



An Ensemble Decision Trees Model to Predict Traffic Pattern for Maritime Traffic Management

Downloaded from: <https://research.chalmers.se>, 2025-06-11 08:34 UTC


Citation for the original published paper (version of record):

Liu, Z., Zuo, W., Shi, H. et al (2025). An Ensemble Decision Trees Model to Predict Traffic Pattern for Maritime Traffic Management. IET Intelligent Transport Systems, 19(1).
<http://dx.doi.org/10.1049/itr2.70049>

N.B. When citing this work, cite the original published paper.

ORIGINAL RESEARCH OPEN ACCESS

An Ensemble Decision Trees Model to Predict Traffic Pattern for Maritime Traffic Management

Zhao Liu^{1,2} | Weipeng Zuo^{1,2} | Hua Shi^{1,2} | Wanli Chen^{1,2} | Xiao Lang³ | Wengang Mao³ | Mingyang Zhang⁴ 
¹School of Navigation, Wuhan University of Technology, Wuhan, China | ²Hubei Key Laboratory of Inland Shipping Technology, School of Navigation, Wuhan University of Technology, Wuhan, China | ³Department of Mechanics and Maritime Sciences, Chalmers University of Technology, Göteborg, Sweden | ⁴School of Engineering, Department of Mechanical Engineering, Aalto University, Espoo, Finland

Correspondence: Mingyang Zhang (mingyang.0.zhang@aalto.fi)

Received: 20 February 2025 | **Revised:** 30 April 2025 | **Accepted:** 16 May 2025

Funding: This work is funded by the National Natural Science Foundation of China (No. 52171351).

ABSTRACT

This study presents a traffic pattern prediction model using ensembles of decision trees, leveraging AIS data to classify maritime traffic patterns. The model integrates static information, such as origin and destination, with dynamic data, including ship speed, course and spatial position, to define and extract relevant traffic features. By combining traditional algorithms with a decision tree ensemble model, a stacked predictive framework is constructed and trained on these extracted traffic characteristics. The model is applied and validated using data from the Fujiangsha waters of the Jiangsu section of the Yangtze River. Comparative analysis reveals that this model consistently outperforms traditional algorithms and ensemble models, maintaining stable accuracy above 98% across diverse scenarios. Testing on unseen ship data further confirms the model's predictive reliability, aligning well with actual navigation patterns. The findings suggest that this model has strong potential to (1) forecast navigation routes for improved traffic management, (2) infer ship behaviour based on predicted traffic patterns and (3) support future applications in intelligent ship navigation.

1 | Introduction

1.1 | Background

Maritime transportation is one of the most dynamic yet high-risk sectors in global marine logistics [1]. With the substantial rise in global economic and trade activities, maritime traffic has become increasingly dense and complex, creating heightened demands for safety and operational efficiency in maritime environments [2]. Traditional, passive maritime traffic management methods face significant limitations, including response delays, limited foresight, inefficiencies and challenges in addressing urgent issues. In contrast, proactive traffic management can markedly improve supervision efficiency, optimise resource allocation and reduce accident risks [3, 4]. Predicting and inferring ship

behaviours is fundamental to enabling active management. This requires advanced analytical capabilities in three key domains: vessel movement patterns, inland waterway traffic characteristics and scientific behavioural reasoning. Such systematic analysis enables critical supervisory functions, including early warning, judgement, analysis and decision-making capabilities.

The increasing integration of maritime big data, particularly Automatic Identification System (AIS) data, has proven invaluable in recent years [5]. AIS data, containing extensive details of ship movements and traffic information, offers a robust resource for characterising maritime traffic patterns, predicting ship behaviour, estimating collision probabilities and detecting anomalies in ship trajectories [6–11]. However, due to the vastness of AIS data and the low density of actionable insights within it,

This is an open access article under the terms of the [Creative Commons Attribution-NonCommercial-NoDerivs](https://creativecommons.org/licenses/by-nc-nd/4.0/) License, which permits use and distribution in any medium, provided the original work is properly cited, the use is non-commercial and no modifications or adaptations are made.

© 2025 The Author(s). *IET Intelligent Transport Systems* published by John Wiley & Sons Ltd on behalf of The Institution of Engineering and Technology.

the inherent complexity of maritime traffic situations and the critical need to elevate safety standards, it has become essential to leverage data mining techniques [8, 9]. These methods can analyse the large-scale AIS datasets to uncover ship behaviour patterns and operational rules, which can then be applied to reduce navigation risks, enhance supervision efficiency and ensure waterway safety. Such analytical capabilities are vital for advancing modern inland waterway systems and supporting the intelligent development of maritime transportation [12].

Inland waterway traffic management inherently benefits from known origin and destination information, yet many existing methods primarily rely on dynamic characteristics such as speed and heading. This neglects potentially valuable static features (e.g., origin and destination points), often resulting in lower classification accuracy and higher computational requirements. Addressing these challenges, this study targets intelligent inland waterway traffic management, focusing on AIS data to conduct an in-depth analysis of ship navigation histories. By extracting key traffic characteristics, this study constructs a predictive traffic pattern model based on ensembles of decision trees, providing a foundational tool to support proactive and effective maritime safety supervision.

1.2 | Literature Review

Maritime traffic pattern classification involves extracting representative and recurring traffic behaviours and navigation paths from extensive ship trajectory data, primarily sourced from AIS and similar systems [5]. By clustering trajectories with similar movement patterns, these methods enable analysis of traffic flow characteristics and ship behaviours, contributing to maritime traffic safety, route planning, behaviour prediction and other applications [3, 13, 14].

The primary approaches for accurately analysing segmented maritime traffic patterns include grid-based, vector-based and statistical methods. Grid-based methods divide a maritime area into spatially indexed grids to represent traffic attributes, thus reducing the problem scale and enhancing knowledge storage efficiency [14, 15]. Vector-based methods conceptualise traffic routes as waypoint-connected paths, compactly represented as graph-based networks that facilitate visualisation of ship movements and patterns [16–18]. Statistical methods focus on quantifying traffic characteristics to determine distribution patterns and thresholds, which can aid in distinguishing between normal and abnormal navigation behaviours [19, 20].

Given the predictive potential of traffic pattern and ship behaviour analysis, these areas have become central research focuses in maritime traffic management. Ship behaviour prediction methods are generally categorised as dynamic-based, machine learning-based, or neural network-based.

Dynamic-based prediction approaches rely on dynamic equations derived from ship manoeuvrability parameters and dead reckoning, using algorithms like Kalman filters, Markov chains, grey prediction and vector analysis [21–23]. However, these models often depend on idealised motion assumptions, which may limit their applicability in real-world conditions.

Machine learning-based prediction models are trained on historical data, allowing for future state predictions based on current conditions. Common algorithms include decision tree regression [24], support vector machines [25], random forests (RFs) [26, 27] and Gaussian process regression [28]. Although machine learning-based models provide high accuracy and can autonomously learn features from data, they require careful algorithm and parameter selection for optimal results.

Neural network-based prediction methods, due to their distributed parallel processing capabilities, excel at handling complex, variable trajectory data with high predictive accuracy. For instance, Gao et al. [29] employed long short-term memory (LSTM) neural networks to predict ship behaviour by converting AIS data into sequential inputs for recurrent neural networks (RNN). Later, Ma et al. [30] introduced the accumulated long short-term memory (ALSTM) model, which uses skip connections and adaptive memory modules to predict navigation intentions in intersecting waterways, thereby addressing limitations of traditional LSTM. Additionally, convolutional neural networks (CNN) have been effectively applied in classification and image recognition, providing insights into ship behaviour and collision risk [26, 27]. Liang et al. [31] developed a multi-view feature fusion framework that combines motion and morphological features through convolutional auto-encoder (CAE) and bidirectional gated recurrent unit (BiGRU) networks for ship classification.

With advancements in artificial intelligence, the maritime industry is increasingly adopting intelligent and automated solutions for trajectory prediction [32]. Machine learning and deep learning-based prediction methods have gained popularity. While deep learning achieves high predictive accuracy, challenges such as model complexity, data requirements, interpretability and parameter tuning remain prevalent in ship behaviour prediction, underscoring the need for further methodological innovation.

1.3 | Contributions

With the increasing complexity of inland waterway traffic due to global trade expansion and growing vessel activities, effective traffic management has become a critical challenge. Traditional approaches, relying on passive monitoring and reactive measures, struggle to meet modern demands for safety, efficiency and sustainability. The integration of advanced data-driven techniques into maritime traffic management presents a promising solution to these challenges.

This study presents a novel traffic pattern prediction model based on ensemble decision trees, integrating both static (origin and destination) and dynamic (speed, heading and spatial position) features to enhance classification accuracy and efficiency in inland waterway traffic analysis. The proposed model achieves over 98% accuracy across diverse scenarios, ensuring robust performance in real-time maritime applications, including route prediction and proactive traffic supervision. By enabling precise traffic classification and ship behaviour prediction, this research advances intelligent inland waterway management, supporting autonomous navigation and proactive traffic control, ultimately improving safety, efficiency, and sustainability in maritime operations.

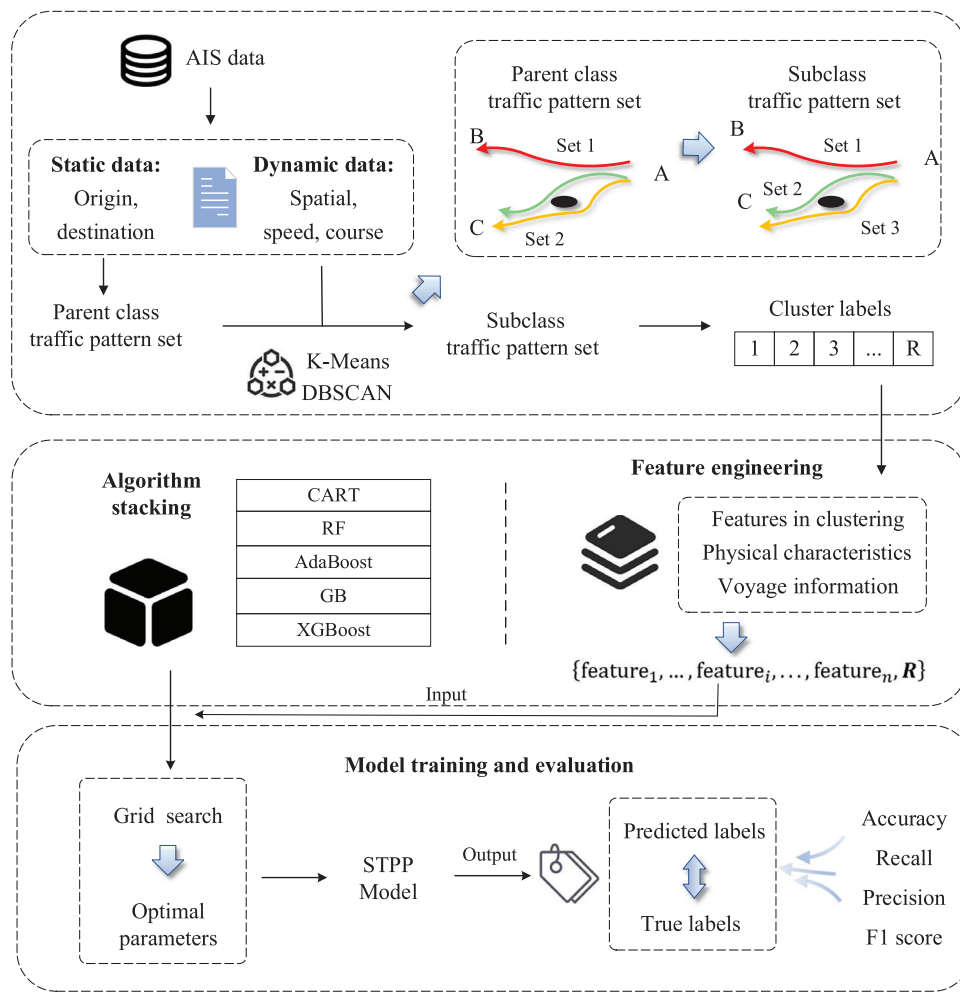


FIGURE 1 | Flowchart of traffic pattern prediction method using AIS data.

The rest of the study is organised as follows. Section 2 presents a two-stage trajectory clustering method for traffic pattern classification and a prediction model based on ensembles of decision trees. Section 3 demonstrates the advantages of this model through a case study. Section 4 concludes the study.

2 | Methodology

This study presents a traffic pattern prediction model based on ensemble decision trees, utilising AIS data to classify maritime traffic patterns, as illustrated in Figure 1. The study consists of the following steps:

Step i: data pre-processing and classification. Historical AIS data are pre-processed to extract relevant trajectory information, enabling the classification of traffic patterns. Traffic patterns are then encoded into categorical labels, facilitating further analysis (details of AIS data processing are provided in Appendix A).

Step ii: feature set construction and model development. By mining static and dynamic information from ship trajectories, a comprehensive feature set is created. Meanwhile, a stacking model based on decision tree algorithms is

constructed, tailored to effectively capture traffic pattern characteristics.

Step iii: model training and evaluation. The feature set is input as independent variables into the stacking model, which is then trained to determine optimal parameters. Model performance is evaluated based on the probability of accurately predicting the correct traffic pattern labels.

2.1 | Traffic Pattern Classification Based on Two-Stage Trajectory Clustering Method

Traffic pattern classification is a methodological approach that employs data mining and related analytical techniques to categorise ship trajectories into distinct clusters based on AIS data. This process aims to delineate and characterise typical maritime traffic patterns. Significant variations in ship traffic characteristics are observed between different trajectory clusters, whereas minimal differences are noted within the same cluster [6, 7].

The concept of parent class traffic pattern encompasses the aggregation of all ship trajectories that share identical departure and arrival areas. Subclass traffic pattern refers to the collection of ship trajectories that, while sharing the same departure

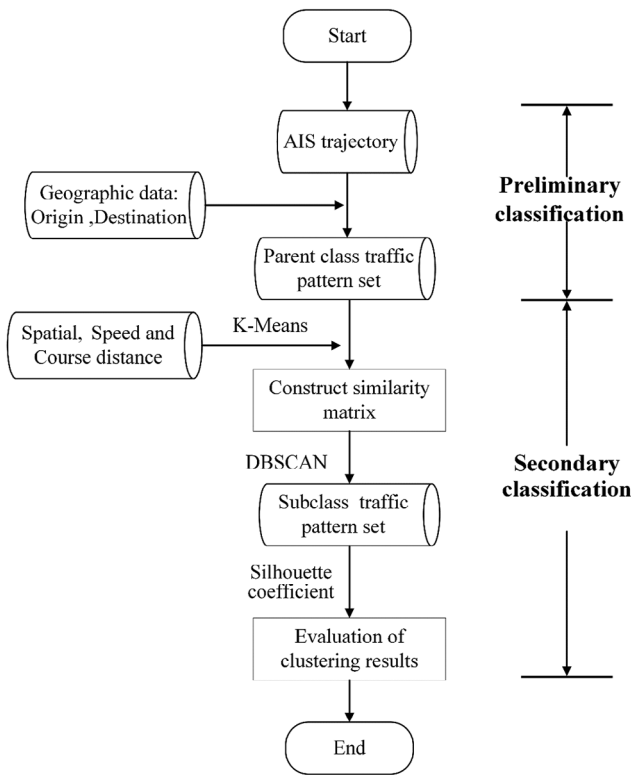


FIGURE 2 | Traffic pattern classification for two-stage trajectory clustering.

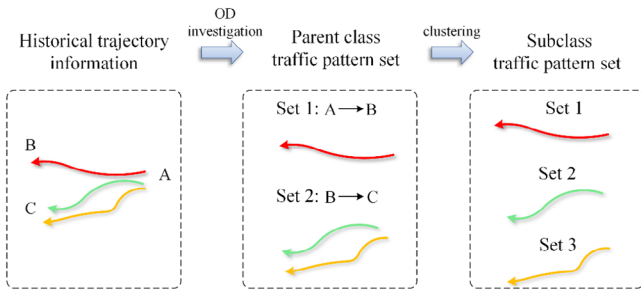


FIGURE 3 | The process of traffic pattern classification.

and arrival areas, exhibit divergent navigation routes. Notably, subclass traffic patterns are subsets of parent class traffic patterns. The classification process begins with an investigation of origin and destination (OD) points, followed by clustering, as illustrated in Figures 2 and 3.

2.1.1 | Preliminary Classification Based on Origin and Destination Investigation

In inland waterways, a traffic pattern is defined by the clustering of ship trajectories that originate from the same departure area and conclude at the same arrival area via a consistent route. Consequently, to effectively extract traffic patterns within inland waterways, it is essential to identify the potential departure and arrival areas within the study region. These areas are composed of two components: (1) the intersection of the waterway boundary and the channel, referred to as the entrance and exit; and (2)

TABLE 1 | The pseudocode for area judgement algorithm based on origin and destination investigation.

Algorithm I: Area Judgement

Input: Coordinates of area, ship trajectory

Output: Trajectory cross results

P_1 : The origin of trajectory; P_n : The destination of trajectory

1: **Begin**

2: **If** $\min(\text{area A.longitude}) < P_1.\text{longitude} < \max(\text{area A.longitude})$

3: and $\min(\text{area A.latitude}) < P_1.\text{latitude} < \max(\text{area A.latitude})$, **then**

4: P_1 in area A == True

5: **End if**

6: **If** $\min(\text{area B.longitude}) < P_n.\text{longitude} < \max(\text{area B.longitude})$

7: and $\min(\text{area B.latitude}) < P_n.\text{latitude} < \max(\text{area B.latitude})$, **then**

8: P_n in area B == True

9: **End if**

10: **If** P_1 in area A == True, P_n in area B == True, **then**

11: trajectory cross area A, area B

12: **END**

the ports, docks, or anchorages within the waterways [8]. The entrances and exits function as the primary OD nodes of the maritime transportation network in the area, whereas the ports, docks, or anchorages serve as the secondary OD nodes.

The preliminary classification of traffic patterns is conducted as follows: First, geographical information is utilised to delineate the spatial boundaries of the entrances and exits within the study water area. Subsequently, the entrances and exits are matched with the origin and destination points of the ship trajectories to determine whether these points fall within the designated entrance and exit areas. If a ship trajectory is found to pass through a specific group of entrances and exits, it is assigned to the parent traffic pattern associated with that group. The pseudocode of the traffic pattern classification method based on origin and destination investigation is shown in Table 1 [8].

2.1.2 | Secondary Classification Based on Cluster Analysis

Within the same parent class, traffic patterns defined by identical departure and arrival areas, there may exist significant variations in trajectory subclasses, necessitating further classification of these parent class patterns. The static characteristics of the ship, such as origin and destination, have already been analysed. Therefore, further exploration is required based on the dynamic characteristics of the ship. This study mainly considers three aspects: the spatial distance of the trajectory Sp_s , the speed distance of the trajectory Sp_v and the course distance of the trajectory Sp_c . The spatial distance, speed distance, and course

distance are detailed in Appendix B. For each pair of trajectories, a similarity matrix is computed based on Sp_s , Sp_v and Sp_c , respectively. These similarity measurement parameters are then normalised and combined to form a comprehensive similarity measurement matrix SM . This matrix is subsequently input into a clustering algorithm to complete the trajectory clustering process.

During the trajectory clustering phase, the K-Means algorithm is utilised to calculate the weight allocation of the similarity matrix. Then the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) algorithm applied for trajectory clustering analysis. The specific descriptions of K-Means algorithm and the DBSCAN algorithm are as follows.

1. K-Means algorithm

K-Means algorithm is a typical partition-based clustering method [33]. K-Means minimises the sum of intra-cluster distances by iteratively optimising cluster centres and assigning cluster members [6, 7]. The objective function for K-Means is the sum of square error (SSE) within the cluster, as shown in Equation (1),

$$SSE = \sum_{i=1}^k \sum_{x \in C_i} \text{dist}^2(m_i, x), \quad (1)$$

where k is the number of clusters, C_i is the i -th cluster, x is the data and m_i is the center of cluster C_i .

K-Means is simple and efficient, and can be used to calculate the weight allocation of similarity matrix and identify the optimal weight combination that maximises the clustering effectiveness.

2. DBSCAN algorithm

DBSCAN is one of the best-known density-based clustering algorithms [34]. In the process of trajectory clustering, especially in the analysis of ship dynamic characteristics, DBSCAN algorithm can well deal with the noise data occasionally appearing in the speed or course. By effectively identifying and isolating such outliers, DBSCAN ensures that the clustering results remain robust and unaffected by anomalous data points [6, 7]. Therefore, for the traffic pattern parent set which has already undergone preliminary classification, DBSCAN can be applied to further separate trajectories based on their distinct dynamic characteristics.

The evaluation of trajectory clustering results is essential for assessing the performance of clustering algorithms on a given dataset. One widely used evaluation metric is the Silhouette Coefficient [35]. This metric quantifies the degree of cohesion and separation among clusters by calculating the average Silhouette Coefficient for all samples in the dataset, as shown in Equation (2),

$$S = \frac{\sum_{i=1}^m s_i}{m}, \quad (2)$$

where S is the mean of the Silhouette coefficients of all sample sets and the range of values is $[-1, 1]$. s_i is the Silhouette Coefficient

TABLE 2 | The pseudocode for DBSCAN algorithm.

Algorithm II: DBSCAN

Input: Dataset $D = \{x_1, x_2, x_3, \dots, x_m\}$

Output: Clustering division $C = \{c_1, c_2, c_3, \dots, c_k\}$

Process:

- 1: Mark the D as unprocessed trajectories;
- 2: **For** $i = 1, 2, 3, \dots, m$;
- 3: Check the neighborhood $\epsilon(x_i)$;
- 4: **If** the number of objects in $\epsilon(x_i) \geq \text{Minpts}$:
- 5: Mark x_i as core point and set up a new class c and add objects in $\epsilon(x_i)$ to N ;
- 6: **For** p in N :
- 7: Check the neighborhood $\epsilon(p)$;
- 8: **If** the number of objects in $\epsilon(p) \geq \text{Minpts}$;
- 9: Add objects not be classified in $\epsilon(x_i)$ to N and add p to c ;
- 10: **Else:**
- 11: Add p to c ;
- 12: **End if**
- 13: **End for**
- 14: **End if**
- 15: **If** the number of objects in $\epsilon(x_i) < \text{Minpts}$:
- 16: Mark x_i as boundary point or noise point;
- 17: **End if**
- 18: **End for**
- 19: **Output** $C = \{c_1, c_2, c_3, \dots, c_k\}$.
- 20: **End.**

of a single sample and its calculation formula is shown in Equation (3),

$$s_i = \frac{b - a}{\max(a, b)}, \quad (3)$$

where a is the average distance between the data point and samples of the same class, and b is the average distance between the data point and samples of the different classes with the closest distance. Generally, the closer the distance between samples of the same class, the farther the distance between samples of different classes, the higher the Silhouette Coefficient score, the better the clustering effect. The pseudocode for DBSCAN algorithm is shown in Table 2.

2.2 | Traffic Pattern Prediction Method Based on Ensembles of Decision Trees

When a ship enters controlled waters, its departure area is typically known, but the destination information may not be immediately available to the shore-based command centre. While the destination and navigation route can be inferred through data mining methods, such as clustering analysis of historical ship trajectories, these approaches often suffer from time lag.

To address this limitation, a more proactive solution involves learning the characteristic information of historical trajectories and matching it with the existing information of the current target ship. This approach can help reduce the uncertainty associated with the target ship's navigation intentions. To evaluate the uncertainty of a ship's navigation destination and route, this study proposes a traffic pattern prediction model based on ensembles of decision trees.

2.2.1 | Decision Tree Algorithm

Decision trees, a widely used supervised learning algorithm in machine learning [36–38], employ a tree structure to classify instances based on specified features. Essentially, decision trees function as a series of 'if-then' rules or a conditional probability distribution across the feature and class spaces. In a decision tree model, each internal node represents an attribute-based decision, each branch represents an outcome and each leaf node represents a classification (for a classification tree) or a regression output (for a regression tree). This algorithm is straightforward, interpretable and can handle both numerical and categorical data effectively.

Three primary decision tree algorithms include ID3, C4.5 and Classification and Regression Tree (CART). The ID3 algorithm, which uses information gain for feature selection, tends to favour features with many values, risking overfitting and lacking support for continuous features and missing values. C4.5, an improvement over ID3, uses the information gain ratio to address overfitting and can handle both continuous features and missing values. The CART algorithm uses the Gini index for feature selection, supports classification and regression and provides pre-pruning and post-pruning options to manage model complexity [39]. While decision trees are inherently simple and interpretable, they can suffer from overfitting, instability and susceptibility to local optima, leading to poor generalisation. Single decision trees can become overly complex by fitting noise in the training data, making them sensitive to minor data changes and tend to follow a greedy approach that doesn't guarantee a global optimal solution.

Ensemble learning addresses these limitations by combining multiple models to improve accuracy and stability. The core principle of ensemble learning is to aggregate many weak learners into a strong learner, with two primary methods for decision trees: bagging and boosting. Bagging: This method builds multiple independent models trained on different subsets of data sampled with replacement from the original dataset. Final predictions are made by voting (for classification) or averaging (for regression). A popular example is RF [40], which reduces overfitting by using random feature selection and sampling, making it suitable for both classification and regression tasks. Boosting: Boosting creates models sequentially, with each model attempting to correct errors from its predecessors, focusing on misclassified samples. Common examples include Gradient Boosting (GB) [41], Adaptive Boosting (AdaBoost) [42] and Extreme Gradient Boosting (XGBoost) [43].

2.2.2 | Traffic Pattern Prediction Model

In this study, the characteristic parameters of the target ship are input into the trained traffic pattern prediction model to

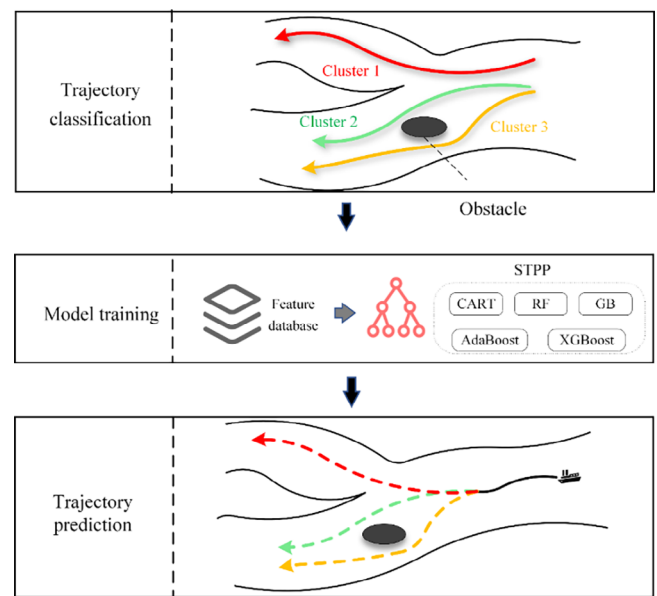


FIGURE 4 | The operation principle of the traffic pattern prediction model.

determine the traffic pattern associated with the target ship. This enables the inference of the target ship's navigation intention. The operation framework of the traffic pattern prediction model based on the ensembles of decision trees is shown in Figure 4. The specific description is as follows:

- Traffic patterns are extracted from historical trajectories and each traffic pattern is labelled manually. This process has been accomplished using the two-stage trajectory clustering method.
- Construct a feature dataset and train the decision tree classification model. This phase represents the core process of the model.
- The trained classification model is applied to new trajectories within the study water area. By analysing the characteristic parameters of the new trajectory, the model assigns it to a specific traffic pattern category. This stage represents the practical application of the model.

Since the target variables are discrete label variables of multiple historical trajectory clusters, it is necessary to construct a multi-classification decision tree. The traffic pattern prediction model based on ensembles of decision trees is divided into six steps as follows:

1. Data preparation

The data used for traffic pattern prediction consist of pre-processed AIS trajectory data. Additionally, the trajectory data must be manually labelled to assign a trajectory category label, which serves as the target variable for the machine learning algorithm.

2. Feature engineering

TABLE 3 | The equations of descriptive statistics.

Descriptive statistics	Equation and description
Maximum value	$x_{max} = \max(x_1, x_2, \dots, x_n)$
Minimum value	$x_{min} = \min(x_1, x_2, \dots, x_n)$
Mean value	$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$
Standard deviation	$\sigma = \sqrt{\frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2}$
Median value	Sort data from smallest to largest, $x_{median} = \begin{cases} x_{\frac{n+1}{2}}, n \text{ is odd} \\ \frac{x_{\frac{n}{2}} + x_{\frac{n}{2}+1}}{2}, n \text{ is even} \end{cases}$

Feature engineering is a critical step in constructing the traffic pattern prediction model. It involves selecting and transforming relevant features from the dataset to serve as decision indicators. These indicators not only act as training features for the historical trajectory dataset but also as input features for the target ship's trajectory. While AIS data provides a wealth of characteristic information, such as dynamic trajectory features, other factors such as the physical characteristics of the ship and voyage information should also be considered. The following characteristics are mainly considered:

1. Features in trajectory clustering: Since the historical trajectory clusters are derived from the results of trajectory clustering, the features used in the decision tree model need to consider the features employed during the clustering process. This ensures that the model can effectively learn the distinctions between different trajectory clusters. As previously mentioned, the primary features include speed, course and their associated statistical measures, such as the mean, interval value (maximum minus minimum), standard deviation and median. The mean and median represent the central tendency of the trajectory characteristics, while the interval value and standard deviation capture the variability and movement patterns of the trajectory over time. The mathematical formulations for these descriptive statistics are detailed in Table 3.
2. Physical characteristics of the ship: Given that navigation regulations often impose specific requirements on large vessels, the physical characteristics of the ship must also be considered in the analysis. This study primarily focuses on the length and width of the ship.
3. Voyage information: Trajectory clusters are defined by the departure area, arrival area and navigation route. Since the departure area is known information, it should be taken into account.
4. Features obtained from feature mining: After the division and extraction of trajectory clusters, further feature mining is conducted to uncover additional traffic-related information hidden in the trajectories.

The general form of the feature parameter set finally formed by feature engineering is shown in Equation (4),

$$\{\text{feature}_1, \dots, \text{feature}_i, \dots, \text{feature}_n, R\}, \quad (4)$$

where feature_i is the i -th feature constructed from the dataset; n denotes the total number of features constructed; R is the label, which corresponds to the traffic pattern labels obtained through the two-stage trajectory clustering process.

3. Data set splitting

Dataset splitting involves dividing the dataset into distinct subsets to ensure the availability of an independent sample set for performance evaluation during model training. This process is crucial for accurately estimating the model's ability to generalise to unseen data. In this study, the dataset is split into a training set and a test set at a ratio of 8:2.

In the training set, $\{\text{feature}_1, \dots, \text{feature}_i, \dots, \text{feature}_n\}$ serves as the independent variable and R acts as the dependent variable. The decision tree algorithm of traffic pattern prediction model automatically finds the relationships and rules between independent variable and dependent variable through inductive inference. In the test set, the independent features are input into the trained algorithm and the predicted R is generated as the output. This predicted R represents the traffic pattern prediction result for the target ship.

4. Algorithm selection

CART is the most effective algorithm for traffic pattern classification among ID3, C4.5 and CART algorithms of a single tree. Consequently, all base learners in the ensemble learning of decision trees are derived from the CART algorithm. However, determining a definitive advantage for a specific model among CART, RF, AdaBoost, GB and XGBoost in various case scenarios remains challenging. To address this, stacking ensemble learning is employed. Stacked ensemble learning is a technique where multiple models are first trained independently, and then a meta-learner is used to synthesise their outputs into a final prediction.

Stacking structure has two layers of algorithms connected in series: (1) Level 0: This layer may include one or more strong learners. (2) Level 1: This layer contains a single learner, typically one with strong interpretability and simpler learning capabilities.

During training, the data is first input into the Level 0 algorithms for training. After training, each algorithm in Level 0 generates its corresponding prediction results. These predictions are then combined into a new feature matrix, which serves as the input for the Level 1 algorithm. The final prediction output is produced by the Level 1 learner.

In the study, the Level 0 base learners include CART, RF, AdaBoost, GB and XGBoost, while the Level 1 meta-learner is multi-logistic regression (MLR). The stacking for traffic pattern prediction (STPP) structure for traffic pattern prediction is shown in Figure 5.

5. Model training

The model is trained using the training set, where the inputs consist of the features in the feature set and the desired outputs are the corresponding labels. During this process, grid search

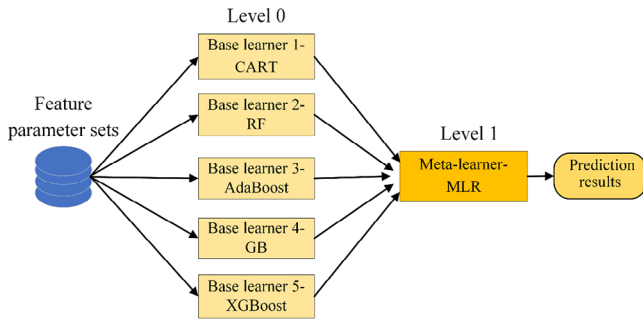


FIGURE 5 | The STPP structure composed of CART, RF, AdaBoost, GB and XGBoost.

TABLE 4 | Classification model evaluation metrics.

Accuracy	Precision	Recall	F1 score
$\frac{TP + TN}{TP + TN + FP + FN}$	$\frac{TP}{TP + FP}$	$\frac{TP}{TP + FN}$	$\frac{2}{\frac{1}{Precision} + \frac{1}{Recall}}$

TABLE 5 | The main parameters for calculating metrics.

	Predicted label is positive	Predicted label is negative
True label is positive	TP	FN
True label is negative	FP	TN

is employed to systematically explore and identify the optimal parameters for the model.

6. Model evaluation

The test set is fed into the trained model to generate predictions for the test data. These predictions are then compared with the true labels to evaluate the model's performance.

2.2.3 | Model Performance Evaluation

To demonstrate model performance on the test set, four typical classifier performance metrics are used: accuracy, precision, recall and F1 score. In general, the four metrics are calculated as shown in Table 4. The main parameters for calculating metrics are shown in Table 5.

Since traffic pattern classification is a multi-classification problem, it is essential to evaluate the classification performance by considering all categories. To achieve this, both macro- and micro-

averaging methods are employed when calculating accuracy, recall and F1 scores. Macro averaging assigns equal weight to all categories. It independently calculates the performance metrics for each category and then computes the arithmetic mean of these metrics across all categories. Micro aggregates the confusion matrices of all categories into a single combined confusion matrix. Performance metrics are then calculated based on this merged matrix. Macro-averaging is particularly useful when the dataset is imbalanced and includes small sample categories. Micro-averaging is more suitable when the dataset involves categories with large sample sizes.

Table 6 illustrates the calculation method for each metric [8]. In Table 6, macro-average metrics are arithmetic averages of metrics for each category. In the micro-average metric equation, l_i denotes the number of samples predicted by the model as class i and actually belonging to class i , m_i denotes the number of samples predicted by the model as class i and n_i denotes the number of samples actually belonging to class i .

3 | Case study

3.1 | Data Preparation and Experimental Environment

This study selects the Fujiangsha water area in the Jiangsu section of the Yangtze River as the research object and conducts method validation analysis based on the AIS data collected from this region. The Fujiangsha water area is located between latitudes 31.877°N and 32.136°N and longitudes 120.275°E and 120.600°E. This area is primarily divided into three distinct waterways: the Fujiangsha North Channel, the Fujiangsha Middle Channel and the Fujiangsha South Channel. The AIS trajectory density in Fujiangsha waters is illustrated in Figure 6.

This study utilises AIS data from China's Northern Navigation Service Center to analyse commercial vessel navigation in the Yangtze River's Fujiangsha section during the 2019 flood season (June–August). During this period, hydrological conditions were characterised by elevated water levels and intensified current velocities. These changes resulted in distinct speed variations between upstream and downstream vessels: downstream ships generally achieved higher speeds due to favourable currents, while upstream ships experienced reduced speeds as they contended against stronger flows.

3.2 | Traffic Patterns in Fujiangsha Water Area

To extract the traffic patterns of the Fujiangsha water area, it is essential to identify the departure areas and arrival areas for

TABLE 6 | Metrics calculation of macro-average and micro-average in multi-classification model.

Metric	Accuracy	Precision	Recall	F1 score
Macro-average	$\frac{TP + TN}{TP + TN + FP + FN}$	$\frac{1}{k} \sum_{i=1}^k P_i$	$\frac{1}{k} \sum_{i=1}^k R_i$	$\frac{1}{k} \sum_{i=1}^k F1_i$
Micro-average		$\frac{\sum_{i=1}^k l_i}{\sum_{i=1}^k m_i}$	$\frac{\sum_{i=1}^k l_i}{\sum_{i=1}^k n_i}$	$\frac{2}{\frac{1}{P_{micro}} + \frac{1}{R_{micro}}}$

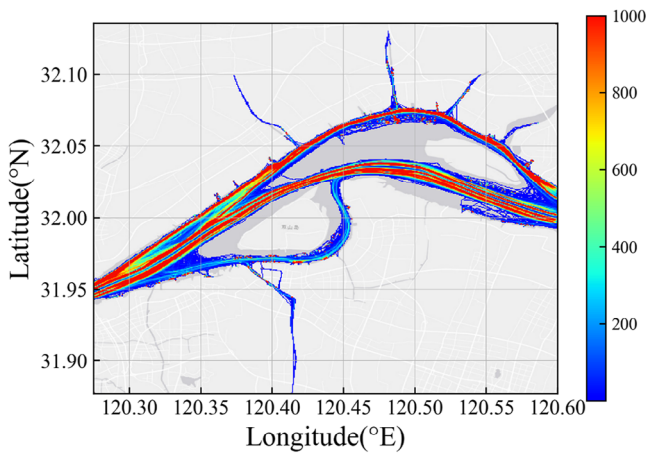


FIGURE 6 | AIS trajectory density in Fujiangsha waters.

ships navigating within this region. Based on the geographical characteristics of the water area, seven judgment areas are delineated. These areas are located at key junctions, including: the intersection of the upstream of the Yangtze River and the study area, the intersection of the downstream of the Yangtze River and the study area, the confluence points of the four tributaries and the main stream and the intersection of the natural harbour waterway and the study area. The seven judgement areas are illustrated in Figure 7. The specific latitude and longitude ranges for these areas are provided in Table 7.

Given the limited number of ship trajectories between the tributaries, this study primarily concentrates on analysing ship trajectories that traverse the upstream and downstream of the Yangtze River, as well as those connecting the Yangtze River and its tributaries. The preliminary classification results are summarised in Table 8. A visual representation of these trajectories is provided in Figure 8.

The ship trajectories between the upstream and downstream of the Yangtze River are further divided, as illustrated in Figure 9.

Using the similarity matrix parameter calculation method described in Section 2.1.2, the values Sp_s , Sp_v , Sp_c are computed. To determine the combined similarity matrix $SM = w_1 * Sp_s + w_2 * Sp_v + w_3 * Sp_c$, the grid search approach is employed in

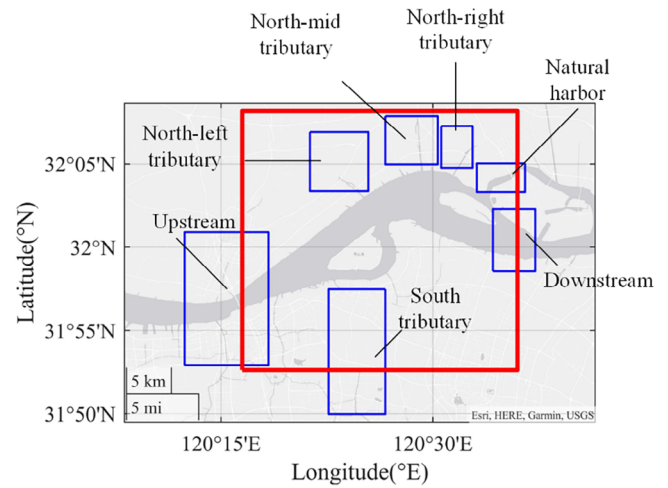


FIGURE 7 | The boundaries of the study area and the judgment areas.

conjunction with the K-Means algorithm. The weights w_1 , w_2 and w_3 are constrained to the range [0,1] with a search step of 0.1 and they must satisfy the condition $w_1 + w_2 + w_3 = 1$. The search parameters for the two types of trajectory sets are detailed in Table 9.

After SM similarity matrix is obtained by using K-Means algorithm, the matrix is input into DBSCAN algorithm for trajectory clustering. To optimise the clustering results, the grid search approach is conducted to determine the values of eps and $MinPts$ that maximise the Silhouette Coefficient of the DBSCAN clustering outcomes. The search step size of eps is 0.05, the search range is [2.6, 3.5], the search step size of $Minpts$ is 1 and the search range is [2,6]. The final values of eps and $Minpts$ for the two types of trajectory sets are presented in Table 10. The statistics of each cluster are shown in Table 11. The visualisations of the clustered trajectories are provided in Figures 10 and 11, respectively.

According to Table 11 and Figure 10, there are three typical routes for ships sailing from upstream to downstream. Among these, the Middle Waterway is the most frequently used, with 4771 trajectories. In contrast, the north waterway and south waterway

TABLE 7 | The latitude and longitude range of the study area and the seven judgment areas.

Area	Range of longitude (°E)		Range of latitude (°N)	
	Minimum	Maximum	Minimum	Maximum
The study area	120.275	120.6	31.877	32.136
Upstream	120.207	120.306	31.882	32.015
Downstream	120.571	120.621	31.976	32.038
South tributary	120.377	120.444	31.833	31.958
North-left tributary	120.355	120.424	32.056	32.115
North-mid tributary	120.444	120.506	32.083	32.131
North-right tributary	120.51	120.547	32.079	32.121
Natural harbor waterway	120.552	120.609	32.055	32.084

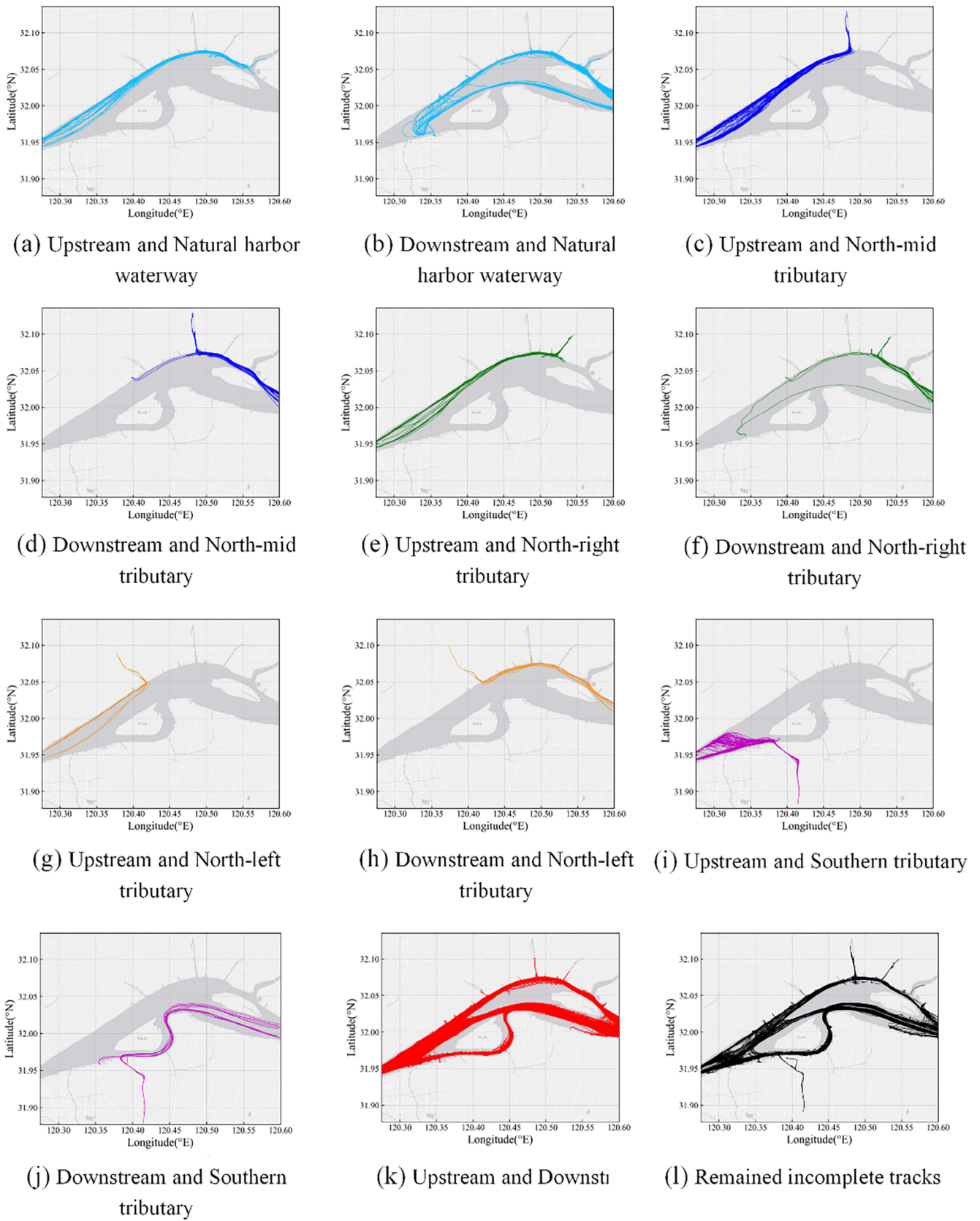


FIGURE 8 | Preliminary classification results of maritime traffic patterns.

TABLE 8 | Preliminary classification results of maritime traffic patterns in Fujiangsha water area.

Trajectories	Number	Proportion
Upstream and south tributary	122	0.99%
Upstream and north-left tributary	8	0.06%
Upstream and north-mid tributary	114	0.92%
Upstream and north-right tributary	32	0.28%
Upstream and natural harbour waterways	48	0.39%
Upstream and downstream	11463	92.89%
Downstream and south tributary	11	0.09%
Downstream and north-left tributary	11	0.09%
Downstream and north-mid tributary	79	0.64%
Downstream and north-right tributary	36	0.29%
Downstream and natural harbour waterways	122	0.99%
Remained incomplete trajectories	289	2.37%

TABLE 9 | Search parameters of similarity matrix weight.

Direction of trajectory sets	Parameter k of K-Means	w_1	w_2	w_3	Contour coefficient
Upstream to downstream	3	0.7	0.1	0.2	0.673
Downstream to upstream	2	0.8	0.1	0.1	0.887

TABLE 10 | Search parameters of DBSCAN clustering.

Direction of trajectory sets	eps	Minpts	Number of clusters	Contour coefficient
Upstream to downstream	2.9	5	3	0.826
Downstream to upstream	3.2	6	2	0.826

are significantly less utilised, with only 8 and 5 trajectories, respectively. This indicates a highly uneven distribution of ship trajectories across the different waterways.

Table 11 and Figure 11 reveal that there are two primary routes for ships sailing from the downstream to the upstream. The distribution of trajectories is more balanced between the middle waterway and the north waterway, suggesting a relatively even utilisation of these routes.

In summary, after filtering out noise trajectories during the clustering process, a total of 15 distinct traffic patterns were identified in the Fujiangsha water area from June to August 2019. The Sankey diagram in Figure 12 visually represents the flow and distribution of these traffic patterns. The specific data is provided in Table 12.

To facilitate the description and analysis of each cluster, the main clusters are assigned labels, which serve as classification targets

TABLE 11 | Statistics of various clusters in clustering results of trajectory sets from upstream to downstream of Yangtze River.

Direction of trajectory sets	Subset of trajectory sets	Number	Total
Upstream to downstream	Cluster through north waterways	8	4790
	Cluster through middle waterways	4771	
	Cluster through south waterways	5	
	Noise	6	
Downstream to upstream	Cluster through north waterways	3627	6194
	Cluster through middle waterways	2552	
	Noise	15	

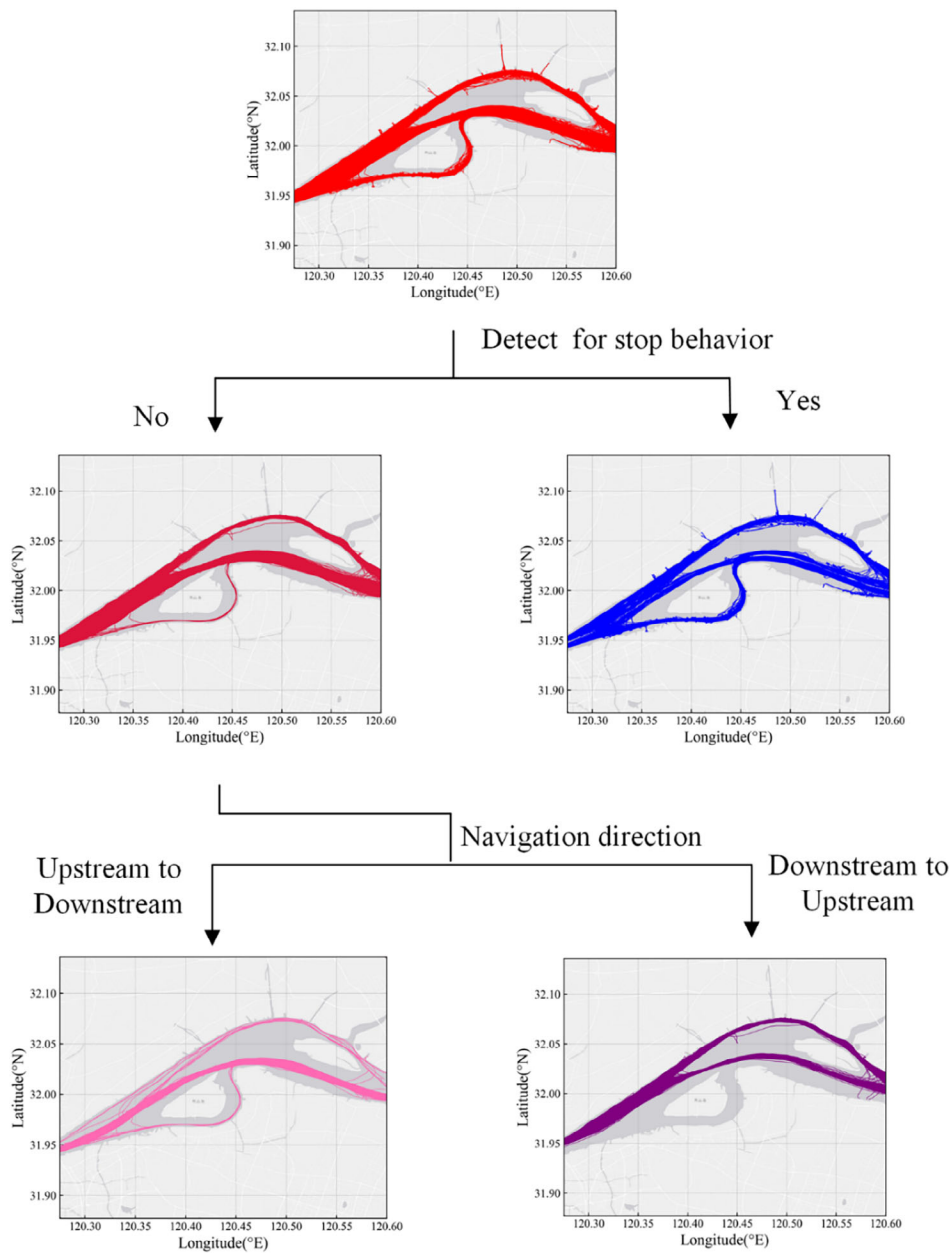


FIGURE 9 | Secondary classification of trajectories between upstream and downstream.

for traffic pattern prediction. These labels are detailed in Table 13. Due to the chaotic and irregular trajectories observed between the upstream and downstream of the Yangtze River and the natural harbour port waterways, these trajectories are excluded from the dataset.

To determine the similarity of the primary traffic patterns, this paper delves into the traffic flow characteristics of the three main traffic patterns. Specifically, the study area is partitioned into 500×500 grids. Within each grid, we calculate traffic-related characteristic information of AIS track points associated with Cluster 2, Cluster 4 and Cluster 5. The results are shown in Figures 13–15 respectively.

As shown in Figure 13(a), the trajectory zone of Cluster 2 forms uniform ribbons with homogeneous density distribution. Figure 13(b) indicates that the course of the vessels changes from

about 50° to approximately 100° . From Figure 13(c), it can be seen that the speed distribution exhibits an obvious edge-like trend, with higher speeds near the midline of the Yangtze River channel compared to the lower side.

In Figure 14(a), the trajectory zone of Cluster 4 is wider in the upper reaches of the Yangtze River and narrower in the lower reaches, with inhomogeneous density distribution and two distinct high-density lines near the upper reaches. Figure 14(b) shows that the course of vessels in this cluster changes from around 300° to approximately 240° . Figure 14(c,d) reveal a similar edge-like speed distribution, where speeds near the midline of the Yangtze River channel are higher than those on the upper side.

Figure 15(a) illustrates that the trajectory zone of Cluster 5 is wider in the upper reaches and narrower in the lower reaches

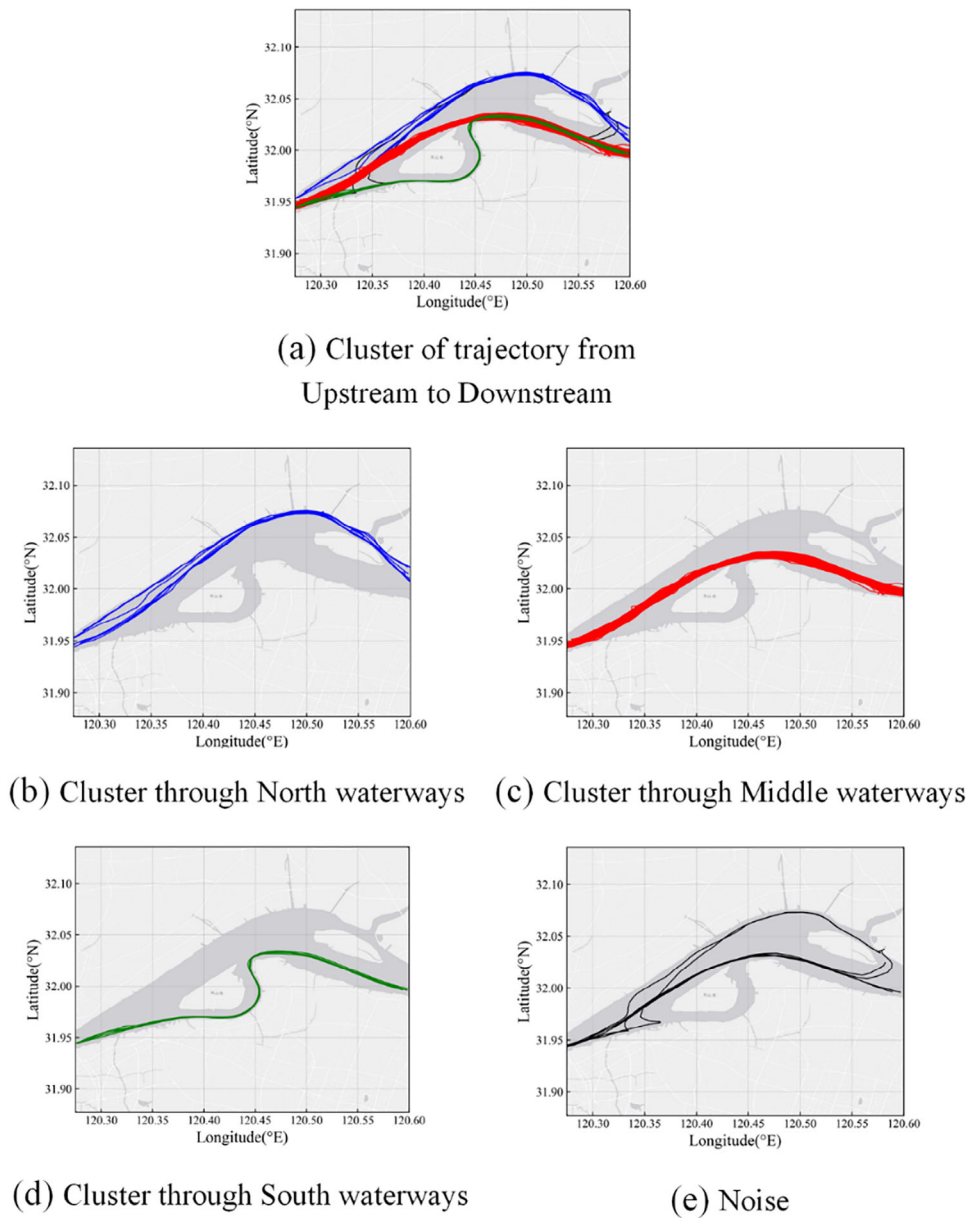


FIGURE 10 | Visualisation of trajectory clustering results from upstream to downstream.

of the Yangtze River, forming a razor-like shape. The density distribution is inhomogeneous with a distinct single high-density track midline. Figure 15(b) reveals that the course of vessels changes from approximately 300° to about 240° . Figure 15(c,d) also demonstrate an edge-like speed distribution, with higher speeds near the midline of the Yangtze River channel compared to the upper side.

In summary, the three main traffic patterns display significant differences in traffic characteristics such as average speed and course.

3.3 | Traffic Patterns Prediction Results

Feature datasets are constructed through manual labelling based on the extracted traffic patterns. Table 13 shows 13 traffic

patterns. In the original dataset, certain traffic patterns (e.g., Clusters 2, 4 and 5) may dominate model training due to their disproportionately high frequencies, causing algorithms to overfit majority-class patterns while neglecting minority-class patterns (e.g., Clusters 3, 10 and 13). This imbalance can be mitigated through dataset balancing, which forces the model to equally prioritise all patterns and enhances its capability to recognise low-frequency traffic modes. By constraining the sample size of each category to approximately 120, this strategy concurrently addresses two critical issues: (1) avoiding parameter estimation biases (e.g., decision tree split instability) caused by insufficient minority-class samples and (2) preventing training efficiency degradation induced by redundant majority-class samples. The final dataset comprises a total of 1601 samples, with the distribution of the dataset is presented in Table 14. The format of dataset is shown in Table 15.

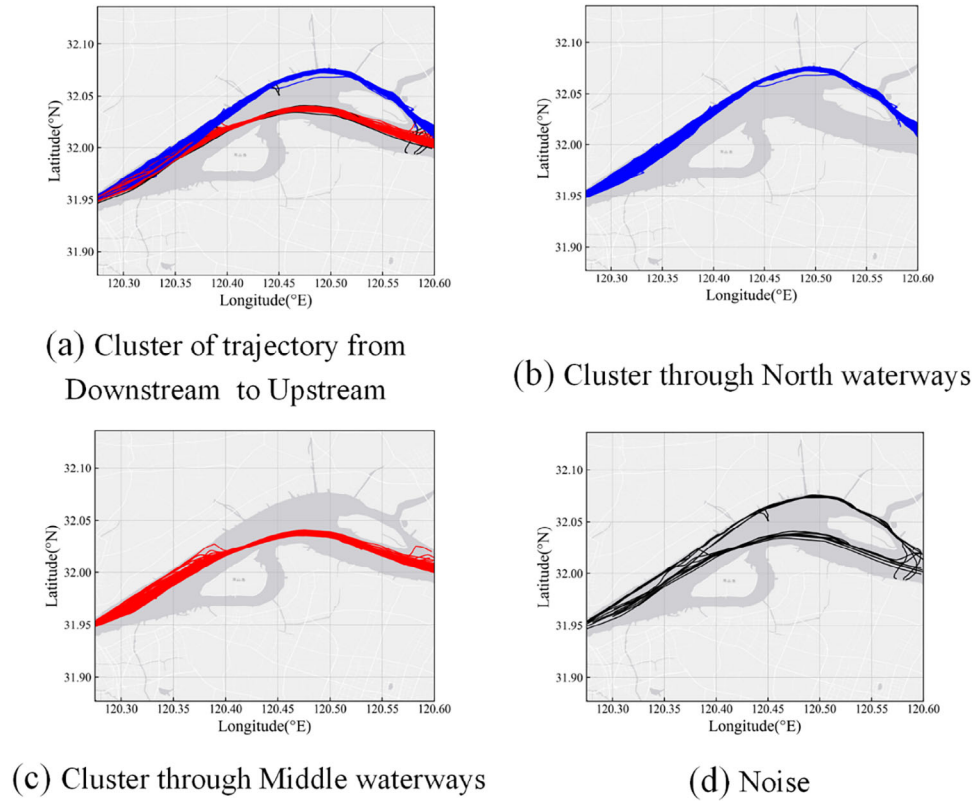


FIGURE 11 | Visualisation of trajectory clustering results from downstream to upstream.

TABLE 12 | Results of traffic pattern classification in Fujiangsha water area.

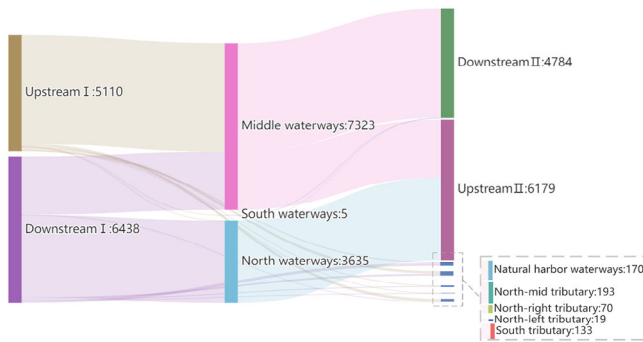
Trajectory cluster	Number	Visualisation
Downward clusters through north waterways	8	Figure 10(b)
Downward clusters through middle waterways	4771	Figure 10(c)
Downward clusters through south waterways	5	Figure 10(d)
Upward clusters through north waterways	3627	Figure 11(b)
Upward clusters through middle waterways	2552	Figure 11(c)
Sailing between upstream and natural harbour waterways	48	Figure 8(a)
Sailing between downstream and natural harbour waterways	122	Figure 8(b)
Sailing between upstream and north-mid tributary	114	Figure 8(c)
Sailing between downstream and north-mid tributary	79	Figure 8(d)
Sailing between upstream and north-right tributary	34	Figure 8(e)
Sailing between downstream and north-right tributary	36	Figure 8(f)
Sailing between upstream and north-left tributary	8	Figure 8(g)
Sailing between downstream and north-left tributary	11	Figure 8(h)
Sailing between upstream and south tributary	122	Figure 8(i)
Sailing between downstream and south tributary	11	Figure 8(j)
Total	11548	

According to Table 15, a total of 11 features are considered. These features including *Origin*, *Cog_{mean}*, *Cog_{range}*, *Cog_{std}*, *Cog_{median}*, *Sog_{mean}*, *Sog_{range}*, *Sog_{std}*, *Sog_{median}*, *L* and *W*. The *R* in the dataset represents the label of the sample, indicating which type of traffic pattern the sample belongs to. The value of *R* ranges from 1 to 13

(i.e., $\{x \in \mathbb{Z} | 1 \leq x \leq 13\}$). The *Origin* in the feature specifically indicates the starting area of the trajectory. The *Cog_{mean}*, *Cog_{range}*, *Cog_{median}* and *Cog_{std}* are the mean, range, standard deviation and median of the course, respectively. The mean, range, standard deviation and median of the speed are represented by *Sog_{mean}*,

TABLE 13 | Labelled ship trajectory clusters for traffic pattern prediction.

Label of cluster	Trajectory cluster	Number
Cluster 1	Downward clusters through north waterways	8
Cluster 2	Downward clusters through middle waterways	4771
Cluster 3	Downward clusters through south waterways	5
Cluster 4	Upward clusters through north waterways	3627
Cluster 5	Upward clusters through middle waterways	2552
Cluster 6	Sailing from upstream to north-mid tributary	44
Cluster 7	Sailing from downstream to north-mid tributary	69
Cluster 8	Sailing from upstream to north-right tributary	9
Cluster 9	Sailing from downstream to north-right tributary	22
Cluster 10	Sailing from upstream to north-left tributary	2
Cluster 11	Sailing from downstream to north-left tributary	8
Cluster 12	Sailing from upstream to south tributary	61
Cluster 13	Sailing from downstream to south tributary	5
	Total	11183

**FIGURE 12** | Sankey diagram of traffic pattern in Fujiangsha water area.**TABLE 14** | The distribution of datasets.

Cluster	Number	Proportion
Cluster 1	120	7.5%
Cluster 2	120	7.5%
Cluster 3	120	7.5%
Cluster 4	120	7.5%
Cluster 5	120	7.5%
Cluster 6	132	8.24%
Cluster 7	138	8.62%
Cluster 8	117	7.3%
Cluster 9	132	8.24%
Cluster 10	120	7.5%
Cluster 11	120	7.5%
Cluster 12	122	7.6%
Cluster 13	120	7.5%
	1601	100%

TABLE 15 | The format of datasets.

Feature	$Traj_1$	$Traj_2$	$Traj_3$
<i>Origin</i>	1	2	2
Cog_{mean}	110	273	270
Cog_{range}	304	42	31
Cog_{std}	77	65	46
Cog_{median}	90	280	265
Sog_{mean}	7	5	4
Sog_{range}	7.7	6	5
Sog_{std}	2.2	2.3	3.1
Sog_{median}	7.1	5.1	4.8
L	100	160	75
W	20	32	8
R	1	2	3

Sog_{median} , Sog_{range} and Sog_{std} , respectively. L represents the length of the ship and W represents the width of the ship. The value of the *Origin* feature is {1, 2}, where 1 indicates that the starting area is upstream of the Yangtze River and 2 indicates that the starting area is downstream.

To enhance the interpretability of the model's decision-making process, this study quantifies and visualises feature importance, explicitly ranking each feature's contribution to prediction outcomes and revealing their mechanistic roles in shaping the model's predictions. Figure 16 illustrates the ranking of feature importance in each model.

As illustrated in Figure 16, the five models exhibit notable variations in feature importance rankings. However, heading-related features are (standard deviation, range, median and

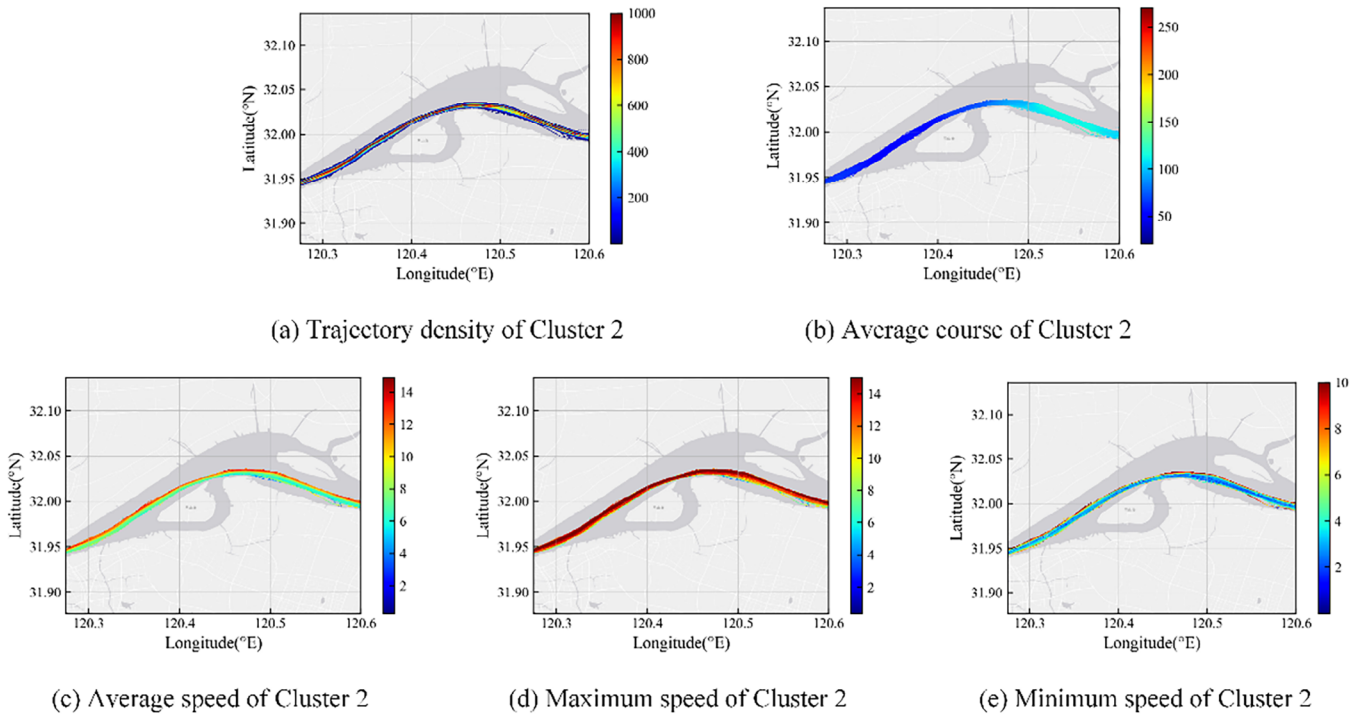


FIGURE 13 | The traffic flow characteristics of Cluster 2.

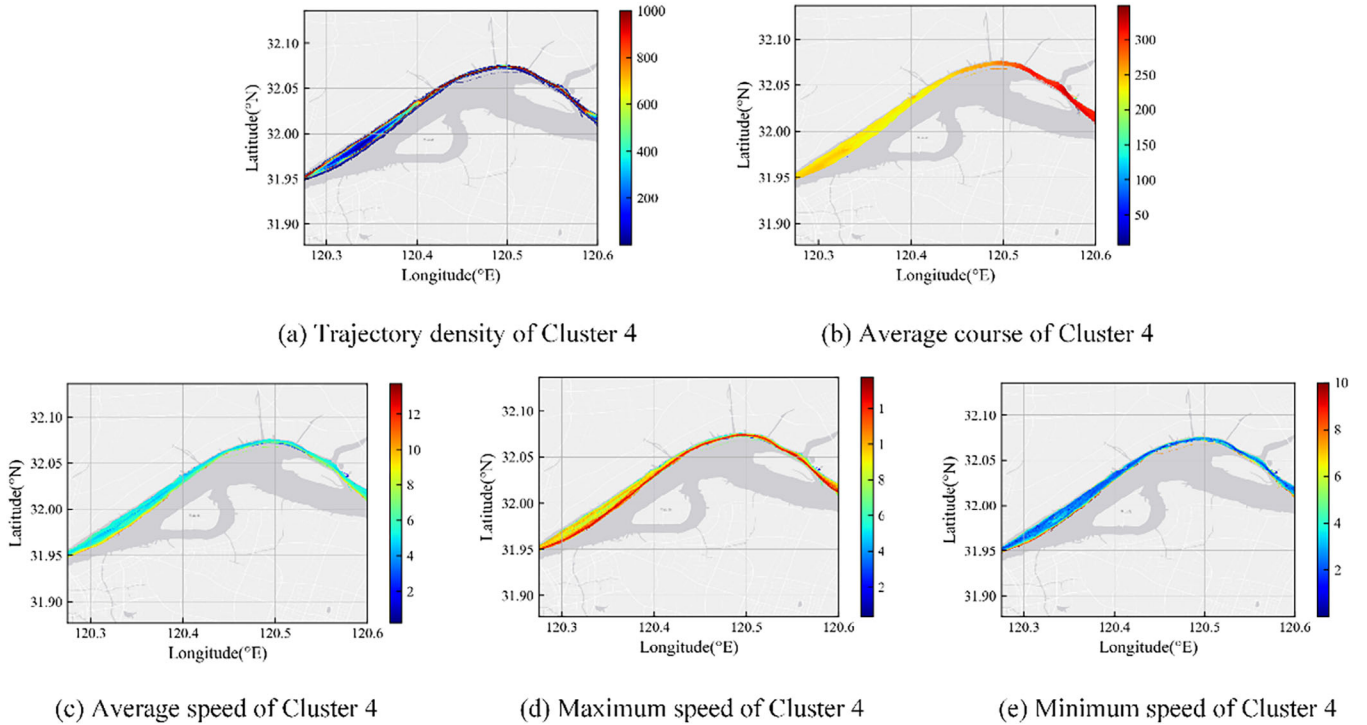


FIGURE 14 | The traffic flow characteristics of Cluster 4.

mean) consistently prioritised in all models. The *Origin* feature demonstrates model-dependent disparities: it significantly contributes to predictions in CART, AdaBoost and XGBoost, yet shows limited influence in RF and even minimal weighting in GB. For vessel dimensions, ship length holds marginally higher importance than ship width, though both rank at moderate-to-low positions across models. These heterogeneous importance

hierarchies collectively demonstrate the interpretability nuances inherent to each model.

To evaluate the performance of the model, the dataset is split into a training set and a test set at a ratio of 8:2, with 80% of the data used for training and 20% reserved for testing.

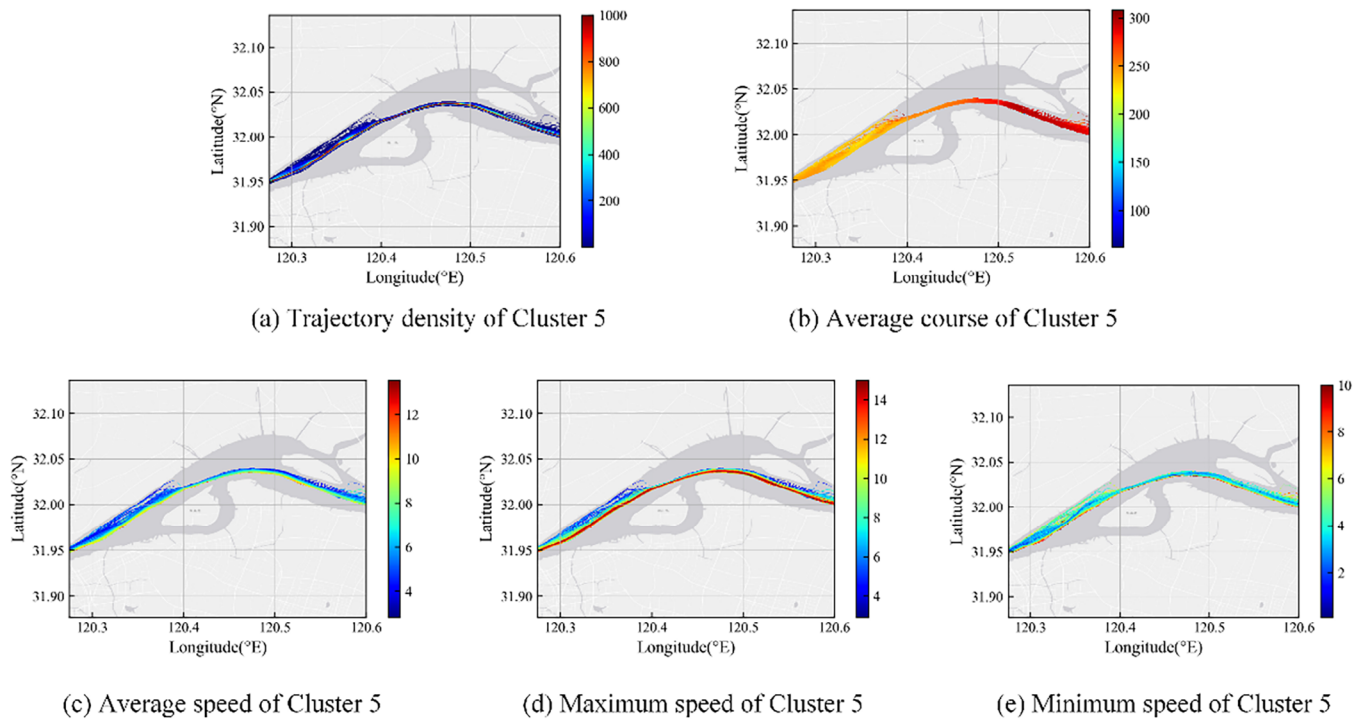


FIGURE 15 | The traffic flow characteristics of Cluster 5.

TABLE 16 | Decision tree model performance metrics for the first 10 min.

Algorithm	Accuracy	Precision	Recall	F1 score
Metrics				
CART	0.8567	0.8769	0.8523	0.8556
RF	0.9844	0.9877	0.9831	0.9847
AdaBoost	0.9782	0.9807	0.9769	0.9773
GB	0.9907	0.9916	0.9908	0.9908
XGBoost	0.9813	0.9848	0.9800	0.9814
STPP	0.9938	0.9942	0.9938	0.9938

Three distinct datasets are created by extracting the first 10 min, 20 min and 30 min of each ship's trajectory. The process of culling and preparing these datasets is illustrated in Figure 17.

1. Decision tree model performance on the first 10 min of data

CART, RF, AdaBoost, GB, XGBoost and STPP are respectively used to train the dataset of the first 10 min. The performance evaluation metrics for these models are summarised in Table 16 and Figure 18.

As shown in Table 16 and Figure 18, the CART tree model exhibits the lowest performance among all models trained on the first 10 min of trajectory data. In contrast, the ensemble methods significantly enhance the model's performance, achieving substantial improvements in accuracy, precision, recall and F1-score. STPP demonstrates the best performance, with evaluation metrics consistently approaching 98%.

2. Decision tree model performance in the first 20 min of data

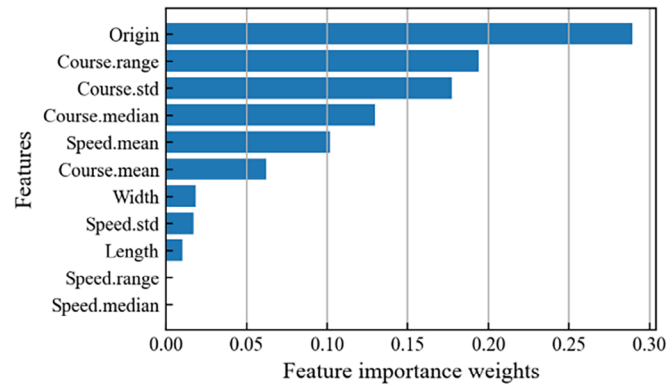
CART tree, RF, AdaBoost, GB, XGBoost and STPP are used to train the dataset of the first 20 min. The model performance evaluation metrics are shown in Table 17 and Figure 19.

As shown in Table 17 and Figure 19, the CART tree model continues to exhibit the poorest performance among all models when trained on the first 20 min of trajectory data. STPP maintains its position as the top-performing model, achieving evaluation metrics consistently around 98%.

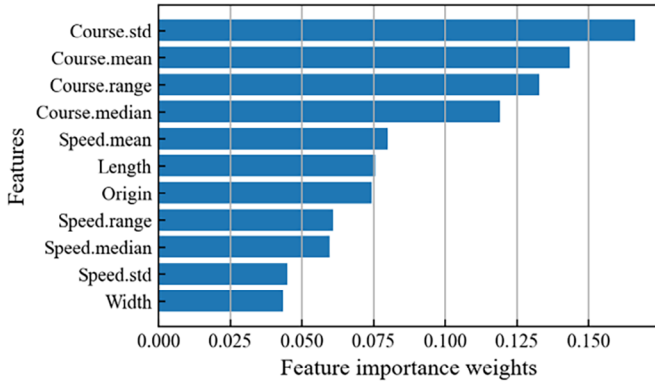
3. Decision tree model performance in the first 30 min of data

CART tree, RF, AdaBoost, GB, XGBoost and STPP are used to train the dataset of the first 30 min, respectively. The model performance evaluation metrics are shown in Table 18 and Figure 20.

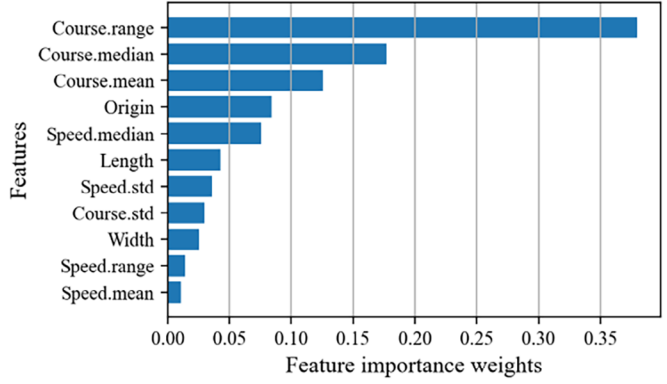
As shown in Table 18 and Figure 20, STPP still demonstrates the best performance, with evaluation metrics consistently around 98%.



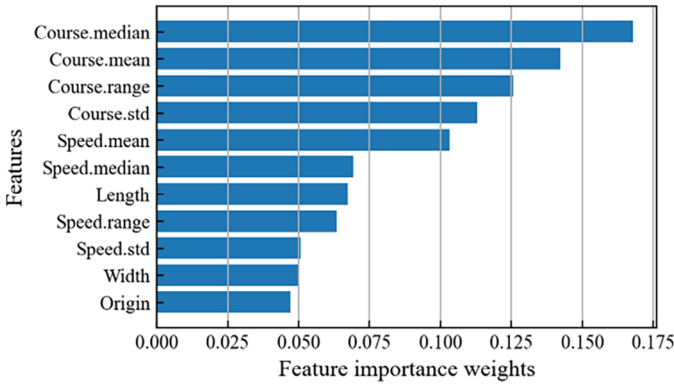
(a) CART



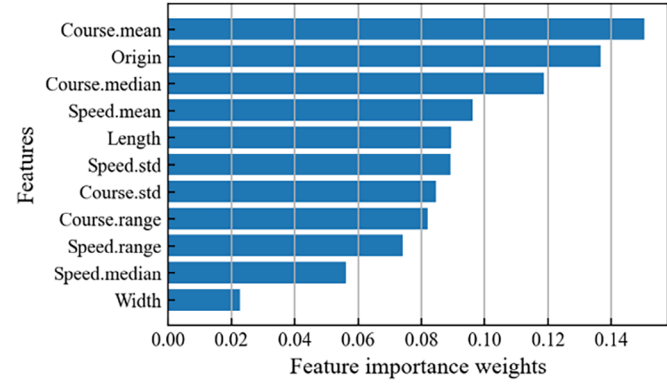
(b) RF



(c) AdaBoost



(d) GB



(e) XGBoost

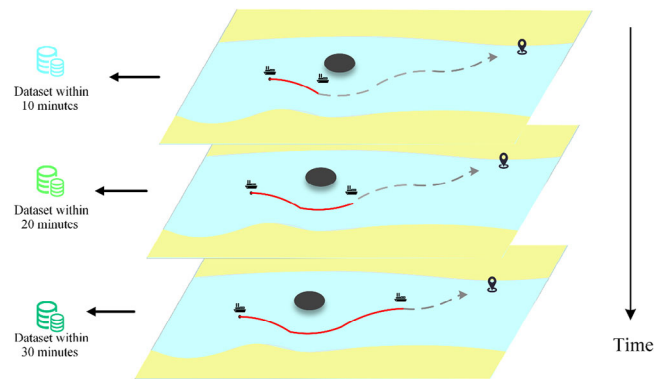
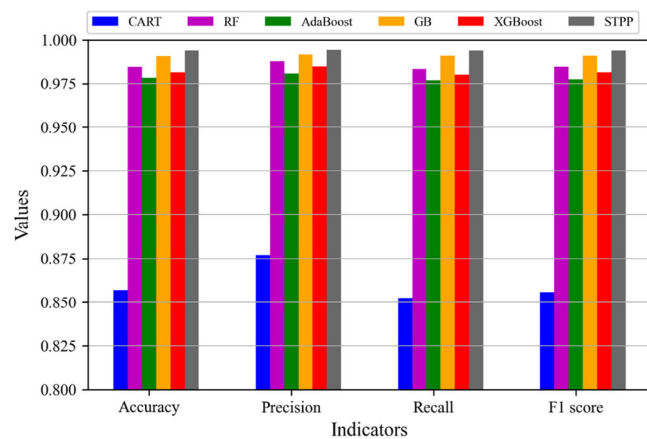
FIGURE 16 | The ranking of feature importance in each model.

TABLE 17 | Decision tree model performance metrics for the first 20 min.

Algorithm	Accuracy	Precision	Recall	F1 score
Metrics				
CART	0.8348	0.8562	0.8236	0.8302
RF	0.9844	0.9857	0.9827	0.9835
AdaBoost	0.9813	0.9836	0.9786	0.9799
GB	0.9875	0.9885	0.9867	0.9871
XGBoost	0.9844	0.9857	0.9827	0.9835
STPP	0.9907	0.9913	0.9898	0.9903

TABLE 18 | Decision tree model performance metrics for the first 30 min.

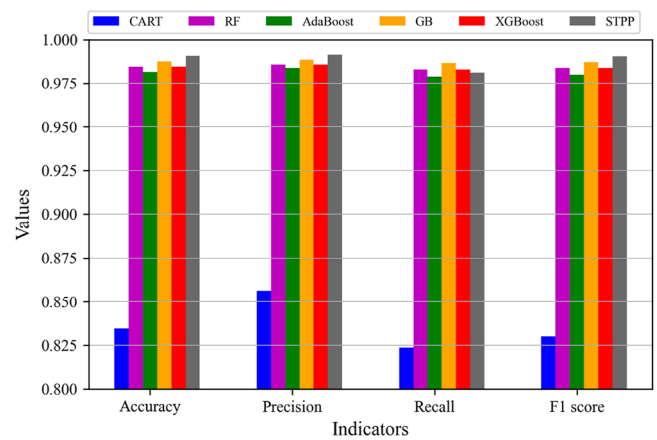
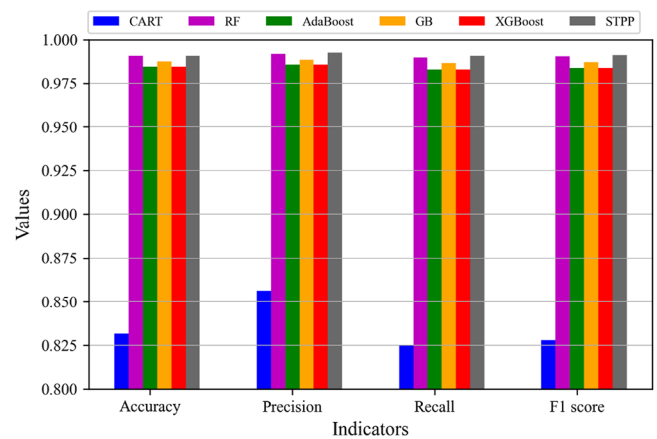
Algorithm Metrics	Accuracy	Precision	Recall	F1 score
CART	0.8318	0.8562	0.8250	0.8279
RF	0.9907	0.9918	0.9898	0.9905
AdaBoost	0.9844	0.9857	0.9827	0.9835
GB	0.9875	0.9885	0.9867	0.9871
XGBoost	0.9844	0.9857	0.9827	0.9835
STPP	0.9907	0.9923	0.9907	0.9910

**FIGURE 17** | The culling of datasets in the first 10 min, 20 min and 30 min.**FIGURE 18** | Column comparison chart of tree model performance in the first 10 min.

4. Comprehensive comparison across different trajectory time lengths

To examine the influence of trajectory time length on model performance, a comprehensive comparison is conducted across datasets representing the first 10 min, 20 min, 30 min and the complete trajectory. The results of accuracy are shown in Figure 21, precision in Figure 22, recall in Figure 23 and F1 scores in Figure 24.

As illustrated in Figures 21–24, the performance metrics of the CART tree model initially decrease and then increase across

**FIGURE 19** | Column comparison chart of tree model performance in the first 20 min.**FIGURE 20** | Column comparison chart of tree model performance in the first 30 min.

the different trajectory time lengths. In contrast, the RF and AdaBoost models show consistent improvement in these metrics. The GB model exhibits no significant trend, while the XGBoost model maintains relatively stable performance. The performance metrics of STPP initially decreased and then improved, ultimately surpassing the performance of all other individual models.

In summary, for trajectory datasets of different time lengths, the ensemble models consistently outperform the single CART tree model. Among the ensemble algorithms, no single model

TABLE 19 | The information for three typical trajectories used to validate traffic pattern prediction models.

	MMSI	Type	Length	Width	Direction	Time interval
Target 1	353465000	Oil tanker	96m	16m	Downward through Middle waterways	30s
Target 2	413358570	cargo ship	99m	16m	Upward through North waterways	30s
Target 3	412762060	cargo ship	134m	18m	Upward through Middle waterways	30s

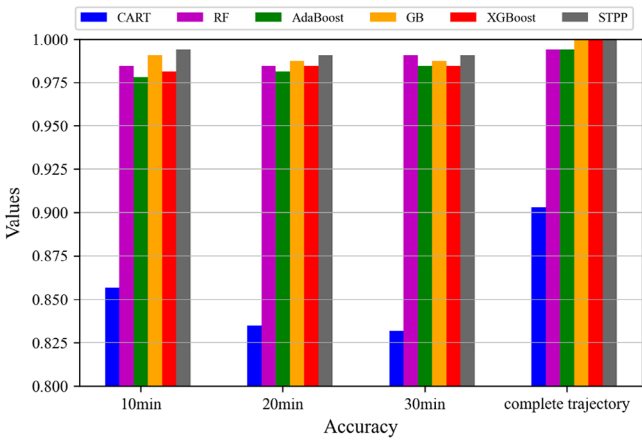


FIGURE 21 | Comparison of tree model accuracy for datasets of different time lengths.

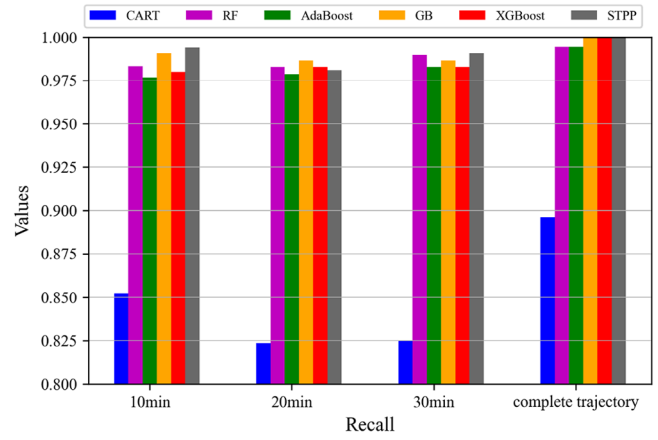


FIGURE 23 | Comparison of tree model recall for datasets of different time lengths.

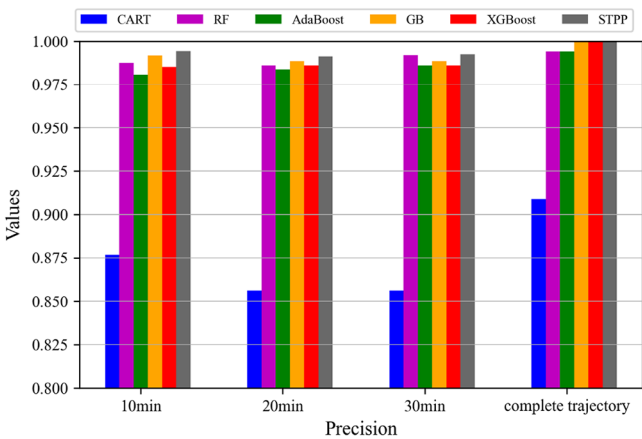


FIGURE 22 | Comparison of tree model precision for datasets of different time lengths.

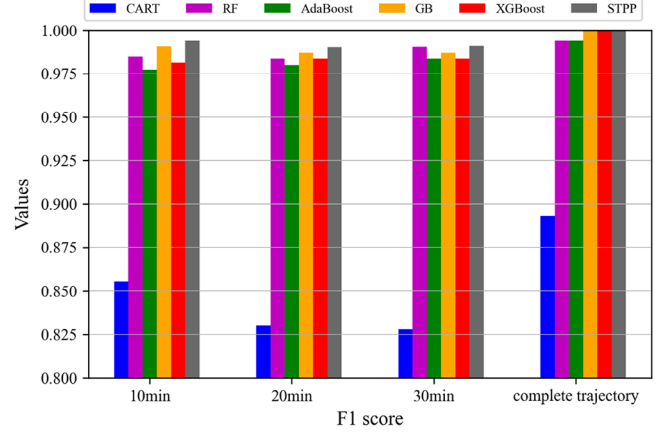


FIGURE 24 | Comparison of tree model F1 score for datasets of different time lengths.

demonstrates a clear advantage over the others. By comparing the performance metrics of the STPP model with other models, it can be seen that the STPP model achieves improvements across all time lengths. This validates the superiority of the proposed STPP model, which combines the strengths of CART, RF, AdaBoost, GB and XGBoost algorithms, in traffic pattern prediction tasks.

3.4 | Validate the Target Ship

The training set and test set utilised in this study are derived from AIS data collected between June and August 2019. To validate the traffic pattern prediction model and assess its ability to predict ship behaviour, three target ship trajectories from 1 September

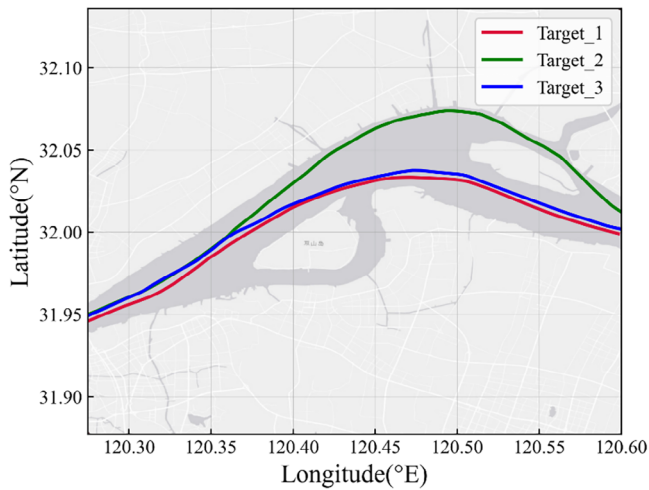
2019, are selected for evaluation. Detailed information about the selected ships is provided in Table 19.

A visualisation of the target ship trajectories is shown in Figure 25. The trajectories include: the downward trajectory of ships passing through middle waterways, the upward trajectory of ships passing through north waterways and the upward trajectory of ships passing through middle waterways.

To evaluate the predictive performance of the model, the first 10 min, 20 min and 30 min of each target ship trajectory are extracted and their corresponding features are input into the CART, RF, AdaBoost, GB, XGBoost and STPP classifiers. To evaluate the predictive performance of the model, the first 10 min, 20 min and 30 min of each target ship trajectory are extracted.

TABLE 20 | Probability prediction results of each case at 10 min of the current trajectory.

Case	Algorithm	Label of traffic pattern												
		1	2	3	4	5	6	7	8	9	10	11	12	13
Case 1	CART	0.70	0.30	0	0	0	0	0	0	0	0	0	0	0
	RF	0	1	0	0	0	0	0	0	0	0	0	0	0
	AdaBoost	0	1	0	0	0	0	0	0	0	0	0	0	0
	GB	0	1	0	0	0	0	0	0	0	0	0	0	0
	XGBoost	0	1	0	0	0	0	0	0	0	0	0	0	0
	STPP	0	1	0	0	0	0	0	0	0	0	0	0	0
Case 2	CART	0	0	0	0.89	0	0	0	0	0.11	0	0	0	0
	RF	0	0	0	0.95	0	0	0	0	0.05	0	0	0	0
	AdaBoost	0	0	0	1	0	0	0	0	0	0	0	0	0
	GB	0	0	0	0.96	0	0	0	0	0.04	0	0	0	0
	XGBoost	0	0	0	1	0	0	0	0	0	0	0	0	0
	STPP	0	0	0	1	0	0	0	0	0	0	0	0	0
Case 3	CART	0	0	0	0	1	0	0	0	0	0	0	0	0
	RF	0	0	0	0	1	0	0	0	0	0	0	0	0
	AdaBoost	0	0	0	0	1	0	0	0	0	0	0	0	0
	GB	0	0	0	0	1	0	0	0	0	0	0	0	0
	XGBoost	0	0	0	0	1	0	0	0	0	0	0	0	0
	STPP	0	0	0	0	1	0	0	0	0	0	0	0	0

**FIGURE 25** | Three target ship trajectories.

The corresponding features of these segments are input into the CART, RF, AdaBoost, GB, XGBoost and STPP classifiers. The probability prediction results are shown in Tables 20–22, respectively (the true labels of the trajectories are highlighted with an orange background).

According to Table 20, when the current trajectory corresponds to the first 10 min of data, the CART algorithm for Case 1 achieves a correct prediction probability of 0.3, while the RF, AdaBoost, GB, XGBoost and STPP models all achieve a correct prediction probability of 1. This demonstrates that, compared to a

single decision tree, the ensemble methods and STPP significantly enhance prediction performance.

For Case 2, the correct prediction probabilities of CART, RF and GB are close to 1, while AdaBoost, XGBoost and STPP achieve perfect prediction probabilities of 1. This indicates that AdaBoost, XGBoost and STPP exhibit superior predictive performance in Case 2.

In Case 3, all algorithms achieve a correct prediction probability of 1, demonstrating that they perform exceptionally well in this scenario.

According to Table 21, when the current trajectory corresponds to the first 20 min of data, both ensemble learning and STPP demonstrate improved prediction performance in Case 1 compared to the single CART tree. In Case 2, AdaBoost, GB and STPP exhibit superior prediction performance, achieving higher accuracy and reliability compared to other models. For Case 3, all algorithms perform well, achieving high prediction accuracy.

According to Table 22, when the current trajectory corresponds to the first 30 min of data, all algorithms demonstrate strong predictive performance in Case 1. RF, AdaBoost, GB, XGBoost and STPP have better predictive performance in Case 2. For Case 3, all algorithms continue to perform well, maintaining high prediction accuracy.

When compared to the results from the first 10 min and 20 min of trajectory data, it is evident that the probability of correct prediction increases as the length of the current trajectory

TABLE 21 | Probability prediction results of each case at 20 min of the current trajectory.

Case	Algorithm	Label of traffic pattern												
		1	2	3	4	5	6	7	8	9	10	11	12	13
Case 1	CART	0.61	0.39	0	0	0	0	0	0	0	0	0	0	0
	RF	0	1	0	0	0	0	0	0	0	0	0	0	0
	AdaBoost	0	1	0	0	0	0	0	0	0	0	0	0	0
	GB	0	1	0	0	0	0	0	0	0	0	0	0	0
	XGBoost	0	1	0	0	0	0	0	0	0	0	0	0	0
	STPP	0	1	0	0	0	0	0	0	0	0	0	0	0
Case 2	CART	0	0	0	0.89	0	0	0	0	0.11	0	0	0	0
	RF	0	0	0	0.87	0	0	0	0	0.13	0	0	0	0
	AdaBoost	0	0	0	1	0	0	0	0	0	0	0	0	0
	GB	0	0	0	1	0	0	0	0	0	0	0	0	0
	XGBoost	0	0	0	0.97	0	0	0	0	0.03	0	0	0	0
	STPP	0	0	0	1	0	0	0	0	0	0	0	0	0
Case 3	CART	0	0	0	0	1	0	0	0	0	0	0	0	0
	RF	0	0	0	0	1	0	0	0	0	0	0	0	0
	AdaBoost	0	0	0	0	1	0	0	0	0	0	0	0	0
	GB	0	0	0	0	1	0	0	0	0	0	0	0	0
	XGBoost	0	0	0	0	1	0	0	0	0	0	0	0	0
	STPP	0	0	0	0	1	0	0	0	0	0	0	0	0

TABLE 22 | Probability prediction results of each case at 30 min of the current trajectory.

Case	Algorithm	Label of traffic pattern												
		1	2	3	4	5	6	7	8	9	10	11	12	13
Case 1	CART	0	1	0	0	0	0	0	0	0	0	0	0	0
	RF	0	1	0	0	0	0	0	0.	0	0	0	0	0
	AdaBoost	0	1	0	0	0	0	0	0	0	0	0	0	0
	GB	0	1	0	0	0	0	0	0	0	0	0	0	0
	XGBoost	0	1	0	0	0	0	0	0	0	0	0	0	0
	STPP	0	1	0	0	0	0	0	0	0	0	0	0	0
Case 2	CART	0	0	0	0.89	0	0	0	0	0.11	0	0	0	0
	RF	0	0	0	1	0	0	0	0	0	0	0	0	0
	AdaBoost	0	0	0	1	0	0	0	0	0	0	0	0	0
	GB	0	0	0	1	0	0	0	0	0	0	0	0	0
	XGBoost	0	0	0	1	0	0	0	0	0	0	0	0	0
	STPP	0	0	0	1	0	0	0	0	0	0	0	0	0
Case 3	CART	0	0	0	0	1	0	0	0	0	0	0	0	0
	RF	0	0	0	0	1	0	0	0	0	0	0	0	0
	AdaBoost	0	0	0	0	1	0	0	0	0	0	0	0	0
	GB	0	0	0	0	1	0	0	0	0	0	0	0	0
	XGBoost	0	0	0	0	1	0	0	0	0	0	0	0	0
	STPP	0	0	0	0	1	0	0	0	0	0	0	0	0

increases. This indicates that the performance of ship navigation intention prediction improves with longer observation times, as more trajectory data provides richer information for the model to make accurate predictions.

While individual ensemble algorithms have already achieved strong prediction performance, the STPP model proposed in this study demonstrates even greater prediction accuracy, particularly in Case 2.

4 | Conclusion

This study introduces a method for classifying maritime traffic patterns using a two-stage trajectory clustering approach. Subsequently, a traffic pattern prediction model is developed, utilising ensembles of decision trees to forecast the traffic pattern of a target ship. The main conclusions of the study are as follows:

1. A two-stage classification method is proposed, which is based on investigation of origin and destination and clustering. Firstly, parent traffic patterns are classified by investigation of origin and destination. Then each parent class trajectory cluster is further refined to identify subclass traffic patterns under the same origin and destination.
2. On the basis of inland traffic pattern classification and feature mining, a traffic pattern prediction model based on ensembles of decision tree algorithms is proposed. Utilising labelled traffic pattern datasets, classifiers are trained using CART, RF, AdaBoost, GB and XGBoost algorithms. These algorithms are then integrated into an STPP framework. The STPP model predicts traffic patterns based on the feature parameters of target ships, demonstrating superior performance compared to individual algorithms.

Author Contributions

Zhao Liu: writing – review and editing, validation, supervision, resources, methodology, investigation, data curation, conceptualisation. **Weipeng Zuo:** writing – review and editing, writing – original draft, visualisation, validation, software, resources, methodology, investigation, formal analysis, data curation, conceptualisation. **Hua Shi:** writing – review and editing, supervision. **Wanli Chen:** writing – review and editing, writing – original draft, validation, software, resources, data curation, conceptualisation. **Xiao Lang:** writing – review and editing, methodology, conceptualisation. **Wengang Mao:** writing – review and editing, supervision, methodology, conceptualisation. **Mingyang Zhang:** writing – review and editing, writing – original draft, visualisation, validation, methodology, investigation, formal analysis, data curation, conceptualisation.

Acknowledgements

This study was supported by the National Natural Science Foundation of China (No. 52171351).

Conflicts of Interest

The authors declare no conflicts of interest.

Data Availability Statement

Data will be made available on request.

References

1. Y. Zhu, O. Gaidai, J. Sheng, A. Ashraf, Y. Cao, and Z. Liu, “Design Prognostics for 4400 TEU Container Vessel by Multi-Variate Gaidai Reliability Approach,” *IET Intelligent Transport Systems* 19 (2025): e12613.
2. M. Zhang, D. Zhang, S. Fu, P. Kujala, and S. Hirdaris, “A Predictive Analytics Method for Maritime Traffic Flow Complexity Estimation in Inland Waterways,” *Reliability Engineering & System Safety* 220 (2022): 108317.
3. Z. Xiao, X. Fu, L. Zhao, et al., “Next-Generation Vessel Traffic Services Systems—From “Passive” to “Proactive.”,” *IEEE Intelligent Transportation Systems Magazine* 15 (2023): 363–377.
4. M. Zhang, G. Taimuri, J. Zhang, et al., “Systems Driven Intelligent Decision Support Methods for Ship Collision and Grounding Prevention: Present Status, Possible Solutions, and Challenges,” *Reliability Engineering & System Safety* 253 (2025): 110489.
5. Z. Xiao, X. Fu, L. Zhang, and R. S. M. Goh, “Traffic Pattern Mining and Forecasting Technologies in Maritime Traffic Service Networks: A Comprehensive Survey,” *IEEE Transactions on Intelligent Transportation Systems* 21 (2020): 1796–1825.
6. C. Liu, M. Musharraf, F. Li, and P. Kujala, “A Data Mining Method for Automatic Identification and Analysis of Icebreaker Assistance Operation in Ice-Covered Waters,” *Ocean Engineering* 266 (2022a): 112914.
7. L. Liu, Y. Zhang, Y. Hu, Y. Wang, J. Sun, and X. Dong, “A Hybrid-Clustering Model of Ship Trajectories for Maritime Traffic Patterns Analysis in Port Area,” *Journal of Marine Science and Engineering* 10 (2022b): 342.
8. Z. Liu, W. Chen, C. Liu, R. Yan, and M. Zhang, “A Data Mining-Then-Predict Method for Proactive Maritime Traffic Management by Machine Learning,” *Engineering Applications of Artificial Intelligence* 135 (2024a): 108696.
9. Z. Liu, W. Yuan, M. Liang, et al., “An Online Method for Ship Trajectory Compression Using AIS Data,” *Journal of Navigation* 77, no. 1 (2024b): 1–22.
10. R. Zhang, H. Ren, Z. Yu, et al., “Self-Supervised Vessel Trajectory Segmentation Via Learning Spatio-Temporal Semantics,” *IET Intelligent Transport Systems* 18 (2024): 2242–2254.
11. L. Huang, C. Wan, Y. Wen, R. Song, and P. van Gelder, “Generation and Application of Maritime Route Networks: Overview and Future Research Directions,” *IEEE Transactions on Intelligent Transportation Systems* 26, no. 1 (2025): 620–637.
12. Q. Ma, X. Du, M. Zhang, H. Wang, X. Lang, and W. Mao, “A Spatial-Temporal Attention Method for the Prediction of Multi Ship Time Headways Using AIS Data,” *Ocean Engineering* 311 (2024): 118927.
13. Z. Liu, H. Gao, M. Zhang, R. Yan, and J. Liu, “A Data Mining Method to Extract Traffic Network for Maritime Transport Management,” *Ocean & Coastal Management* 239 (2023): 106622.
14. Z. Xiao, L. Ponnambalam, X. Fu, and W. Zhang, “Maritime Traffic Probabilistic Forecasting Based on Vessels’ Waterway Patterns and Motion Behaviors,” *IEEE Transactions on Intelligent Transportation Systems* 18 (2017): 3122–3134.
15. P.-R. Lei, T.-H. Tsai, and W.-C. Peng, “Discovering Maritime Traffic Route From AIS Network,” in *2016 18th Asia-Pacific Network Operations and Management Symposium (APNOMS)* (IEEE, 2016), 1–6.
16. C. Huang, X. Qi, J. Zheng, R. Zhu, and J. Shen, “A Maritime Traffic Route Extraction Method Based on Density-Based Spatial Clustering of Applications With Noise for Multi-Dimensional Data,” *Ocean Engineering* 268 (2023): 113036.

17. B. Murray and L. P. Perera, "Ship Behavior Prediction Via Trajectory Extraction-Based Clustering for Maritime Situation Awareness," *Journal of Ocean Engineering and Science* 7 (2022): 1–13.
18. G. Pallotta, M. Vespe, and K. Bryan, "Vessel Pattern Knowledge Discovery From AIS Data: A Framework for Anomaly Detection and Route Prediction," *Entropy* 15 (2013): 2218–2245.
19. L. Kang, Q. Meng, and Q. Liu, "Fundamental Diagram of Ship Traffic in the Singapore Strait," *Ocean Engineering* 147 (2018): 340–354.
20. B. Ristic, B. La Scala, M. Morelande, and N. Gordon, "Statistical Analysis of Motion Patterns in AIS Data: Anomaly Detection and Motion Prediction," in *2008 11th International Conference on Information Fusion* (IEEE, 2008), 1–7.
21. W. Luo and G. Zhang, "Ship Motion Trajectory and Prediction Based on Vector Analysis," *Journal of Coastal Research* 95 (2020): 1183–1188.
22. L. P. Perera, P. Oliveira, and C. Guedes Soares, "Maritime Traffic Monitoring Based on Vessel Detection, Tracking, State Estimation, and Trajectory Prediction," *IEEE Transactions on Intelligent Transportation Systems* 13 (2012): 1188–1200.
23. T. Xiaopeng, C. Xu, S. Lingzhi, M. Zhe, and W. Qing, "Vessel Trajectory Prediction in Curving Channel of Inland River," in *2015 International Conference on Transportation Information and Safety (ICTIS)* (IEEE, 2015), 706–714.
24. M. Abebe, Y. Shin, Y. Noh, S. Lee, and I. Lee, "Machine Learning Approaches for Ship Speed Prediction Towards Energy Efficient Shipping," *Applied Science* 10 (2020): 2325.
25. G. Chen and Z. Li, "Improved Particle Swarm Optimization LSSVM Spatial Location Trajectory Data Prediction Model in Health Care Monitoring System," *Personal and Ubiquitous Computing* 26 (2022): 795–805.
26. C. Zhang, J. Bin, W. Wang, et al., "AIS Data Driven General Vessel Destination Prediction: A Random Forest Based Approach," *Transportation Research Part C: Emerging Technologies* 118 (2020a): 102729.
27. W. Zhang, X. Feng, F. Goerlandt, and Q. Liu, "Towards a Convolutional Neural Network Model for Classifying Regional Ship Collision Risk Levels for Waterway Risk Analysis," *Reliability Engineering & System Safety* 204 (2020b): 107127.
28. H. Rong, A. P. Teixeira, and C. Guedes Soares, "Maritime Traffic Probabilistic Prediction Based on Ship Motion Pattern Extraction," *Reliability Engineering & System Safety* 217 (2022): 108061.
29. D. Gao, Y. Zhu, J. Zhang, Y. He, K. Yan, and B. Yan, "A Novel MP-LSTM Method for Ship Trajectory Prediction Based on AIS Data," *Ocean Engineering* 228 (2021): 108956.
30. J. Ma, C. Jia, Y. Shu, K. Liu, Y. Zhang, and Y. Hu, "Intent Prediction of Vessels in Intersection Waterway Based on Learning Vessel Motion Patterns With Early Observations," *Ocean Engineering* 232 (2021): 109154.
31. M. Liang, Y. Zhan, and R. W. Liu, "MVFFNet: Multi-View Feature Fusion Network for Imbalanced Ship Classification," *Pattern Recognition Letters* 151 (2021): 26–32.
32. X. Han, M. Pan, Z. Liu, H. Meng, H. Sun, and R. Zhang, "Semantics Analysis Model Based on Deep Learning for Vessel Traffic Service Application," *IET Intelligent Transport Systems* 17 (2023): 2089–2102.
33. J. A. Hartigan and M. A. Wong, "Algorithm AS 136: A K-Means Clustering Algorithm," *Journal of the Royal Statistical Society: Series C (Applied Statistics)* 28 (1979): 100–108.
34. M. Ester, H.-P. Kriegel, and X. Xu, "A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases With Noise," in *Second International Conference on Knowledge Discovery and Data Mining (KDD-96)* (AAAI Press, 1996), 226–231.
35. P. J. Rousseeuw, "Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis," *Journal of Computational and Applied Mathematics* 20 (1987): 53–65.
36. L. R. Abreu, I. S. F. Maciel, J. S. Alves, L. C. Braga, and H. L. J. Pontes, "A Decision Tree Model for the Prediction of the Stay Time of Ships in Brazilian Ports," *Engineering Applications of Artificial Intelligence* 117 (2023): 105634.
37. W.-Y. Loh, "Classification and Regression Trees," *WIREs Data Mining and Knowledge Discovery* 1 (2011): 14–23.
38. R. Yan, S. Wang, and Y. Du, "Development of a Two-Stage Ship Fuel Consumption Prediction and Reduction Model for a Dry Bulk Ship," *Transportation Research Part E: Logistics and Transportation Review* 138 (2020): 101930.
39. Y. Li, Z. Guo, J. Yang, H. Fang, and Y. Hu, "Prediction of Ship Collision Risk Based on CART," *IET Intelligent Transport Systems* 12 (2018): 1345–1350.
40. L. Breiman, "Random Forests," *Machine learning* 45 (2001): 5–32.
41. J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals of Statistics* 29 (2001): 1189–1232.
42. Y. Freund and R. E. Schapire, "A Decision-Theoretic Generalization of on-Line Learning and an Application to Boosting," *Journal of Computer and System Sciences* 55 (1997): 119–139.
43. T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16* (Association for Computing Machinery, 2016), 785–794.
44. B. Morris and M. Trivedi, "Learning Trajectory Patterns by Clustering: Experimental Studies and Comparative Evaluation," in *2009 IEEE Conference on Computer Vision and Pattern Recognition* (IEEE, 2010), 312–319.
45. M. Zhang, J. Montewka, T. Manderbacka, P. Kujala, and S. Hirdaris, "A Big Data Analytics Method for the Evaluation of Ship–Ship Collision Risk Reflecting Hydrometeorological Conditions," *Reliability Engineering & System Safety* 213 (2021): 107674.

Appendix A: The content of AIS data processing

AIS data plays a critical role in monitoring ship position and speed, monitoring marine traffic, coordinating ship navigation and optimising route planning. To enhance the accuracy and reliability of data analysis and facilitate computer-based identification and processing, it is essential to clean and pre-process AIS data. The main contents are as follows.

1. Error data: Identify and remove data entries that do not conform to specifications, such as MMSI values, latitude, longitude, speed and heading information that fall outside normal ranges.
2. Duplicate data: Detect and eliminate duplicate records where MMSI, position information and timestamps are identical.
3. Missing information data: Address data entries with missing critical information, such as MMSI, latitude, longitude, heading, speed, or timestamps.

To ensure data quality and usability, the following methods are applied during pre-processing:

1. Data encoding: Encode different traffic pattern categories into numerical labels to represent model prediction results clearly.
2. Data interpolation: Fill in the missing data caused by external factors or data cleaning. Smoothing techniques are employed to improve data quality and ensure the continuity and integrity of ship trajectories.
3. Data normalisation: Normalise the pre-processed data to a specific range, mitigating the adverse effects of outliers and ensuring consistency in the dataset.

Appendix B: Introduction to spatial distance, speed distance and course distance

B.1 | Spatial distance of the trajectory

The spatial distance of the trajectory refers to the difference between two trajectories in terms of their spatial positioning. The methods for calculating spatial distance in trajectory analysis include Euclidean distance, dynamic time warping (DTW) and Hausdorff distance [7]. Euclidean distance requires the length of the compared trajectories to be consistent, which is often not the case for ship trajectories. The computational complexity of DTW distance is high, resulting in slower processing speeds. Hausdorff distance considers the distance between all pairs of points in two sets, which can capture the overall shape difference between sets [44]. Given these considerations, this study employs Hausdorff distance to measure the spatial distance between trajectories.

When calculating the Hausdorff distance between two trajectories, the formula for calculating Sp_s is shown in Equation (B1),

$$\begin{aligned} Sp_s &= \max\{h(Tr_i, Tr_j), h(Tr_j, Tr_i)\} \\ h(Tr_i, Tr_j) &= \max_{p_j^j \in Tr_j} \left(\min_{p_i^i \in Tr_i} (\text{dist}(p_i^i, p_j^j)) \right) \\ h(Tr_j, Tr_i) &= \max_{p_i^i \in Tr_i} \left(\min_{p_j^j \in Tr_j} (\text{dist}(p_j^j, p_i^i)) \right) \end{aligned} \quad (B1)$$

B.2 | Speed distance of the trajectory

The speed distance of the trajectory refers to the difference between different trajectories in terms of the ship speed characteristic. The speed distance of the trajectory Sp_v is used to quantify the dissimilarity between different trajectories in terms of speed [45]. It captures both the central tendency and the temporal variability of speed. The calculation of Sp_v is expressed in Equation (B2),

$$\begin{aligned} Sp_v &= \{Sog_{\text{mean}}, Sog_{\text{median}}, Sog_{\text{range}}, Sog_{\text{std}}\} \\ Sp_v(Tr_i, Tr_j) &= |Sog_{\text{mean}}^i - Sog_{\text{mean}}^j| + |Sog_{\text{median}}^i - Sog_{\text{median}}^j| \\ &\quad + |Sog_{\text{range}}^i - Sog_{\text{range}}^j| + |Sog_{\text{std}}^i - Sog_{\text{std}}^j| \end{aligned} \quad (B2)$$

B.3 | Course distance of the trajectory

Similar to the speed distance of the trajectory, the course distance of the trajectory refers to the difference between different trajectories in terms of the ship course characteristic. The course distance Sp_c of the trajectory is used to quantify the dissimilarity relationship between different trajectories in terms of course [45]. It is also captures both the central tendency and the temporal variability of The calculation of Sp_c is expressed in Equation (B3),

$$\begin{aligned} Sp_c &= \{Cog_{\text{mean}}, Cog_{\text{median}}, Cog_{\text{range}}, Cog_{\text{std}}\} \\ Sp_c(Tr_i, Tr_j) &= |Cog_{\text{mean}}^i - Cog_{\text{mean}}^j| + |Cog_{\text{median}}^i - Cog_{\text{median}}^j| \\ &\quad + |Cog_{\text{range}}^i - Cog_{\text{range}}^j| + |Cog_{\text{std}}^i - Cog_{\text{std}}^j| \end{aligned} \quad (B3)$$

When constructing similarity metric matrix, it is essential to normalize Sp_s , Sp_v and Sp_c respectively. Since these three indicators represent data in different dimensions, normalisation is needed to eliminate the influence of dimensionality. The three indicators are combined and assigned respective weights to form the multi-criteria feature S , as shown in Equation (B4),

$$S = \sum w_i * S_i \in [Sp_s, Sp_v, Sp_c] \quad (B4)$$

where S serves as the multi-criteria feature of similarity measurement, consisting of spatial distance Sp_s , speed distance Sp_v and course distance

Sp_c , with corresponding weights of w_1, w_2, w_3 . These weights satisfy the condition $w_1 + w_2 + w_3 = 1$.

The assignment of weights is achieved through the grid search approach, where a small sample is used to systematically explore parameter combinations and identify the weight values that optimise the performance of the objective function. In this study, the Silhouette coefficient of the K-Means clustering results is adopted as the objective function.

The final constructed similarity matrix SM is shown in Equation (B5),

$$SM = \begin{pmatrix} s_{11} & \dots & s_{1n} \\ \vdots & S_{ij} & \vdots \\ s_{n1} & \dots & s_{nn} \end{pmatrix} \quad (B5)$$

where S_{ij} denotes the multi-criterion feature distance between trajectory i and trajectory j .