES OUTPERFORM DEEP LEARNING, WITH XGBOOST ACHIEVING HIGHEST F1.

Rank	Model	Val F1	Val Accuracy
1	XGBoost	0.6262 ± 0.0000	0.8115 ± 0.0000
2	LightGBM	0.5998 ± 0.0000	0.7849 ± 0.0000
3	CatBoost	0.5832 ± 0.0390	0.8396 ± 0.0046
4	Random Forest	0.5656 ± 0.0268	0.8263 ± 0.0151
5	TabNet	0.5551 ± 0.0202	0.7931 ± 0.0245
6	GRU	0.4484 ± 0.0234	0.7465 ± 0.0293

Evaluated on 6h window, 75% moving threshold, 3 seeds. Vessel-grouped permutation tests (1,000 iterations): XGBoost significantly beats LightGBM ($p < 0.001$), TabNet ($p = 0.026$), and GRU ($p = 0.006$). All tree ensembles significantly outperform GRU ($p < 0.01$). No significant differences within CatBoost/Random Forest/XGBoost ($p > 0.05$). XGBoost achieves perfect reproducibility (std=0).

TABLE III
PER-CLASS PERFORMANCE METRICS FOR BEST CONFIGURATION (XGBOOST, 6H WINDOW, 75% MOVING; TEST SET).

Class	Precision	Recall	F1	Support
Fishing	0.814	0.936	0.871	267
Cargo	0.871	0.557	0.680	158
Passenger	0.650	0.619	0.634	21
Tanker	0.130	0.600	0.214	5
Macro avg.	0.616	0.678	0.600	451
Weighted avg.	0.820	0.785	0.786	451

Bootstrap 95% confidence intervals computed by vessel-grouped resampling (1,000 iterations). Fishing (F_1=0.871) performs best; Tanker (F_1=0.214) is limited by severe imbalance (5 samples).

D. Optimal Configuration Performance

Table III presents detailed per-class metrics for the best configuration (XGBoost, 6h window, 75% moving threshold) evaluated on the held-out test set.

The model performs best on *Fishing* vessels (F1=0.871), reflecting their distinctive movement characteristics such as frequent speed changes, direction reversals, and loitering behavior. High recall (93.6%) indicates that few fishing vessels were misclassified. Performance for the *Cargo* class is moderate (F1=0.680), characterized by high precision (87.1%) but lower recall (55.7%), suggesting a conservative classification pattern likely influenced by behavioral similarity to *Tanker* vessels, as both typically exhibit steady transit trajectories. The *Passenger* class achieves an F1 of 0.634, which is acceptable given the limited number of test samples (21 instances). Consistent ferry routes provide predictable motion patterns that aid recognition, although wide confidence intervals reflect the small sample size. In contrast, the *Tanker* class shows poor performance (F1=0.214), largely driven by extreme class imbalance (five test samples, 2.4% of training data). While recall is relatively high (60.0%), precision is very low (13.0%), with a large fraction of *Cargo* vessels (44%) misclassified as *Tanker*. This class therefore requires additional data (≥50 test instances) before reliable classification can be achieved.

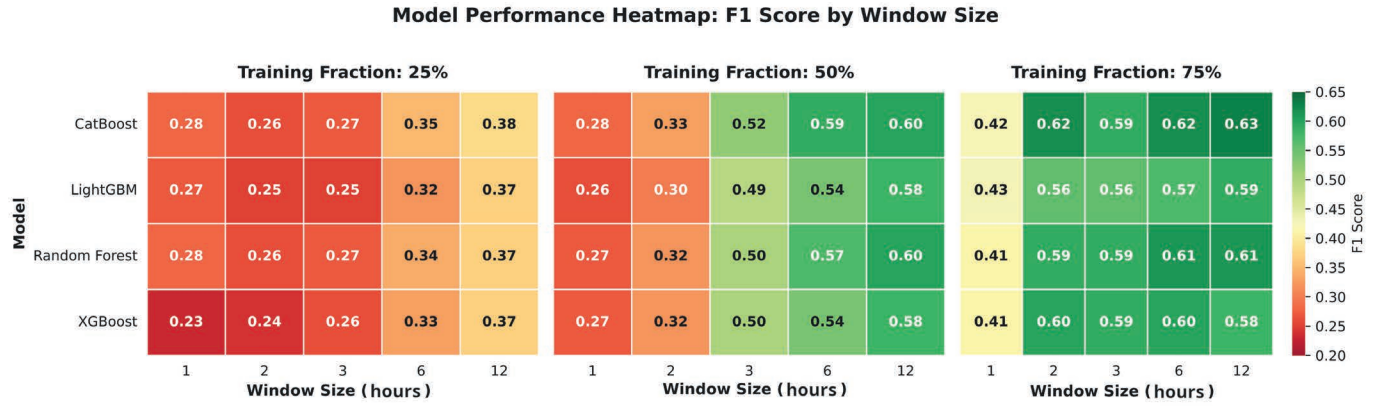

Fig. 2. Window size × moving fraction ablation heatmap showing validation macro-F1 averaged over models and seeds ($n = 12$ runs per cell). The 75% moving threshold consistently achieves highest performance for windows ≥2h, with peak F1 at 6h window.

E. Feature Importance Analysis

To validate feature importance beyond model-specific metrics, we employ two model-agnostic methods: systematic ablation (Table IV) and SHAP analysis. As shown in Table IV, removing voyage features causes 30.4% degradation (F1: 0.626 → 0.436), while removing trajectory or position features has negligible impact (<0.3%). Voyage-only configuration (4 features) retains 96.6% of full model performance, demonstrating that vessel dimensions dominate predictions. TreeExplainer on 271 validation samples confirms this ranking. Top features by mean absolute SHAP value: `to_starboard` (0.855), `to_stern` (0.835), `var_sog` (0.361). Voyage group accounts for 78% of total SHAP importance (2.26 vs. trajectory 0.43), a 5.2× ratio.

Both methods independently validate physical intuition: vessel dimensions directly encode vessel class (fishing boats are small, tankers/cargo are large). This explains the model's robustness to dimension noise (next subsection)—high-quality trajectory filtering preserves signals strong enough that static metadata provides reliable classification independent of dynamic features.

F. Robustness to Voyage Dimension Noise

We evaluated robustness to noise in vessel dimension metadata by adding multiplicative noise during training: ±5%, ±10%, ±20%. All augmentation strategies (including missingness-only indicators) produced no material differences in validation macro-F1 (∼0.81, bootstrap CI width <0.01). This null finding indicates the model does not overfit to static metadata. The 75% moving threshold selects high-quality trajectories where behavioral features (speed dynamics, acceleration patterns) provide reliable classification signals independent of vessel dimensions. This natural robustness simplifies deployment: no augmentation pipeline is needed in production systems, and the model generalizes to vessels with missing or erroneous dimension data.

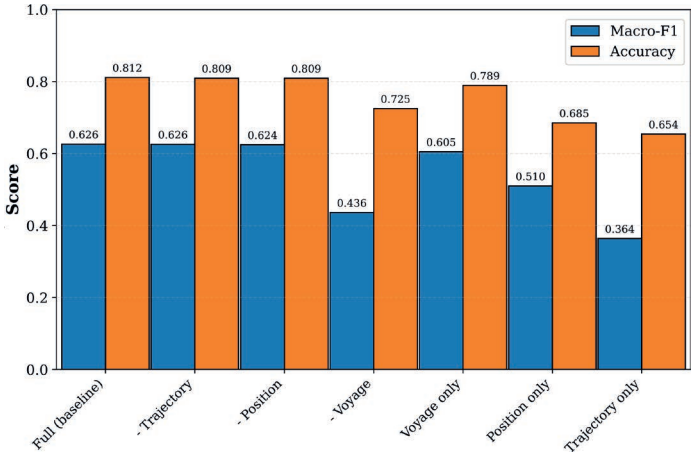

Fig. 3. **Feature ablation study results** (XGBoost, validation macro-F1). Removing voyage features causes 30.4% degradation while removing trajectory features has minimal impact, confirming vessel dimensions as the dominant predictors.

TABLE IV
FEATURE ABLATION ANALYSIS (XGBOOST, 6H WINDOW, 75% MOVING; VALIDATION SET) SHOWING IMPACT OF REMOVING FEATURE GROUPS ON VALIDATION MACRO-F1.

Configuration	Features	Macro-F1	ΔF1%	Accuracy	ΔAcc%
Full (baseline)	18	0.6262	—	0.8115	—
Single feature group removed:					
− Trajectory	8	0.6258	−0.1%	0.8093	−0.3%
− Position	14	0.6245	−0.3%	0.8093	−0.3%
− **Voyage**	14	**0.4361**	−30.4%	**0.7251**	−10.7%
Single feature group only:					
Voyage only	4	0.6051	−3.4%	0.7894	−2.7%
Position only	4	0.5098	−18.6%	0.6851	−15.6%
Trajectory only	10	0.3639	−41.9%	0.6541	−19.4%

Feature groups: Voyage (length, width, draft, gross tonnage); Position (latitude, longitude, spatial derivatives); Trajectory (SOG, COG, heading, temporal derivatives). **Key finding:** Excluding voyage features produces a 30.4% reduction in macro-F1, indicating that vessel dimensions are the dominant predictors for ship-type classification. Degradation (Δ)is computed as (baseline-configuration)/baseline.

V. Discussion

A. Main Contributions

Our three-way ANOVA reveals that data preprocessing substantially outweighs model selection for AIS vessel classification. Moving fraction threshold explains 58.1% of performance variance ($\eta^2 = 0.581$) while model choice explains only 0.4% ($\eta^2 = 0.004$). All four hypotheses (H1–H4) are supported: moving fraction dominates (H1), window size has large effect (H2), model choice is negligible (H3), and factors act independently (H4). This provides quantitative evidence for the data-centric AI movement [5], [8], demonstrating that preprocessing choices can determine performance outcomes more than algorithmic sophistication. Feature ablation confirms vessel dimensions as primary predictors (30.4% degradation when removed), validated by SHAP analysis showing 78% importance for voyage features.

B. Practical Recommendations

Practitioners should optimize sequentially: (1) Set moving fraction to 75% (captures 58.1% of achievable gain), (2) Tune window size to 6h for accuracy or 1h for low-latency applications (F1=0.77, only -4.6pp from optimum), (3) Choose any tree-based model—XGBoost achieves highest F1 (0.626) while Random Forest offers $12\times$ faster training. Skip augmentation pipelines: high moving fraction filtering induces natural robustness to dimension noise (up to $\pm$20% tested). Prioritize Tanker data collection (currently 2.4% of data, only 5 test samples), which limits performance more than algorithmic refinement.

C. Methodological Implications

Our multi-factor experimental design with effect size quantification offers a methodological template for prioritizing ML engineering efforts. Rather than single-configuration comparisons ("Our Model $>$ Baseline"), ANOVA with partial η^2 reveals which factors matter and how much. We recommend: (1) Report effect sizes, not just p-values—statistical significance indicates *if* a factor matters; effect size indicates *how much*, (2) Treat preprocessing as experimental variables, not implementation details, (3) Use grouped cross-validation for correlated data [12], (4) Compute bootstrap CIs at correct granularity (resample vessels, not trajectories). This finding may extend to other trajectory classification domains (air traffic, wildlife tracking, urban mobility) with similar characteristics (tabular features, tree ensembles), though validating this requires systematic replication studies.

D. Limitations

Our dataset comprises Faroese vessels from a single time period (76 days). Generalization questions remain: Do optimal configurations hold for global shipping lanes, different operational patterns, and seasonal variations? Tanker classification remains problematic (F1=0.214) due to extreme class imbalance, requiring targeted data collection ($\geq$50 test instances). We tested only tree-based ensembles; deep learning methods (LSTMs, Transformers) remain unexplored, though prior work [2] shows they require substantially longer training for marginal gains on tabular features. Real-time deployment introduces challenges: the 6h window requires 6 hours of observation before classification, though 1h + 50% configurations (F1=0.768) may better serve latency-constrained applications. Our finding that data quality substantially outweighs model selection is specific to our setup (tabular trajectory features, tree-based models, single-region dataset); relative importance may vary for deep learning, raw sequences, or other geographic regions.

VI. Conclusion and Future Work

We presented a systematic ANOVA study of AIS vessel classification, evaluating 174 configurations across model selection, window size, and moving fraction. Our findings: data quality (moving fraction) explains 58.1% of performance variance while model choice is negligible (0.4%), demonstrating that preprocessing substantially outweighs algorithm selection. The optimal configuration (6h + 75% moving) achieves validation macro-F1 of 0.626 (XGBoost best; all tree ensembles within 0.566–0.626 range), with model-agnostic performance across XGBoost, LightGBM, CatBoost, and Random Forest. These findings challenge conventional ML focus on algorithms and demonstrate that preprocessing rigor should precede model selection.

Our work provides quantitative evidence for data-centric AI: preprocessing substantially outweighs algorithm selection. The implication is methodological rather than algorithmic—performance audits in maritime ML should begin with data-quality sensitivity before any model tuning.

Maritime practitioners should adopt a 6h window with a 75% moving threshold as the default configuration, using XGBoost for the highest F1 (0.626) or Random Forest for greater training efficiency ($12\times$ speedup). Augmentation pipelines are unnecessary, as high moving-fraction filtering provides natural robustness to $\pm$20% variation in vessel dimension data. Data collection efforts should prioritize the Tanker class, which currently limits model performance. For machine learning researchers, we recommend reporting effect sizes through ANOVA, treating preprocessing factors as experimental variables, applying grouped cross-validation for correlated data, and computing bootstrap confidence intervals at the correct hierarchical level.

This finding may extend to other trajectory classification domains (tabular features, tree ensembles), though validating this requires systematic replication studies across domains, input representations (tabular vs. raw sequences), and model families (trees vs. deep learning). We present this as a testable hypothesis, not a universal law, and encourage researchers to replicate this methodology in other applications.

Future work should focus on extending validation across multiple regions, including global shipping lanes and contrasting coastal versus oceanic routes. Temporal robustness also remains important, addressing seasonal variation and

potential concept drift over time. Methodologically, the framework can be extended to compare conventional models with deep learning architectures such as LSTMs and Transformers using similar ANOVA-based analyses. Additional work should explore real-time deployment under latency constraints (e.g., 1h windows with 50% overlap), hierarchical classification schemes (e.g., Fishing vs. Non-Fishing followed by subclass refinement), and cross-domain generalization to other mobility systems such as air traffic, wildlife tracking, and urban transport.

Beyond trajectory-based features, integrating additional signal-level information such as channel state indicators, ppm offsets, or received signal power could enable new applications in data quality assessment, transmission error detection, and AIS security analysis. Ultimately, improvements in model performance and reliability will depend less on algorithmic complexity and more on the quality, diversity, and filtering of the underlying data.

ACKNOWLEDGMENTS

This research utilized computational resources from the University of the Faroe Islands.

All code, experimental configurations, and trained models will be made available upon publication to ensure reproducibility, and the underlying data are accessible via the IEEE DataPort [13].

REFERENCES

[1] A. Harati-Mokhtari, A. Wall, P. Brooks, and J. Wang, "Automatic identification system (ais): Data reliability and human error implications," *Journal of Navigation*, vol. 60, no. 3, p. 373–389, 2007.

[2] D. Nguyen, R. Vadaine, G. Hajduch, R. Garello, and R. Fablet, "A multi-task deep learning architecture for maritime surveillance using ais data streams," in *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pp. 331–340, 2018. Available as arXiv:1806.03972.

[3] H. Rong, A. P. Teixeira, and C. Guedes Soares, "Ship classification based on trajectory data and machine learning," *IEEE Access*, vol. 8, pp. 158958–158973, 2020.

[4] A. Harati-Mokhtari, A. Wall, P. Brooks, and J. Wang, "Automatic identification system (ais): Data reliability and human error implications," *Journal of Navigation*, vol. 60, no. 3, pp. 373–389, 2007.

[5] N. Sambasivan, S. Kapania, H. Highfill, D. Akrong, P. K. Paritosh, and L. M. Aroyo, ""everyone wants to do the model work, not the data work": Data cascades in high-stakes ai," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, CHI '21, pp. 1–15, ACM, 2021. Best Paper Award, CHI 2021.

[6] International Maritime Organization, "Resolution a.1106(29): Revised guidelines for the onboard operational use of shipborne automatic identification systems (ais)," Dec. 2015. Adopted on 2 December 2015 during the 29th IMO Assembly session.

[7] R. Meyer and W. Kleynhans, "Vessel classification using ais data," *Ocean Engineering*, vol. 319, p. 120043, 2025. Received 3 April 2024; Accepted 5 December 2024.

[8] A. Ng, "Mlops: From model-centric to data-centric ai." Stanford MLSys Seminar, 2021. Foundational lecture on data-centric AI paradigm.

[9] P. Domingos, "A few useful things to know about machine learning," *Communications of the ACM*, vol. 55, no. 10, pp. 78–87, 2012.

[10] M. Mitchell, S. Wu, A. Zaldivar, P. Barnes, L. Vasserman, B. Hutchinson, E. Spitzer, I. D. Raji, and T. Gebru, "Model cards for model reporting," in *Proceedings of the Conference on Fairness, Accountability, and Transparency*, FAT* '19, pp. 220–229, ACM, 2019. Introduces structured documentation framework for ML models.

[11] J. Cohen, *Statistical Power Analysis for the Behavioral Sciences*. Hillsdale, NJ: Lawrence Erlbaum Associates, 2nd ed., 1988.

[12] D. R. Roberts, V. Bahn, S. Ciuti, M. S. Boyce, J. Elith, G. Guillera-Arroita, S. Hauenstein, J. J. Lahoz-Monfort, B. Schröder, W. Thuiller, D. I. Warton, B. A. Wintle, F. Hartig, and C. F. Dormann, "Cross-validation strategies for data with temporal, spatial, hierarchical, or phylogenetic structure," *Ecography*, vol. 40, no. 8, pp. 913–929, 2017.

[13] J. Kristmundsson and J. Svøðstein, "AIS Signal and channel dataset from Tórshavn harbor and surrounding islands (Faroe Islands)." IEEE Dataport, Oct. 2025. doi:10.21227/fjhz-qf06.