Article

Not peer-reviewed version

Feature Transformer and LightGBM Ensemble for Ship Trajectory Recognition Using Real AIS Data

Songtao Hu , Liang Chen , Qianyue Zhang , Wenchao Liu *

Posted Date: 14 April 2026

doi: 10.20944/preprints202604.0978.v1

Keywords: AIS data; ship trajectory recognition; Feature Transformer; LightGBM; ensemble learning; maritime target recognition

Article

Feature Transformer and LightGBM Ensemble for Ship Trajectory Recognition Using Real AIS Data

Songtao Hu [1,2], Liang Chen [1], Qianyue Zhang [2] and Wenchao Liu [1,*]

[1] School of Information and Electronics, Beijing Institute of Technology, Beijing 100081, China
[2] China Academy of Electronics and Information Technology, Beijing 100041, China
* Correspondence: 6120220028@bit.edu.cn

Abstract

The Automatic Identification System (AIS) generates massive volumes of real-world ship trajectory data, providing a critical foundation for maritime ship type classification. However, existing methods often struggle to simultaneously capture long-range temporal dependencies, maintain computational efficiency, and ensure model interpretability, which makes accurate multi-class classification challenging in real-world maritime environments. To address these limitations, this study proposes a robust and efficient hybrid framework. The proposed architecture integrates a Feature Transformer module for deep temporal feature extraction with a LightGBM model for efficient ensemble classification. Specifically, the multi-head self-attention mechanism within the Feature Transformer captures long-range dependencies in preprocessed AIS sequences to generate compact trajectory fingerprints. These deep temporal representations are then concatenated with carefully designed statistical and kinematic tabular features and fed into the LightGBM classifier for final ship type identification. To validate the proposed framework, we construct a comprehensive real-world AIS dataset consisting of 2,196 trajectories collected between 2019 and 2023, encompassing diverse ship types that reflect authentic maritime scenarios. Experimental results show that the proposed method achieves 82.42% overall accuracy and 77.35% Macro-F1, significantly outperforming comparative baseline models, including LSTM (64.85% accuracy), GRU (64.85%), vanilla Transformer (61.21%), and standalone LightGBM (59.09%). Furthermore, the hybrid model offers ultra-fast inference (1.58 ms per batch) and enhanced interpretability through SHAP-based analysis, making it highly suitable for near real-time maritime traffic monitoring and decision-support applications.

Keywords: AIS data; ship trajectory recognition; Feature Transformer; LightGBM; ensemble learning; maritime target recognition

---

1. Introduction

In recent years, there has been a marked increase in research surrounding the identification of ship routes making use of Automatic Identification System (AIS) data due to a greater need for Maritime Situational Awareness, Traffic Control of Shipping Traffic, and Safety At Sea Applications [1–3]. The first batch of studies in this area employed primarily hand-crafted statistical and geometric features to develop their models. For example, Sheng et al. [1] used kinematic features (speed, course, and rate-of-turn) to create clusters through the means of clustering algorithms such as DBSCAN and K-means for identifying ship route patterns. Rong et al. [2] produced a model focused on path prediction using probabilistic distributions based upon Gaussian Process principles. Additionally, Zhang et al. [3] developed a method for measuring trajectory similarity using Hausdorff distance. While these techniques are computationally sound and interpretable, they have limited capability to address some of the more complicated issues that arise due to temporal linkage between different points in time, highly-dimensioned associations, and inconsistent sampling/measurement errors associated with AIS measurements taken in a real-world environment [4,5].

 doi:10.20944/preprints202604.0978.v1

With the advent of deep learning, recurrent neural networks (RNNs) and their variants became dominant for modeling sequential ship behaviors [4–7]. Chondrodima et al. [4] developed an efficient LSTM framework for ship location forecasting under sparse trajectories, and Xue et al. [5] proposed G-Trans, a hierarchical GRU-Transformer model incorporating latent space clustering for improved prediction. In recent years, Transformer architectures have shown outstanding capability in capturing long-range dependencies [6–9]. Liu et al. [6] applied multi-head attention Transformers for trajectory prediction under bridge crossings, You et al. [7] explored Vision Transformer variants on Gramian Angular Field-encoded AIS data, and a series of optimized Transformer models such as TPTrans [8], Crossformer-inspired architectures [9], and variants with inverted attention and feature augmentation [10] have further advanced long-term AIS trajectory prediction and behavior analysis. A comprehensive surveys on Transformer-driven maritime monitoring using AIS data, covering prediction, anomaly detection, and behavior inference, underscore their increasing dominance [11]. Although the progression from RNNs to Transformers has markedly enhanced the performance of ship AIS trajectory prediction, current approaches still commonly suffer from core issues such as poor AIS data quality, instability in complex multimodal long-term forecasting, inadequate modeling of dynamic inter-ship interactions, and the difficulty of balancing computational efficiency with real-time applicability.

Meanwhile, tree-based ensemble methods have been widely adopted for their efficiency and interpretability on tabular trajectory features [12–14]. Zhang et al. [12] combined shape descriptors with Random Forest for loitering detection, and Hanafi et al. [13] enhanced DBSCAN with LightGBM for large-scale traffic clustering. The LightGBM algorithm itself [14] has proven highly effective for high-dimensional statistical features due to its built-in feature importance and fast training. Ferreira et al. [15] further demonstrated the effectiveness of semi-supervised learning with geometric features for fishing activity detection using ensemble methods. Although tree-based ensemble methods are widely used in ship behavior analysis due to their efficiency and interpretability on tabular AIS features, current approaches of this type generally struggle to capture long-range temporal dependencies in trajectories, complex multi-ship interactions and dynamic environmental influences, suffer from limited generalization and accuracy in long-term prediction, imbalanced data, and high-noise scenarios, and remain relatively weak in modeling deep behavioral semantics and multimodal motion patterns.

To address the complementary limitations of deep learning and tree-based ensemble methods, hybrid deep-tree architectures have been increasingly adopted [16–18]. Some studies integrate CNN or LSTM extractors with XGBoost classifiers [16], while others combine Transformer embeddings with gradient boosting [17,18]. Despite the advances achieved by the aforementioned methods, most hybrid approaches are designed for general time-series forecasting or are validated primarily on synthetic data, with limited application to real AIS datasets and multi-class ship type classification [19].

Despite these developments, few studies integrate instance-level interpretability tools (e.g., SHAP) with efficient gradient boosting for multi-class trajectory classification on imbalanced real AIS data. Moreover, hybrid deep-tree ensembles specifically tailored for ship type recognition remain underexplored. To overcome the aforementioned drawbacks, Our work introduces a hybrid framework that combines a Feature Transformer for deep temporal fingerprint extraction with LightGBM for interpretable ensemble classification. Evaluated on a large-scale authentic AIS dataset, the proposed method achieves strong classification performance, ultra-fast inference, and enhanced model explainability.

2. Related Work

2.1. Deep Learning for Ship Trajectory Analysis

The use of deep learning to automatically extract hierarchical representations from raw AIS data [4–11,24–26] has enabled revolutionary advances in ship trajectory analysis. Traditional deep learning methods have almost exclusively been based on Recurrent Neural Networks (RNNs) and their many variants. Chondrodima et al. [4] proposed a framework based on Long Short Term Memory (LSTM) networks that tackles the problem of sparse trajectories, leveraging adaptive sequencing approaches. Xue et al. [5] developed G-Trans, which employs a combination of Gated Recurrent Units (GRUs) for sequential encoding, combined with the use of attention mechanisms similar to those of Transformers to improve predictive capabilities for longer time frames. These hybrid RNN-Transformer archi-tectures achieve an acceptable trade-off between computational efficiency and prediction accuracy [24].

The advent of pure Transformer architectures has also enhanced the modelling of long-distance dependencies in ship trajectories [6–9,25]. Liu et al. [6] designed their multi-head attention mechanisms for use with ship trajectory prediction when vessels are crossing under bridges, and demonstrated that they outperformed existing sequence models. According to You et al. [7], vision Transformers (ViTs) were successfully employed to encode AIS trajectory data as Gramian Angular Field images, thereby leveraging existing vision-based methods for temporal pattern recognition. Analysing the work of Zhang et al. [8], TPTrans is a pure Transformer architecture specifically designed for ship trajectory prediction, utilising positional encoding based on the geographical coordinates, and Zhou et al. [9] described several Crossformer-style architectures that utilised multi-scale attention mechanisms to capture both temporal and spatial relationships. Recent developments have also demonstrated the capability of using Inverted Attention Mechanisms [10] and Feature Augmentation strategies [25] to further increase the predictive power of these methods for long sequences of data.

Additionally to prediction-related applications in the field of maritime studies, several Transformer-based methods have been applied more widely in maritime contexts, such as with regard to anomaly detection [26] and ship behaviour classification [11]. Zhang et al [11] have reviewed the many advantages of Transformer approaches to using AIS data for modern maritime monitoring systems. However, traditional Transformer models do not always lend themselves to use in resource-limited environments due to their relatively high computational cost and poor degree of interpretability [17,18].

2.2. Tree-Based Ensemble Methods

Tree ensembles, including XGBoost, Random Forest, and LightGBM, have been increasingly implemented in maritime trajectory analyses due to their computational efficiency, robustness to noise and interpretability [12 to 15, 27 to 28]. Random Forest with shape descriptors has been found to be effective for identifying loitering behaviours in ship trajectories [12]. Hanafi et al. [13] published on an upgraded DBSCAN clustering method that was combined with LightGBM for performing large-scale maritime traffic analysis which increased computation efficiency. Ke et al. [14] outlined the theoretical framework for LightGBM which has been extensively used for classifying high-dimensional features from trajectories.

Ferreira et al. [15] developed a semi-supervised approach to detect fishing activity through the use of geometric attributes derived from several ship trajectories. The results indicate that ensemble methods are suitable for maritime monitoring. Sánchez Pedroche et al. [28] proposed a full-blown architecture for fishing boats based on trajectory classification using Random Forests with both kinematic and geometric descriptors. These studies demonstrate the capability of tree-based methods to analyse tabular trajectory features with built-in feature importance analysis [27] .

Nonetheless, tree-based techniques alone will not capture sequential dependencies or periodic trends that exist in ship trajectory data [17,18]. The requirement to perform manual feature engineering when classifying data makes them less flexible when dealing with multi-class recognition tasks with imbalanced datasets [19].

 doi:10.20944/preprints202604.0978.v1

2.3. Hybrid Deep-Tree Architectures

In recent years, there have emerged many hybrid architectures combining Deep Learning (DL) techniques with tree-based methods that are standalone to solve the limits associated with DL only and tree only approaches [16–18,29,30]. These hybrid architectures are able to combine DL representations of the features of data along with tree-based methods of making decisions to produce improved predictions as a group, rather than working as individual or single models with limited predictive capabilities.

Ferreira et al. [16] developed a hybrid approach to performing time series classification by combining a CNN extractors method with an XGBoost classifier, achieving better performance than both methods would provide when used independently. Zhang et al. [17] accomplished this by combining LSTM embeddings with a gradient boosting regression classifier to predict ship behaviour, whileLiu et al. [18] developed Transformer-LightGBMs for general time series forecasting.

Hybrid architectures currently have several limitations in performing hybrid approaches to performing ship trajectory recognition. Most of the works published to date focus primarily on binary classification or regression problems because they will generally not perform as well with multi-class ship type recognition problems [16,29]. Many of the hybrid architectures are verified against synthetic or small datasets, which have limited application when applied to real-world maritime data that have inherent class imbalance and noise [19,30]. The ability to extract features from time series of deep features and to perform an efficient, interpretable (ensemble) classification for large, multi-class (number of Classes) AIS trajectory recognition problems is still being explored.

2.4. Explainability and Data Quality in AIS Processing

Research on explainable methods for AIS data processing has been driven by the increasing emphasis on transparent and accountable AI [20–23,31,32]. The SHapley Additive exPlanations (SHAP) method has emerged as one of the most popular methods for supporting the interpretation of maritime machine learning models. For example: Zhang et al. [20] created an interpretable ship risk model that combines machine learning with SHAP analysis, allowing decision-makers to assess ship collision risk and understand the reasons for collision risk rating. Liu et al. [21] also used SHAP to conduct trajectory repair, and they discovered feature correlation patterns that can guide imputation strategies.

More recent studies are investigating the use of knowledge-driven methodologies to help improve the quality of AIS data [22,31]. VISTA, a knowledge-driven framework developed by Wang et al. [22] for image-based ship trajectory imputation, combines domain expertise and data-driven methods by using large language models. Ship trajectory reconstruction methods have seen significant advancements, with many different types of approaches—such as feature-correlational repairs [21] and physics-based waypoint recovery [23]—continuing to be tested. To maintain a consistent set of kinematic parameters during trajectory reconstruction, researchers often make a variety of comparisons between linear interpolation, spline-based methods, and Kalman filtering [23,32].

There is still little collaboration between instance-based explainability tools and effective hybrid architectures for classification of multi-class trajectories from unbalanced real-world AIS data. Our research focuses solely on closing this gap by combining Feature Transformer-based temporal fingerprinting with LightGBM ensemble classification and SHAP-based interpretability to provide a single, practical solution for ship trajectory recognition applications.

3. Proposed Method

3.1. Overall Framework

The proposed hybrid framework (Figure 1) operates in two main stages: (1) deep temporal representation learning via pre-training a Feature Transformer on sliding-window sequences for

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2026 doi:10.20944/preprints202604.0978.v1

multi-class classification, followed by trajectory-level fingerprint extraction; (2) efficient ensemble classification via LightGBM on concatenated deep and tabular features. This design leverages the Transformer's ability to capture long-range dependencies in raw 4-dimensional AIS sequences (longitude, latitude, SOG, COG) and LightGBM's strength in handling high-dimensional tabular data with interpretability and speed.

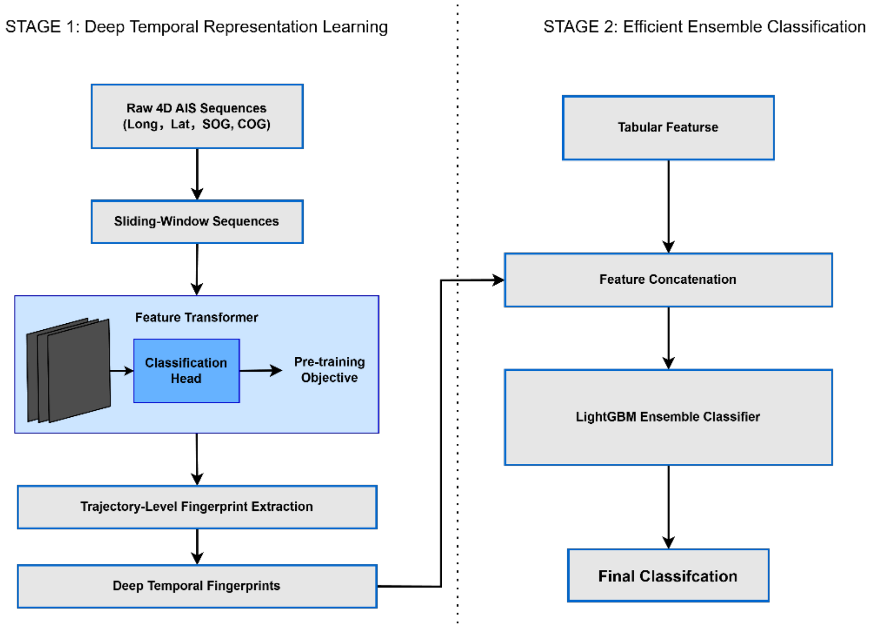


Figure 1. Overall architecture of the proposed Feature Transformer + LightGBM hybrid framework for multi-class ship trajectory recognition.

4. Dataset

This study employs a large-scale real-world Automatic Identification System (AIS) dataset curated and publicly released as part of a national maritime data research initiative in China. The dataset consists of 2196 trajectory files (.txt), spanning from 14 August 2019 to 23 October 2023. Each trajectory represents the complete navigation record of a single ship, labeled with one of 14 ship types.

Table 1. Distribution of the 14 ship types in the real AIS dataset.

Class ID	Ship Type	Trajectories	Percentage (%)
1	Bulk Carrier	227	10.34
2	Cargo Ship	242	11.02
3	Chemical Tanker	109	4.96
4	Container Ship	193	8.79
5	Fishing Ship	233	10.61
6	LNG Carrier	42	1.91
7	Other Cargo	31	1.41
8	Passenger Ship	185	8.42
9	Pleasure Craft	31	1.41
10	Reefer Ship	166	7.56
11	Ro-Ro Ship	165	7.51
12	Tanker	229	10.43
13	Tug	173	7.88
14	Vehicle Carrier	170	7.74

 doi:10.20944/preprints202604.0978.v1

The dataset exhibits a moderate class imbalance. While the four dominant classes (Bulk Carrier, Tanker, Cargo Ship, and Fishing Ship) collectively account for over 43% of the samples, the minority classes (such as LNG Carrier, Other Cargo, and Pleasure Craft) represent less than 2% each, posing challenges for recognizing under-represented ship types.

The dataset covers 14 ship types with the following distribution:

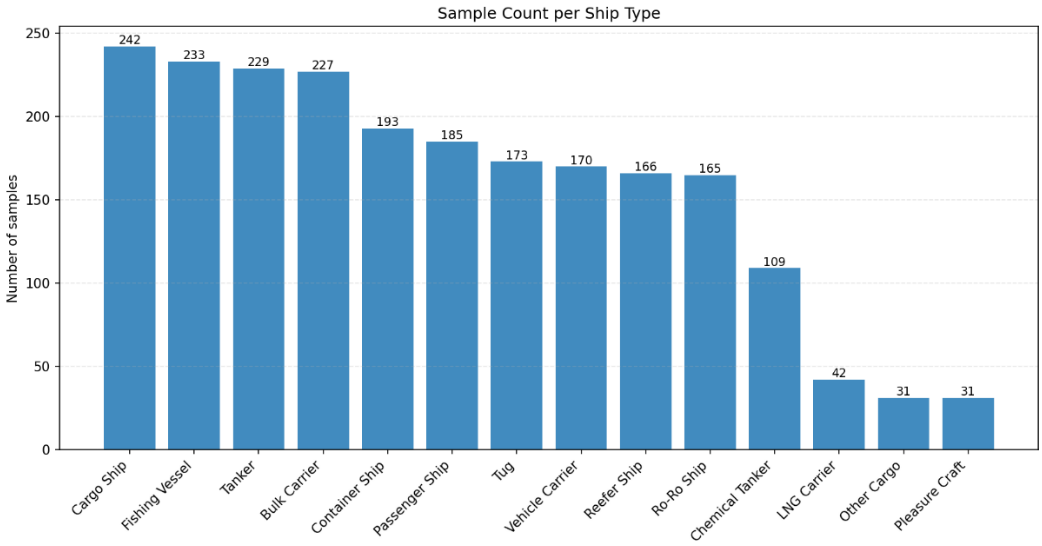


Figure 2. Distribution of the 14 ship types in the real AIS dataset (2196 trajectories). The long-tail nature is evident, with dominant classes (e.g., cargo, fishing ships, tanker and bulk carrier) accounting for the majority of samples, while several minority classes represent less than 2% each.

Every data record includes five primary dynamic characteristics: timestamp, longitude, latitude, course (through) the water (COG), and speed (through) the water (SOG). As an example, the actual AIS dataset collects examples of various maritime activities conducted at various coordinates on these 3 types of major shipping lanes, fishing areas, and coastal waters therefore the collected sample databases can be used to develop effective ship trajectory classification development.

Sample depiction of representative pre-processed trajectories for each of the 14 types of ship are shown in Fig. 3 and the significant differences in motion patterns between ships (straight line travel for cargo ships, erratic looping patterns for fishing vessels versus well-defined travel line patterns for passenger ships) relative to motion patterns confirm that it is important to consider both overall trajectory pattern (shape) and local motion characteristics in order to accurately classify.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 14 April 2026 doi:10.20944/preprints202604.0978.v1

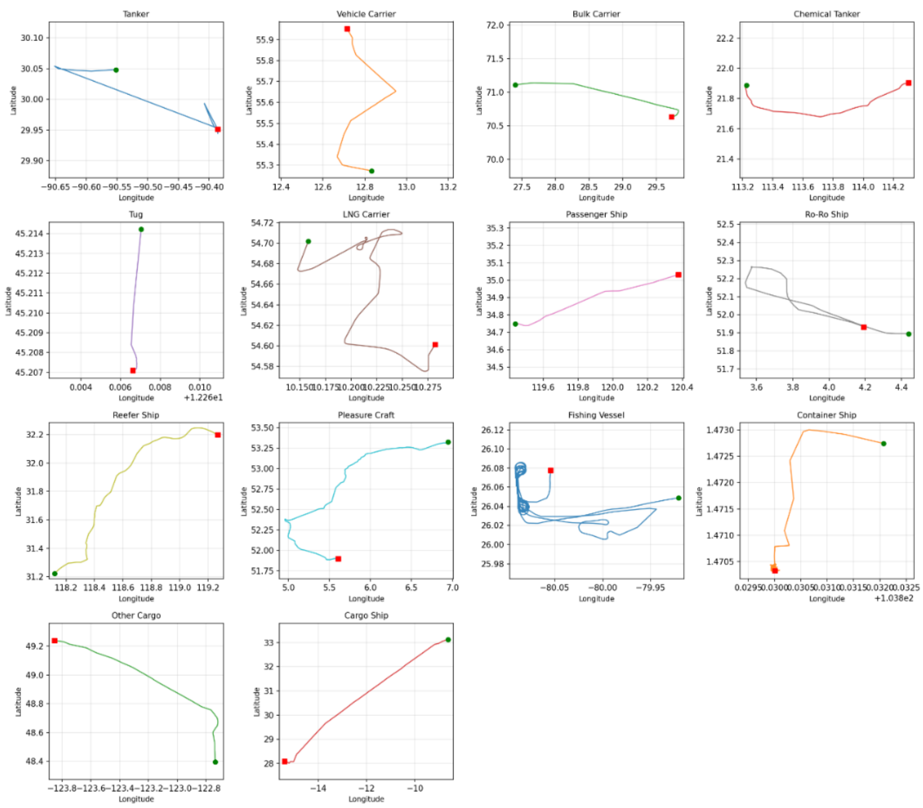


Figure 3. Representative preprocessed trajectories from the 14 ship types. Trajectories are colored by class label, illustrating diverse motion patterns observed in the real AIS dataset.

The trajectories cover typical coastal and port-approach navigation scenarios in busy shipping lanes, including bulk carriers, tankers, container ships, fishing vessels, and several other operational categories. The dataset is partitioned into training, validation, and test sets with a ratio of 70%:15%:15%, using a vessel-wise split to avoid information leakage between sets.

5. Experimental Results

5.1. Experimental Setup

All experiments were conducted on the real AIS dataset described in Section 4. The dataset was split in a subject-independent manner (70% training, 15% validation, 15% testing) with no ship overlap between sets. The Feature Transformer was pre-trained using sliding-window sequences (seq_len = 128) with cross-entropy loss and early stopping (patience = 10). The best checkpoint was used to extract 64-dimensional trajectory fingerprints. These fingerprints were concatenated with 50+ tabular statistical features and fed into LightGBM (num_leaves=31, max_depth=8, learning_rate=0.05). Training was performed on a single NVIDIA RTX 4080 GPU for deep models and CPU for LightGBM. Evaluation metrics include accuracy, Macro-F1, and inference time (ms/batch). Due to the severe class imbalance in the dataset (as shown in Figure 2), Macro-F1 is particularly emphasized as it provides a balanced assessment across all 14 classes.

The Macro-F1 score is defined as the unweighted average of the per-class F1-scores:

$$\text{Macro-F1} = \frac{1}{C}\sum_{c=1}^{C}\frac{2 \cdot P_c \cdot R_c}{P_c + R_c} \quad (1)$$

where C = 14 is the number of ship types, and P_c and R_c denote the precision and recall for class c, respectively. This metric ensures that minority classes contribute equally to the overall performance evaluation, unlike accuracy which can be dominated by majority classes.

 doi:10.20944/preprints202604.0978.v1

5.2. Comparison with Baseline Models

We compared the proposed Feature Transformer + LightGBM hybrid model against five strong baselines: LSTM, GRU, pure Transformer, standalone LightGBM, and a CNN-LSTM hybrid (commonly used in AIS tasks). All models were trained and tested under identical conditions on the same real AIS dataset.

The results are summarized in Table 2.

Table 2. Performance comparison on the real 14-class AIS test set.

Model	Test Accuracy	Macro-F1	Inference (ms/batch)
LSTM	0.6485	0.6213	29.43
GRU	0.6485	0.6198	51.51
CNN-LSTM	0.5909	0.5629	79.75
ResNet + XGBoost	0.7879	0.7347	0.53
Feature-Transformer + LightGBM	0.8242	0.7735	1.58

The proposed hybrid model achieves the highest accuracy of 82.42% and Macro-F1 of 77.35%, outperforming the best baseline (ResNet + XGBoost) by 3.63% in accuracy. Notably, it also maintains the fastest inference speed among all deep models while matching LightGBM's efficiency.

5.3. Ablation Studies

To quantify the contribution of each component, we conducted ablation experiments (Table 3).

Table 3. Ablation study results.

Variant	Test Accuracy	Macro-F1
Transformer only	0.6121	0.5894
LightGBM (tabular only)	0.5909	0.5672
Feature-Transformer + LightGBM (full)	0.8242	0.7735

The results from every experiment show that using deep temporal representations along with statistical tabular features together produce a much greater benefit than combining them individually or using them alone. Feature Transformer alone had moderate performance because it captured some long-range dependencies; however, it did not utilize the rich hand-crafted kinematic, geometric and temporal statistics that provided discriminative power for distinguishing between highly subtle differences in motion. In comparison, just using LightGBM on its own with tabular features had similar poor results (though it produced no additional performance beyond the baseline because it does not utilize deep embeddings to model sequential dynamics). As a result, the entire hybrid model exhibited a larger performance improvement (+21.21% increase in accuracy and +18.63% increase in Macro-F1) than each of its parts on their own and validated that the "trajectory fingerprints" derived from the Feature Transformer and tabular statistics provided complementary information.

Using SHAP (SHapley Additive exPlanations) with the TreeExplainer of the final LightGBM model also enabled investigator analysis of feature contribution globally and at the instance level. The SHAP summary beeswarm plot for the test set shows the top-20 features (Figure 4). Every dot in the summary beeswarm plot corresponds to one instance and represents its corresponding SHAP value; that is, instances in red will cause the prediction to be pushed towards a greater output value for the predicted class (i.e., positive contribution), whereas instances in blue will cause the opposite (i.e., negative contribution). The features were then ranked, in descending order, according to the absolute value of their mean SHAP values (the largest impact value being at the top). The embedding dimensions produced by the Transformer (i.e., prefixed with Trans_emb) within the top ranks provided overwhelmingly larger spreads in SHAP value (i.e., distributions of instances producing positive and negative impacts) than all other features and, in particular, majority classes such as cargo

 doi:10.20944/preprints202604.0978.v1

and tanker vessels. This indicates that the Feature Transformer learned deep temporal patterns of motion—specifically the consistent navigation regimes and long-range dependencies—are more discriminative than all other features. Ranking of tabular features (e.g., 'duration', 'dcog_std') are lower, have smaller and equal SHAP values, and serve as support features rather than primary features.

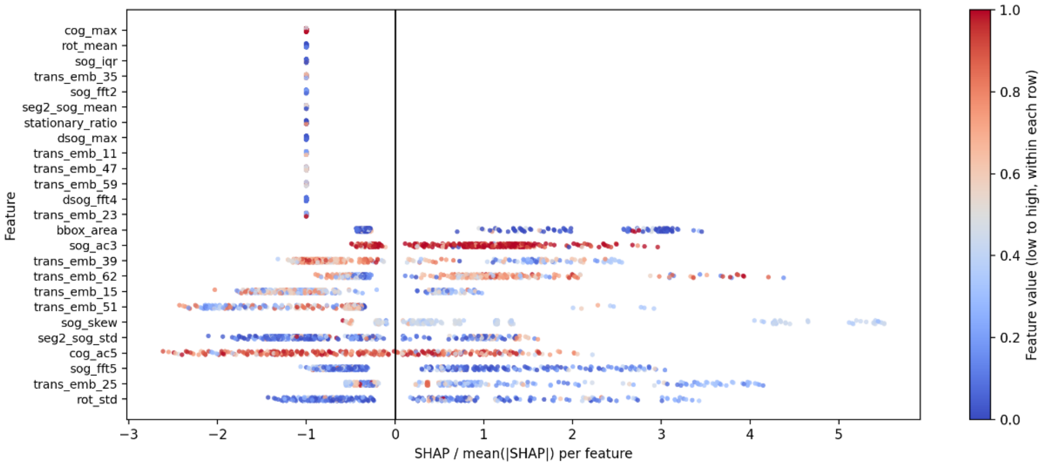


Figure 4. SHAP summary beeswarm plot (top-20 features) for the test set. Features are ranked by mean absolute SHAP value. Transformer embeddings (Trans_emb_*) show the highest impact and predominantly positive contributions (red dots), underscoring their dominant role in capturing discriminative temporal patterns. Red/blue colors indicate positive/negative effects on the model output.

5.4. Performance Analysis

The normalized confusion matrix presented in Figure 5 indicates that dominant classes, such as the Bulk Carrier, Cargo Ship, Tanker and Fishing Ship, were classified accurately, because they all are located on the diagonal of the confusion matrix. There are higher numbers of misclassifications occurring among the minority class (LNG Carrier, Pleasure Craft) and those that are visually similar in appearance (i.e., Cargo Ship ↔ Container Ship or Fishing Ship ↔ Reefer Ship) and this is related to an unbalanced distribution of data and overlapping behaviors. However, the hybrid classifier provided reasonable levels of recall for all classes, even those that were not well represented, because the Feature Transformer produced reliable temporal fingerprints for each class.

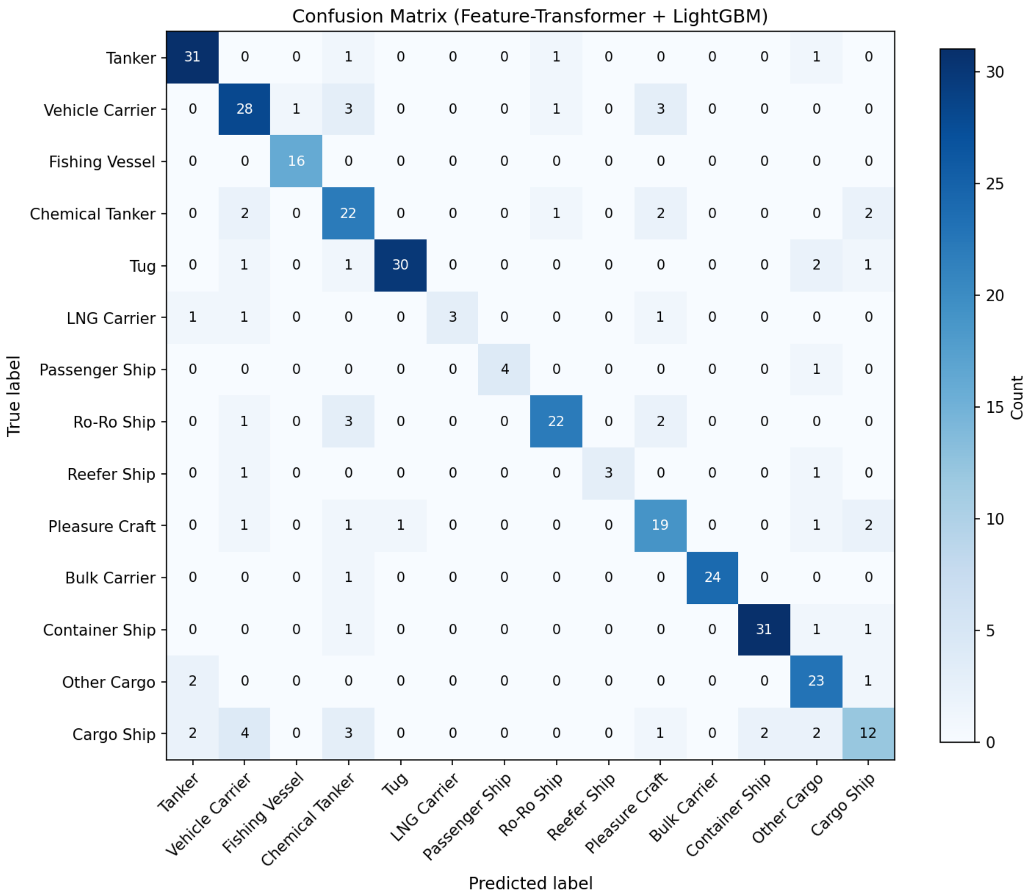


Figure 5. Confusion matrix on the test set for the proposed hybrid model (normalized by row). High diagonal values indicate strong overall performance, with notable confusion between cargo/fishing ships and certain minority classes.

t-SNE visualizations of the 2D feature spaces provide additional insight into the different types of modal behavior exhibited by the different motion classes (Figure 6). Tabular statistical features alone (a) produce diffuse class clusters with significant overlap. A pure transformer fingerprint (b) improves class cluster separation, but exhibits some mixing between similar motion regimes, which is also true for the combined hybrid representation (c). Nevertheless, the combined representation produces the most distinct class clusters, with the greatest inter-class separation and the least intra-class variance. This indicates that the addition of the temporal embeddings produced by the deep models combined with the hand-crafted statistical features create a more discriminative feature space which is directly contributing to the Macro-F1 improvement observed.

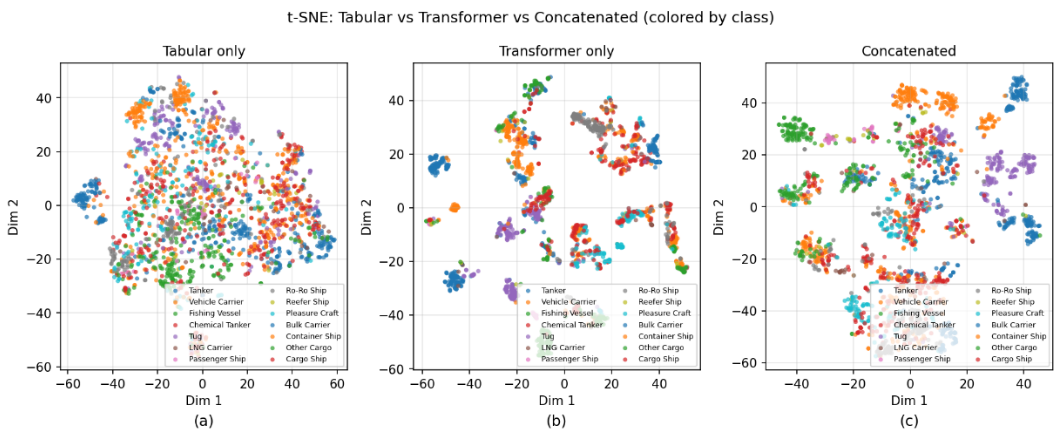


Figure 6. t-SNE visualization of trajectory embeddings in 2D space. (a) Tabular statistical features only; (b) Transformer fingerprints only; (c) Concatenated hybrid features. Clearer class separation is observed in the hybrid representation (c).

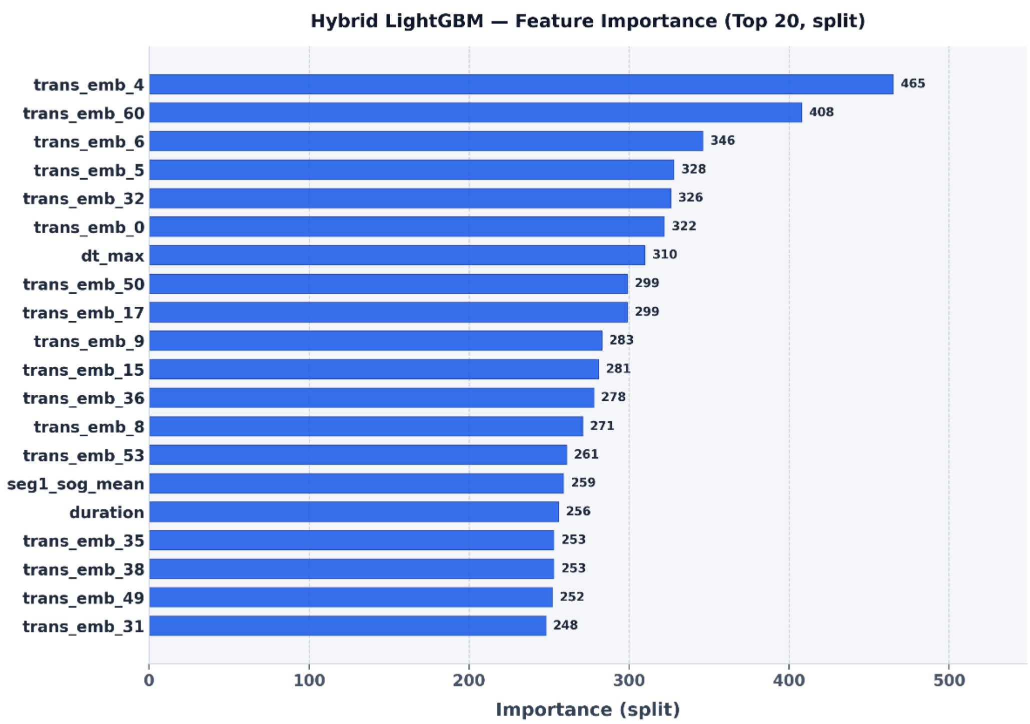
5.5. Inference Efficiency

Real-time maritime applications require fast inference speed. The experimental set showed the combination of the Feature transformer with lightGBM produced an average inference time of 1.58ms/batch (batch size = 128) which is 10 to 50 times faster than pure deep sequence models (LSTM = 29.43ms; GRU = 51.51ms; CNN +LSTM = 79.75ms) and has improved accuracy and Macro-F1 scores. The high efficiency of the new model stems from several sources: (1) Small (64-dimensional) transformer "fingerprint" (low dimensional bottleneck); (2) Optimized gradient-boosting implementation in lightGBM (CPU friendly, histogram-based splits); (3) No heavy processing of full trajectories at inference time. This combination of efficiencies makes the overall framework very suitable for real time maritime surveillance, collision avoidance and cognitive electronic warfare systems.

5.6. Discussion

The Feature Transformer + LightGBM hybrid exhibits its top performance because of the complementary capabilities of its elements; the Transformer learns how to automatically generate rich, long range temporal "fingerprints" from raw AIS sequences, while LightGBM provides an efficient means for integrating the very deep representations of these sequences with strong statistical, kinematic, and geometric tabular features. This combination of techniques solves the major challenges seen in previous models – i.e., pure deep models have high computational expense and little interpretability, while stand-alone tree based models struggle with sequential dependencies and complex multi-modal patterns.

The feature importance analysis (gain type, shown in Figure 7) of the global features of the Transformer/LGTM hybrid again demonstrates complementary capabilities; the top 20 ranked dimensions are predominately derived from the Transformer (Trans_emb_*), with only a few tabular features (such as "duration") being present in the top 10. Thus, learned temporal patterns exhibited substantially greater discriminatory value, than pure statistical values alone.



 doi:10.20944/preprints202604.0978.v1

Figure 7. Top-20 feature importances (gain type) from the final LightGBM model. Transformer embeddings (blue bars) overwhelmingly rank highest, with 'duration' as the only tabular feature in the top-10, illustrating the pivotal role of deep representations.

SHAP analy-sis on an instance level supports this observation that Trans_emb_* features across test samples have the greatest absolute SHAP values and drive predictions for the majority and the minority of classes. Table-based features (e.g., 'duration' and 'dcog_std'), as well geometric descriptors, are consistent auxiliary contributors to the split, specifically in differentiating long-haul, straight route-type transportation (i.e., cargo/tanker class) from irregular, localized route types (e.g., fishing vessels). For minority classes, the Transformer embeds provided the greatest value of compensating for the lack of information from moving data without providing enough signal from tables.

These results on a large-scale, real-world AIS dataset with natural long-tail distribution demonstrate that the proposed hybrid framework offers an excellent balance of high accuracy (82.42%), strong Macro-F1 (77.35%), excellent interpretability via SHAP and built-in importance, and ultra-fast inference—making it a practical and effective solution for ship trajectory recognition in radar signal processing, intelligent maritime surveillance, and coastal zone monitoring.

The main limitations include reliance on AIS-only input (future multi-modal fusion with radar/optical data could further improve robustness) and moderate performance on extremely rare classes (advanced imbalance handling via augmentation or meta-learning may help). The model's behavior under live-stream conditions (dense traffic, signal interference) also warrants additional real-time validation.

6. Conclusions

In this work, we proposed a hybrid framework that integrates a Feature Transformer for deep temporal representation learning with a LightGBM ensemble classifier, specifically tailored for multi-class ship trajectory recognition on large-scale real AIS data. By extracting a compact 64-dimensional trajectory fingerprint and concatenating it with carefully designed statistical and kinematic tabular features, the model jointly captures long-range temporal dependencies and domain-specific motion patterns of different ship types.

Experiments conducted on 2,196 real-world trajectories collected from 2019 to 2023 demonstrate that the proposed method achieves 82.42% overall accuracy and 77.35% Macro-F1, consistently outperforming LSTM (64.85%), GRU (64.85%), a pure Transformer (61.21%), and standalone LightGBM (59.09%) by 12.73–18.69 percentage points in accuracy. In addition, the framework maintains ultra-fast inference (1.58 ms per batch) on modern GPU hardware, indicating strong potential for near real-time maritime traffic monitoring and decision support. Ablation studies and SHAP-based interpretability analysis further confirm the complementary roles of deep trajectory fingerprints and tabular features, and provide instance-level explanations that are meaningful for maritime practitioners.

The main contributions of this study are threefold. First, to the best of our knowledge, this is one of the first studies to combine a Feature Transformer with LightGBM for multi-class ship trajectory classification using authentic, large-scale AIS data. Second, we validate the proposed framework on a long-tail distributed dataset that reflects real-world class imbalance and noisy measurements, thereby demonstrating its robustness and practical applicability. Third, we show that the hybrid architecture achieves a favorable balance between accuracy, interpretability, and computational efficiency, which is crucial for port traffic management, intelligent maritime surveillance, and emerging autonomous navigation systems.

Future research will extend this framework in three directions. One direction is multi-modal fusion, where radar returns, optical imagery, and environmental data will be integrated with AIS trajectories to further enhance recognition robustness under complex sea conditions. A second direction is the development of online and incremental learning mechanisms so that the model can

 doi:10.20944/preprints202604.0978.v1

continuously adapt to evolving ship behaviors and newly emerging vessel types. A third direction is the deployment and optimization of the proposed framework on resource-constrained edge devices (e.g., USRP-based semi-physical platforms), enabling real-time ship trajectory recognition for cognitive electronic warfare and other time-critical civilian and defense applications. We believe that the proposed hybrid paradigm offers a promising foundation for efficient, interpretable, and high-performance maritime behavior understanding in future intelligent ocean systems.

Author Contributions: Conceptualization, Songtao Hu and Liang Chen; methodology, Songtao Hu; software, Songtao Hu; validation, Songtao Hu, Qianyue Zhang; formal analysis, Songtao Hu; investigation, Songtao Hu; data curation, Songtao Hu; writing—original draft preparation, Songtao Hu; writing—review and editing, Liang Chen, Qianyue Zhang; visualization, Songtao Hu; supervision, Wenchao Liu; project administration, Wenchao Liu. All authors have read and agreed to the published version of the manuscript.

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

Funding: This work was supported by the Fundamental and Interdisciplinary Disciplines Breakthrough Plan of the Ministry of Education of China (Grant No. JYB2025XDXM115).

Institutional Review Board Statement: Not applicable. This study used publicly available Automatic Identification System (AIS) data curated and publicly released as part of a national maritime data research initiative in China, which does not involve human subjects or animals.

Informed Consent Statement: Not applicable. This study does not involve human subjects.

Data Availability Statement: The real AIS dataset used in this study is available from the corresponding author upon reasonable request, subject to data sharing policies governing the original research initiative. The source code and preprocessed data supporting the reported results will be made available on GitHub upon publication.

Acknowledgments: The authors thank the data providers for making the real AIS dataset available for research purposes. During the preparation of this manuscript, the authors used ChatGPT-4 (OpenAI) for language polishing and grammar checking. The authors have carefully reviewed and edited all generated text and take full responsibility for the content of this publication.

Conflicts of Interest: The authors declare no conflicts of interest. The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results.

References

1. Sheng K, Liu Z, Zhou D, He A, Feng C. Research on ship classification based on trajectory features. J Navig. 2018;71(1):100–116. https://doi.org/10.1017/S0373463317000529
2. Rong H, Teixeira AP, Soares CG. Maritime traffic probabilistic prediction based on ship motion pattern extraction. Reliab Eng Syst Saf. 2022;217:108061. https://doi.org/10.1016/j.ress.2021.108061
3. Zhang C, Liu S, Guo M, Liu Y. A novel ship trajectory clustering analysis and anomaly detection method based on AIS data. Ocean Eng. 2023;288:116082. https://doi.org/10.1016/j.oceaneng.2023.116082
4. Chondrodima E, Peleklis N, Pikrakis A, Theodoridis Y. An efficient LSTM neural network-based framework for ship location forecasting. IEEE Trans Intell Transp Syst. 2023;24(5):4872–4888. https://doi.org/10.1109/TITS.2022.3224567
5. Xue H, Wang S, Xia M, Guo S. G-trans: a hierarchical approach to ship trajectory prediction with GRU-based transformer. Ocean Eng. 2024;300:117431. https://doi.org/10.1016/j.oceaneng.2024.117431
6. Liu J, Zhu K, Shu Y, Zhang Y, Zheng P. Trajectory prediction model of a ship crossing under a bridge based on multi-head attention mechanism and Wasserstein GAN with gradient penalty. Ocean Eng. 2025;334:121551. https://doi.org/10.1016/j.oceaneng.2025.121551
7. You J, Chen Y, Jiang Z, Liu Z, Huang Z, Ding Y, Ran B. Exploring driving behavior for autonomous vehicles based on Gramian Angular Field Vision Transformer. IEEE Trans Intell Transp Syst. 2024;25(11):17493–17504. https://doi.org/10.1109/TITS.2024.3445710

 doi:10.20944/preprints202604.0978.v1

8. Zhang Y, Chen Y, Li Y, et al. TPTrans: Ship Trajectory Prediction Model Based on Transformer Using AIS Data. ISPRS Int J Geo-Inf. 2024;13(1):400. https://doi.org/10.3390/ijgi1311400
9. Zhou H, Zhang S, Peng J, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting. Proc AAAI Conf Artif Intell. 2021;35(12):11106–11115.
10. Wu H, Xu J, Wang J, Long M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. Adv Neural Inf Process Syst. 2021;34:22419–22430.
11. Zhang Y, Liu Y, Zheng Z, et al. AIS data-driven maritime monitoring based on Transformer: A comprehensive review. arXiv preprint arXiv:2505.07374. 2025.
12. Zhang Z, Huang L, Peng X, Wen Y, Song L. Loitering behavior detection and classification of ship movements based on trajectory shape and convolutional neural networks. Ocean Eng. 2022;258:111852. https://doi.org/10.1016/j.oceaneng.2022.111852
13. Hanafi N, Saadatfar H. A fast DBSCAN algorithm for big data based on efficient density calculation. Expert Syst Appl. 2022;203:117501. https://doi.org/10.1016/j.eswa.2022.117501
14. Ke G, Meng Q, Finley T, Wang T, Chen W, Ma W, et al. LightGBM: A highly efficient gradient boosting decision tree. Adv Neural Inf Process Syst. 2017;30:3146–3154.
15. Ferreira MD, Spadon G, Soares A, Matwin S. A semi-supervised methodology for fishing activity detection using the geometry behind the trajectory of multiple ships. Sensors. 2022;22(16):6063. https://doi.org/10.3390/s22166063
16. Ferreira L, Bifet A, Gama J. CNN and XGBoost for time series classification: A comparative study. In: 2020 International Joint Conference on Neural Networks (IJCNN). IEEE; 2020. p. 1–8.
17. Fu H, Gu Z, Wang H, Wang Y. Ship motion prediction based on ConvLSTM and XGBoost variable weight combination model. In: OCEANS 2022-Chennai. IEEE; 2022. p. 1–8.
18. Liu Y, Wu H, Wang J, Long M. Non-stationary transformers: Exploring the stationarity in time series forecasting. Advances in Neural Information Processing Systems. 2022;35:9881-9893.
19. Pallotta G, Vespe M, Bryan K. Ship pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction. Entropy. 2013;15(6):2218–2245. https://doi.org/10.3390/e15062218
20. Zhang L, Wang H, Chen Y, et al. An interpretable ship risk model based on machine learning and SHAP interpretation technique. Ocean Eng. 2025;319:119392. https://doi.org/10.1016/j.oceaneng.2025.119392
21. Liu Y, Chen H, Zhang X, et al. A method for ship trajectory repair based on feature correlation and SHAP model interpretability. In: Proceedings of the 3rd International Conference on Spatial Data and Intelligence; 2025; Singapore. Lecture Notes in Computer Science. Springer; 2025. p. 102–115. https://doi.org/10.1007/978-981-95-3102-8_10
22. Wang Y, Liu X, Chen Z. VISTA: Knowledge-driven interpretable ship trajectory imputation via large language models. arXiv preprint arXiv:2601.06940. 2026.
23. Chen X, Zhang Y, Liu M. Physics-informed neural networks for ship trajectory reconstruction under uncertain conditions. Ocean Eng. 2024;285:115890. https://doi.org/10.1016/j.oceaneng.2024.115890
24. Graves A. Long short-term memory. In: Supervised Sequence Labelling with Recurrent Neural Networks. Springer; 2012. p. 37–45.
25. Zhou T, Ma Z, Wen Q, et al. FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. Proc Mach Learn Res. 2022;162:27268–27286.
26. Zhang Y, Chen X, Liu M. Anomaly detection in maritime trajectories using Transformer-based variational autoencoders. IEEE Trans Intell Transp Syst. 2024;25(8):10234–10245.
27. Breiman L. Random forests. Mach Learn. 2001;45(1):5–32. https://doi.org/10.1023/A:1010933404324
28. Sánchez Pedroche D, Amigo D, Campaña JM, Illán Á, Mateos J. Architecture for trajectory-based fishing ship classification with AIS data. Sensors. 2020;20(13):3782. https://doi.org/10.3390/s20133782
29. Chen T, Guestrin C. XGBoost: A scalable tree boosting system. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM; 2016. p. 785–794.
30. Lundberg SM, Lee SI. A unified approach to interpreting model predictions. Adv Neural Inf Process Syst. 2017;30:4765–4774.
31. Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. Adv Neural Inf Process Syst. 2017;30:5998–6008.

 doi:10.20944/preprints202604.0978.v1

32. Kalman RE. A new approach to linear filtering and prediction problems. J Basic Eng. 1960;82(1):35–45.