Weather-Enhanced Vessel Arrival Time Prediction at Hong Kong Port: A Machine Learning Framework with Multi-Source Data Fusion

Author:

Junpeng Li (5921856)

Supervisors:

Prof.dr.ir. L.A. (Lóri) Tavasszy
F. (Frederik) Schulte

2025.07

Contents

Abstract

Global maritime trade carries over 80% of world cargo, yet vessel arrival time (VAT) prediction remains highly inaccurate. Hong Kong Port experiences average ETA-ATA deviations of 13.8 hours, causing massive congestion costs and supply chain disruptions.

Current methods rely exclusively on either static ETA reports or dynamic AIS data, missing the complete picture. This fragmented approach ignores how ships actually navigate—constantly responding to weather conditions, sea states, and their own physical capabilities.

This study develops a multi-source data fusion framework that integrates four key streams: ETA baselines, real-time AIS movements, marine weather data (wave height, wind speed, swell patterns), and vessel physical parameters (VPP). OpenFE automatic feature engineering handles complex data interactions, while six machine learning models (XGBoost, Random Forest, LightGBM, LSTM, Transformer, TabPFN) are systematically compared.

Testing on Hong Kong Port data shows TabPFN achieves optimal performance with 2.88–3.42 hour prediction errors, which means 43%–47% improvement over ETA baselines. Weather factors occupy 3 of the top 15 important features, contributing 20% of predictive power. Surprisingly, traditional machine learning consistently outperforms deep learning on this structured maritime data. These advances enable optimised berth allocation, reduced port congestion, and more reliable logistics planning, supporting the maritime industry's digital transformation.

Keywords Ship arrival time prediction; multi-source data fusion; machine learning; AIS data; weather data; TabPFN; OpenFE; Hong Kong Port;

1 Introduction

1.1 Background

As an important carrier of international trade, the development scale and growth trend of global maritime trade directly reflect the degree of world economic integration. According to the latest data from the United Nations Conference on Trade and Development (UNCTAD), the global maritime trade volume will reach 1.23 billion tons in 2023, an increase of 2.4% over 2022, showing a strong recovery trend (United Nations Conference on Trade and Development 2024). Maritime transport occupies a dominant position in international trade, undertakes more than 80% of the world's cargo transportation volume, and is an important link connecting the economies of various continents (United Nations Conference on Trade and Development 2023a). With the deepening development of the globalization process, the volume of maritime trade is expected to maintain an average annual growth rate of more than 2% during the period 2024-2028 (United Nations Conference on Trade and Development 2023b).

As a key node in the shipping network, the operation efficiency of ports directly determines the ship turnover speed, cargo handling capacity and logistics costs. Port congestion has become an important

factor restricting the efficiency of the global supply chain. According to McKinsey research, by December 2021, port congestion had reduced global container ship capacity by about 16%, and ocean freight rates had risen to 4-5 times the 2019 level (McKinsey & Company 2022). According to research by economists David Hummels and Georg Schaur, each day of delay can cause economic losses of 0.6%-2.3% of the value of the cargo on board (Hummels and Schaur 2012). In addition, container detention fees are usually between US$75-300 per container per day. These direct costs accumulate rapidly, causing huge economic pressure on the global supply chain (Investopedia 2024).

The uncertainty of Vessel Arrival Time (VAT) has become a major challenge for port management. This uncertainty is most intuitively reflected in the deviation between the estimated time of arrival (ETA) and the actual time of arrival (ATA), which not only affects the efficiency of port scheduling, but may also lead to supply chain disruptions, cargo delays and additional economic costs (Notteboom and Rodrigue 2008). The accuracy of VAT prediction has received widespread attention in academic research. Chu et al. found in their study of port arrival time prediction that the prediction error of traditional methods is usually within the root mean square error range of 25.5 hours, while the machine learning method that integrates AIS data and port call records can reduce this error to 15.5 hours (Chu, Yan, and Wang 2022). In a case study of Hong Kong Port, the world's busiest container hub, the study showed that the accuracy of VAT prediction has a significant impact on port operating efficiency (Lam et al. 2023). This study verified the universal problem of significant deviations between ETA and ATA by analysing the actual operating data of Hong Kong Port. Therefore, using data-driven methods to accurately predict VAT will not only help optimise port operating efficiency, but also provide strong decision-making support for shipping companies, shippers and logistics service providers.

1.2 Significance

This study addresses three critical research gaps identified in the current literature on VAT prediction and makes significant scientific and practical contributions to the field.

Addressing the insufficient application of weather data in VAT prediction: Current research demonstrates inadequate integration of weather data into VAT prediction models (Rahman et al. 2025; J. Yang et al. 2024). Nasir et al. 2024 notes that while many approaches have relied solely on AIS data, limited research has comprehensively integrated diverse data sources including Weather Data. This study constructs a comprehensive multi-source data framework that systematically integrates ETA, AIS, weather and VPP data. By addressing the technical challenges of meteorological data acquisition and processing, including temporal resolution differences and spatial scale mismatches, this research overcomes the limitations of existing studies that often ignore meteorological factors to avoid technical complexities. The framework will fully explore the inherent correlations between various data types through advanced feature fusion methods, enabling more accurate predictions.

Overcoming traditional feature engineering limitations: Traditional manual feature engineering methods in VAT prediction are time-consuming, require extensive domain expertise, and often lead to suboptimal feature sets (Pecan AI 2025). Saleh, Hassan, and Al-Rashid 2023 observes that traditional methods rely heavily on human skills and judgement, making them susceptible to human error and less efficient in handling large volumes of data. This study addresses these limitations by im-

plementing automated feature engineering techniques that can generate a large number of candidate features from raw data and identify the most valuable feature combinations through intelligent selection mechanisms. This approach moves beyond basic ship motion parameters to discover complex feature combinations and interactions that traditional methods often miss, particularly when integrating multi-source data.

Advancing deep learning applications for tabular maritime data: While deep learning methods have been widely used in VAT prediction, most studies focus on traditional neural network architectures like LSTM and CNN, which are primarily designed for sequence or image data (Li, Jiao, and Z. Yang 2023). Hollmann, Müller, Purucker, et al. 2025 demonstrates that specialized tabular foundation models like TabPFN significantly outperform traditional methods on structured data. This study addresses the insufficient application of advanced deep learning methods specifically designed for tabular data by employing TabPFN alongside traditional approaches, enabling optimal utilisation of multi-dimensional structured maritime data including static ship characteristics, dynamic navigation parameters, meteorological conditions, and port information.

The study employs a comprehensive methodological approach using multiple machine learning methods, including Random Forest, XGBoost, LightGBM, Transformer, LSTM, and TabPFN, leveraging the advantages of each model to capture complex patterns in the integrated multi-source data. This multi-model ensemble approach not only enhances prediction accuracy but also provides robust performance across different maritime scenarios and conditions.

The VAT prediction model developed in this study has significant practical implications for the shipping industry. By accurately predicting ship arrival times, the model helps optimise port scheduling, reduce congestion, and improve berth efficiency, thereby minimising operational costs caused by delays such as fuel waste, idle time, and late penalties. The model provides enhanced planning and resource allocation support for logistics operators, port authorities, and shipping companies, improving overall supply chain coordination efficiency.

The implementation of this research enables more intelligent port operations, reduces environmental impact, enhances stakeholder competitiveness in the global shipping industry, and supports the development of smart ports and smart shipping by providing actionable insights for industry digital transformation.

In summary, this study delivers the following core values:

- **Scientific and Methodological Contribution**: Addresses three key research gaps by proposing a comprehensive multidimensional feature fusion framework that integrates weather data effectively, implements automated feature engineering, and applies advanced tabular deep learning methods. This multi-model ensemble approach enhances VAT prediction accuracy and establishes new research directions.
- **Practical Application**: High-precision models optimize port operations, reduce delays, and improve supply chain efficiency by providing reliable arrival time predictions under diverse weather conditions and operational scenarios.
- **Industry Impact**: Supports smart port and smart shipping development by providing stakeholders with actionable insights derived from comprehensive multi-source data analysis, enabling data-driven decision making and digital transformation.

1.3 Research Questions

Main Research Question How can high-precision VAT prediction be achieved through multi-source data fusion and machine learning models?

Sub-Research Questions

- How can fragmented records be effectively connected?
- What model architecture optimally combines ETA with AIS data for spatio-temporal prediction?
- How can the integration of meteorological data and VPP data improve the accuracy of predicting VAT?

The Fig.1 shows a brief process diagram of this paper.

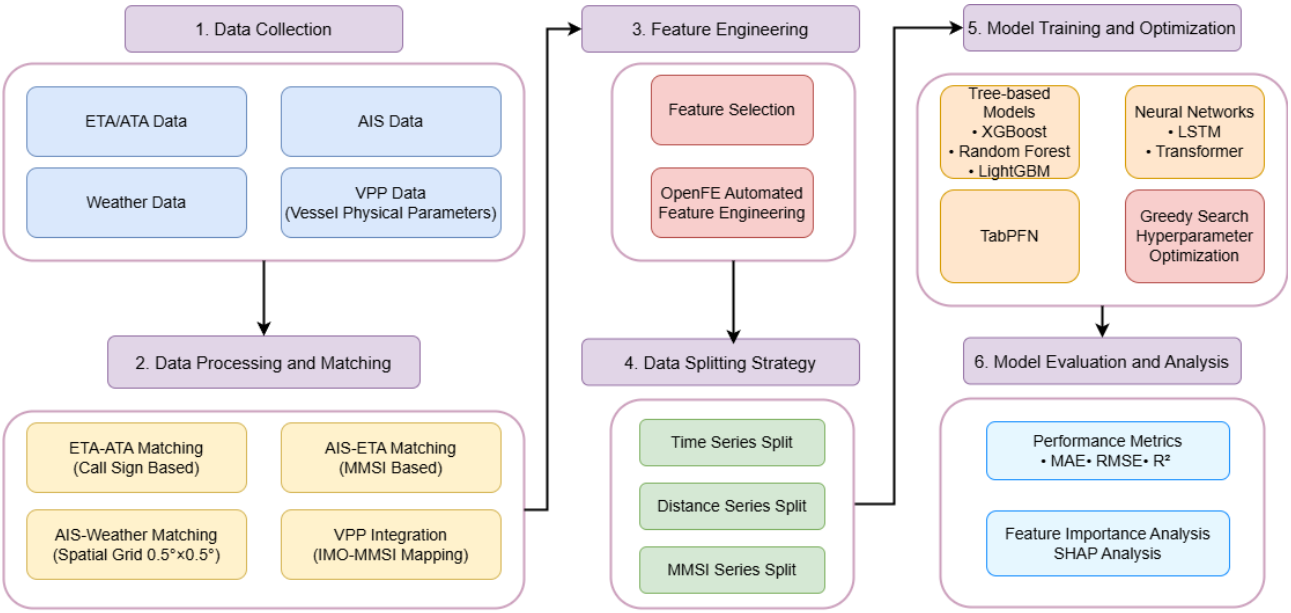

Figure 1: Thesis Structure Overview

Following the research framework shown in Fig.1, this study unfolds in a systematic progression: The first phase involves collecting and integrating multi-source data including ETA/ATA data, AIS data, weather data, and VPP data. The data processing and matching phase then establishes comprehensive data associations through ETA-ATA matching, AIS-ETA matching, AIS-Weather matching, and VPP integration. The feature engineering stage applies feature selection techniques and OpenFE automatic feature engineering to build high-quality feature sets. A three-dimensional data splitting strategy based on Time Series Split, Distance Series Split, and MMSI Series Split ensures reliable model validation. The model training and optimization phase deploys various algorithms including tree-based models like XGBoost, Random Forest, and LightGBM, neural networks such as LSTM and Transformer, plus advanced models like TabPFN, all optimized through greedy search hyperparameter optimization. Finally, comprehensive model evaluation and analysis uses performance metrics including MAE, RMSE, and R^2, while diving deep into the prediction mechanisms through

feature importance analysis and SHAP analysis to build a complete vessel arrival time prediction framework.

1.4 Contribution

- **Systematic weather data fusion**

  This study systematically integrates multi-dimensional marine meteorological data (wave height, wind speed, swell, etc.) into the VAT prediction framework, and a 0.5°×0.5° spatial grid strategy is used to achieve efficient weather data acquisition and matching, filling the scientific gap in the insufficient application of meteorological factors in the maritime forecasting field.

- **Application of OpenFE automatic feature engineering**

  The OpenFE framework is introduced to realize automatic feature generation and selection, breaking through the limitations of traditional manual feature engineering, and automatically discovering complex feature interaction patterns through operations such as numerical transformation, feature combination, statistical features, and time window features, bringing a 2-6% stable performance improvement to traditional machine learning models.

- **The application of TabPFN in the maritime field**

  For the first time, the pre-trained Transformer model TabPFN designed for tabular data is applied to ship arrival time prediction. Competitive performance can be obtained on small-scale data sets without additional training, achieving a 43.6%-46.5% improvement in prediction accuracy.

- **Multi-source heterogeneous data matching system**

  A systematic data matching framework for ETA-ATA, AIS-ETA, AIS-Weather, and AIS-VPP was built. Through Call Sign group matching, MMSI index association, and time window constraints, the technical problems of scattered maritime data sources and inconsistent formats were effectively solved.

- **Three-dimensional data segmentation verification strategy**

  A three-dimensional ordered segmentation strategy based on MMSI, timestamp, and reverse cumulative distance was designed to ensure spatio-temporal continuity and the logical order of ship navigation, avoid the destruction of the intrinsic correlation of data by traditional random sampling, and improve the reliability of model verification.

2 Literature Review

This chapter will review existing research on VAT prediction to build the theoretical foundation for this study and identify key research opportunities. Literature review will cover three main areas: first, looking at how prediction methods have evolved from traditional statistical approaches to modern deep learning techniques to understand where we stand today; second, examining the four main types of influencing factors (ETA, AIS, weather, and VPP data) to see their impact on prediction accuracy; and third, identifying the research gaps in current research to find critical gaps that present opportunities for innovation.

2.1 Evolution of VAT Prediction Methods

2.1.1 Traditional Statistical Methods

Before a series of automated processing methods matured, VAT prediction mainly relied on traditional statistical models and manual experience. Statistical models such as linear regression based on historical average time were difficult to adapt to complex scenarios such as port congestion and faced significant challenges in the actual port environment(Hong Kong Maritime Authority 2022). The prediction methods of this period were mainly based on deterministic mathematical models, assuming that the ship's navigation process followed a fixed pattern, ignoring the complexity and uncertainty of the marine environment.

The traditional prediction model has the following main limitations: First, it cannot effectively capture the non-linear characteristics of ship behaviour. In the actual navigation process, ships may make dynamic decisions such as speed adjustment and route changes due to emergencies. These behaviours have significant non-linear characteristics, resulting in prediction errors generally exceeding 20% (Yu 2021).

Second, the traditional model relies too much on a single data source and lacks systematic integration of multi-source information. Modern port operations involve multi-dimensional information such as meteorological conditions, traffic flow, and port operation efficiency, which have a significant impact on VAT prediction (Filom et al. 2023).

2.1.2 Machine Learning Methods

With the gradual application of machine learning in VAT prediction, it provides new solutions for VAT prediction in port operations. Researches marked an important shift from traditional statistical methods to data-driven methods in VAT prediction. The core advantage of machine learning methods is that they can automatically learn complex patterns and laws from a large amount of historical data without pre-assuming the form of the relationship between variables.

Ensemble learning methods such as Random Forest and XGBoost are particularly prominent in port operations. The Random Forest model based on ship size, speed and departure port characteristics can effectively capture the complex non-linear relationship during the ship's voyage and reduce the VAT prediction error to 12% (Zhang et al. 2022). As an optimised version of the gradient boosting algorithm, XGBoost performs well in handling feature importance evaluation and missing value processing, and is particularly suitable for dealing with the common problem of incomplete data in the marine environment.

The applicability of support vector machines (SVM) and neural networks in VAT prediction has been further verified. Studies have shown that SVM performs well when processing small sample data, especially in high-dimensional feature space, it can find the optimal separation hyperplane, effectively avoiding the over-fitting problem (Flapper 2022). Neural networks can learn complex non-linear patterns through multi-layer structures and show strong learning ability when processing large-scale data sets. However, the performance of these models is highly dependent on feature selection, and unreasonable feature engineering may lead to over-fitting or under-fitting of the model. Feature se-

lection and combination, such as historical ship navigation trajectories, real-time speed, and channel congestion, will significantly affect the prediction effect of the model.

2.1.3 Deep Learning methods

Deep learning methods have shown strong capabilities in the field of VAT prediction, especially in processing complex spatio-temporal sequence data. The research characteristics of this period are the complexity of the model architecture and the enhancement of multi-source data fusion capabilities. Deep learning methods can automatically learn the hierarchical feature representation of data without manually designing features, which has significant advantages when processing high-dimensional spatio-temporal data such as AIS data. Long short-term memory (LSTM) networks are widely used in ship trajectory prediction. Tang et al. constructed a ship trajectory prediction model based on LSTM neural network and achieved good results in long-term position prediction by learning the historical AIS data of Tianjin Port (Tang, Yin, and Shen 2022). LSTM can effectively handle long-term dependency problems through its unique gating mechanism, which is particularly important for tasks with long time series characteristics such as ship navigation. Compared with traditional recurrent neural networks, LSTM can avoid the gradient vanishing problem and perform more stably when processing long sequence data.

At the same time, the latest review studies show that Transformer-based AIS data-driven maritime monitoring is becoming an important research direction (Sun et al. 2024). The application of Transformer models in the maritime field mainly focuses on trajectory prediction methods, behaviour detection and prediction techniques. Its powerful sequence modelling capabilities, especially the ability to capture long-distance dependencies and complex temporal dynamics, make Transformer an effective tool for processing AIS data. These studies provide important technical support for data-driven maritime monitoring tasks and promote the digital transformation of the maritime industry.

In recent years, new deep learning methods for tabular data have also begun to show potential in the field of maritime forecasting. Tabular Prior Data Fitting Network (TabPFN), a Transformer architecture designed specifically for tabular data, has shown unique advantages in processing structured maritime data (Hollmann, Müller, Eggensperger, et al. 2023). TabPFN is pre-trained on a large amount of synthetic tabular data through meta-learning, and can be directly applied to new ship operation data prediction tasks without additional training. This method is particularly suitable for VAT prediction tasks that include multi-dimensional tabular data such as ship static characteristics, dynamic navigation parameters, and environmental conditions. Compared with traditional deep learning methods that require a large amount of labelled data and long training time, TabPFN can quickly obtain competitive prediction performance on small to medium-sized ship datasets, which is of great significance for port application scenarios where data is scarce.

To systematically summarise the methodological progression in VAT prediction research, Table 1 presents a comprehensive comparison of different approaches across three distinct periods, highlighting their key characteristics, advantages, and limitations.

Table 1: Evolution of VAT Prediction Methods

Period	Method Category	Reference	Main Method/-Model	Advantages	Limitations
Before 2010	Traditional Statistical Methods	(Hong Kong Maritime Authority 2022)	Linear regression based on historical averages	Simple computation, easy to understand	Difficult to adapt to complex scenarios like port congestion
		(Yu 2021)	Deterministic mathematical models	Based on fixed pattern assumptions	Prediction errors generally exceed 20%, cannot capture nonlinear characteristics
		(Filom et al. 2023)	Traditional statistical models	Basic statistical approaches	Over-reliance on single data source, lack of multi-source information integration
2010-2018	Machine Learning Methods	(Zhang et al. 2022)	Random Forest (ship size, speed, departure port features)	Effectively capture complex non-linear relationships, reduce VAT prediction error to 12%	Highly dependent on feature selection, unreasonable feature engineering may cause overfitting
		(Flapper 2022)	Support Vector Machine (SVM)	Good performance with small sample data, find optimal separation hyperplane in high-dimensional space, avoid overfitting	Performance highly dependent on feature selection, requires large labeled datasets
2018-Present	Deep Learning Methods	(Tang, Yin, and Shen 2022)	LSTM-based ship trajectory prediction model	Effectively handle long-term dependencies through gating mechanism, avoid gradient vanishing	High computational complexity, requires large training data
		(Sun et al. 2024)	Transformer-based AIS data-driven maritime monitoring	Powerful sequence modeling capabilities, capture long-range dependencies and complex temporal dynamics	Poor interpretability, high computational resource requirements
		(Hollmann, Müller, Eggensperger, et al. 2023)	TabPFN (Transformer architecture for tabular data)	Pre-trained through meta-learning, directly applicable to new tasks without additional training, fast competitive performance on small-medium datasets	Adaptability to large-scale complex scenarios needs verification

2.2 Analysis of influencing factors

2.2.1 ETA Factor

ETA is a basic information for prediction, and its accuracy directly affects the final prediction results. Traditional ship ETA records are usually used for port berth planning, but lack the accuracy required to implement effective plans (Kim et al. 2023). This lack of accuracy mainly stems from the static nature of ETA information, that is, the ETA determined by the ship at the time of departure often cannot reflect the dynamic changes during the voyage.

The reliability of ETA is highly dependent on the accuracy of the target prediction. The ETA prediction of a ship is closely related to the target prediction, and accurate target information is the prerequisite for accurate ETA prediction (Rong et al. 2020). In a complex shipping network, a ship may change its destination during the voyage, and this destination uncertainty will directly affect the accuracy of ETA. The global nature of the shipping industry allows ships to sail from any port to any other port, and can adopt different sailing speeds and routes. This flexibility increases the complexity of ETA prediction.

In practical applications, ships report their estimated departure time (EDT) 36 hours before leaving the port, providing an important time reference point for ETA prediction (Thunberg et al. 2023). At the same time, the quality of ETA information is also affected by human factors. Crews' estimates of arrival times are often based on experience and intuition, lacking a scientific calculation basis.

Modern ETA prediction methods pay more and more attention to dynamic adjustment and real-time updating. By integrating real-time AIS data, weather information and port status, dynamic correction of ETA is achieved to improve the timeliness and accuracy of predictions.

2.2.2 AIS Factor

Automatic Identification System (AIS) data is the core data source for modern VAT prediction. Since the AIS system was mandatory by the International Maritime Organization (IMO) in 2004, it has provided a rich data basis for ship trajectory prediction (Park, Sim, and Bae 2021). The implementation of this global standard enables maritime researchers to obtain unprecedented ship motion data.

AIS data contains three main categories of information: static information such as Maritime Mobile Service Identity (MMSI), ship type, etc. and dynamic information such as ship position, heading, speed, timestamp, etc. Its advantage lies in its fully automated nature, which does not require human intervention. Unlike radar systems, the AIS system uses a longer wavelength and is not affected by weather or sea conditions. (Park, Sim, and Bae 2021), while its feature of updating every few minutes enables researchers to capture subtle changes in ship motion.

The latest research shows that the application of AIS data in complex traffic environments is effective. Analysis of AIS data sets from different water characteristics shows that AIS data-driven ship trajectory prediction can effectively assist in identifying abnormal ship behaviour and reducing maritime risks (Liu et al. 2023).

2.2.3 Weather Factor

Weather conditions have a significant impact on ship navigation behaviour and arrival time, and are a key factor that cannot be ignored in VAT prediction. Quantitative analysis shows that when the wind speed exceeds 10 m/s, the probability of ship arrival delay will increase by 35% (Zhou et al. 2022). This significant statistical relationship reveals the important impact of meteorological conditions on shipping operations.

Research shows weather conditions such as wind speed, wave height and current have a direct impact on ship speed. Correlation analysis shows that significant wave height and swell significant wave height have a moderate correlation with ship speed (18%-19%) (Dorsser et al. 2015). Wind speed has a low negative correlation with ship speed (about -8%). Although the correlation is not as significant as wave height, its impact may be amplified under extreme conditions.

Visibility is another important meteorological factor. When visibility is less than 1 nautical mile, the uncertainty of the ship's arrival time increases significantly (Brandt 2023). There are significant differences in the sensitivity of different types of ships to meteorological conditions, which provides the possibility of combining VPP data. Large container ships are usually more resistant to strong winds and waves than small bulk carriers. This ship type specificity requires the prediction model to be able to adjust the weight of meteorological effects according to the specific ship type.

2.2.4 VPP Factor

As an important factor affecting navigation performance, VPP plays a key role in arrival time prediction. Static characteristics of ships such as ship type, dead weight tonnage (DWT), gross tonnage (GT), length, ship width and ship construction year constitute the basic framework of ship performance (Rahman et al. 2025).

Ship type is one of the most important classification parameters. Different types of ships (such as container ships, bulk carriers, tankers, etc.) have very different design characteristics and operating modes. Container ships are usually designed to sail at high speeds to meet the time requirements of regular liner services; while bulk carriers focus more on cargo capacity and have relatively low sailing speeds. This type difference directly affects its speed-power relationship, fuel consumption characteristics and sensitivity to environmental conditions.

There are significant differences in the sensitivity of different types of ships to meteorological conditions. Large container ships are usually more resistant to strong winds and waves than small bulk carriers due to their greater draft, lower centre of gravity and better stability design. The load state of a ship will significantly affect its sailing performance. There are obvious differences in the sailing speed, fuel consumption and manoeuvrability of fully loaded ships and empty ships under the same conditions.

Building upon the factor analysis discussion, Table 2 provides a detailed examination of the four primary data categories that influence VAT prediction, analysing their mechanisms, advantages, challenges, and current application status.

Table 2: Analysis of Key Factors Influencing VAT Prediction

Factor Category	Data Characteristics	Reference	Impact Mechanism on VAT	Main Advantages	Main Challenges
ETA Factors	Ship-reported estimated arrival time; Static prediction information	(Kim et al. 2023)	Provides basic time reference for port berth planning	Wide coverage; Easy to obtain	Lacks accuracy required for effective planning; Static nature
		(Rong et al. 2020)	ETA reliability highly dependent on target prediction accuracy	Provides basic time framework	Destination uncertainty directly affects ETA accuracy
		(Thunberg et al. 2023)	Ships report EDT 36 hours before departure, providing important time reference	Provides early warning time	Based on experience and intuition, lacks scientific calculation basis
AIS Factors	MMSI, position, heading, speed and other dynamic information; Updated every minute	(Park, Sim, and Bae 2021)	Provides rich data foundation since IMO mandate in 2004	Fully automated; Weather-independent; Rich data content	Massive data volume; Requires complex processing algorithms
		(Liu et al. 2023)	Effective application in complex traffic environments, identifies abnormal vessel behavior	Effectively assists in reducing maritime risks	Requires advanced data processing and analysis methods
Weather Factors	Wind speed, wave height, visibility, ocean current and other marine environmental data	(Zhou et al. 2022)	35% increase in delay probability when wind speed >10m/s	Direct impact on navigation efficiency	Spatio-temporal resolution differences; Complex data quality control
		(Dorsser et al. 2015)	Significant wave height has moderate correlation with vessel speed (18-19%)	Provides environmental constraint information	Correlation may be amplified under extreme conditions
		(Brandt 2023)	Arrival time uncertainty significantly increases when visibility <1 nautical mile	Visibility is important safety indicator	Different vessel types show significant differences in meteorological sensitivity
VPP Factors	Ship type, deadweight tonnage, gross tonnage, length, beam and other physical parameters	(Rahman et al. 2025)	Affects speed-power relationship; Determines environmental sensitivity	Reflects vessel performance differences; Relatively stable	Static characteristics; Limited marginal effects

2.3 Research gaps and opportunities

2.3.1 Insufficient application of weather data in VAT prediction

In current research on VAT prediction, the integrated application of weather data is still insufficient. Rahman et al. 2025 emphasises that by improving the integration of environmental data into ETA models, the maritime industry can achieve more reliable and precise arrival time forecasts, leading to better voyage planning and reduced operational risks. However, J. Yang et al. 2024 points out that while some studies have incorporated weather conditions into their frameworks, many existing approaches still rely primarily on AIS data and vessel information without fully leveraging diverse environmental data sources.

Nasir et al. 2024 notes that while many prior approaches have relied solely on AIS data, and some incorporated a combination of AIS and vessel information, limited research has integrated diverse data sources including Maritime Weather Data (MWD) comprehensively. Even when weather factors are considered, they are often limited to simple weather data and lack the comprehensive use of multidimensional meteorological data.

Existing studies also face technical challenges in the acquisition and processing of meteorological data. Issues such as differences in temporal resolution, spatial scale mismatch, and data quality control between different data sources make the effective integration of weather data complicated. In order to avoid these technical complexities, most studies choose to ignore the processing of meteorological factors, which results in the VAT prediction model being unable to fully reflect the real marine environmental impact.

In addition, there are significant differences in the sensitivity of different types of ships to meteorological conditions, but existing studies lack personalised modelling methods for such differences. This insufficient application of weather data limits the accuracy and practicality of the VAT prediction model, especially the prediction performance under severe weather conditions.

2.3.2 Traditional feature engineering methods limit the potential of model performance

Current machine learning methods generally use traditional artificial feature engineering methods in VAT prediction, which has obvious limitations. Pecan AI 2025 explains that feature engineering is the process of taking a dataset and constructing explanatory variables — features — that can be used to train a machine learning model for a prediction problem. Traditional manual feature engineering is time-consuming, requires domain expertise, and often leads to suboptimal feature sets. Traditional feature engineering mainly relies on the knowledge and experience of domain experts, which is easy to miss potential valuable features, and it is also difficult to discover complex feature combinations and interactions.

(Saleh, Hassan, and Al-Rashid 2023) observe that traditional methods of marine traffic management rely heavily on human skills and judgment and are susceptible to human error, have limitations in handling large volumes of data, and can be less efficient in predicting potential situations. Existing studies often only use basic ship motion parameters, such as position, speed, heading and other original features, while ignoring high-order features that may be generated by feature combination and transformation. This inadequate feature representation directly limits the machine learning model

from its due potential. Especially when integrating multi-source data, artificial feature engineering methods are difficult to effectively handle the complex relationships between different data sources.

(Koehrsen 2018) demonstrates that automated feature engineering tools can streamline the process by cleaning up data, constructing features, and surfacing relevant variables specific to your data and business problem. The benefits of automated feature engineering include efficiency, bias detection, consistency, and deeper exploration of data. However, the application of automated feature engineering technology in the field of VAT prediction is relatively rare. These technologies can automatically generate a large number of candidate features from raw data and identify the most valuable feature combinations through intelligent selection mechanisms, providing new possibilities for breaking through the limitations of traditional feature engineering.

2.3.3 Insufficient application of advanced deep learning methods for tabular data

In the field of VAT prediction, although deep learning methods have been widely used, most studies still focus on traditional neural network architectures, such as LSTM, CNN, etc. Li, Jiao, and Z. Yang 2023 notes that ship trajectory prediction based on Automatic Identification System (AIS) data has attracted increasing interest as it helps prevent collision accidents and eliminate potential navigational conflicts. However, these methods are mainly designed for sequence data or image data, and have long operation time when processing structured tabular data, which may not be able to fully exert their advantages.

Ship operation data is essentially multi-dimensional structured tabular data, including static ship characteristics, dynamic navigation parameters, meteorological conditions, and port information. However, existing studies rarely explore the application potential of deep learning methods designed specifically for tabular data in VAT prediction. Hollmann, Müller, Purucker, et al. 2025s demonstrate that TabPFN, a tabular foundation model, outperforms all previous methods on datasets with up to 10,000 samples by a wide margin, using substantially less training time. In 2.8 s, TabPFN outperforms an ensemble of the strongest baselines tuned for 4 h in a classification setting. The limitations of this method selection may result in the model being unable to optimally utilise the information in the data.

Especially when dealing with complex prediction tasks with a large number of features, traditional deep learning methods may face challenges such as large training data requirements and high computational complexity. Wu et al. 2025 explains that TabPFN operates in two stages: pre-training and inference. During the pre-training stage, the model is pre-trained on a diverse set of synthetic datasets. In the inference stage, given a new task and a set of labelled examples as a "prompt," TabPFN directly predicts the labels of test samples using in-context learning, without requiring further parameter updates. New deep learning methods designed specifically for tabular data, such as the TabPFN model, have unique advantages in processing structured data, but their application in the field of VAT prediction is relatively rare.

In addition, existing studies are often limited by the computation and performance of traditional machine learning methods when using large-scale feature sets for prediction. How to effectively utilise the rich feature sets generated by automated feature engineering and give full play to the value of

large-scale feature data remains an important issue that needs to be addressed in current research.

To identify specific opportunities for methodological innovation, Table 3 systematically categorises the major research gaps identified in current VAT prediction literature, detailing their manifestations and potential innovation opportunities.

Table 3: Research Gaps and Innovation Opportunities

Research Gap	**Specific Manifestations**	**Technical Challenges**	**Innovation Opportunities**
Insufficient Weather Data Application	Most studies still mainly rely on AIS data; Lack of comprehensive utilization of multi-dimensional meteorological data	Temporal resolution differences; Spatial scale mismatches; Data quality control	Systematic meteorological data integration; Personalized meteorological sensitivity modeling
Traditional Feature Engineering Limitations	Relies on domain expert knowledge; Easy to miss potentially valuable features; Difficult to discover complex interactions	Time-consuming manual feature design; Complex multi-source data relationships	Automated feature engineering techniques; Intelligent feature selection mechanisms
Insufficient Application of Advanced Deep Learning Methods for Tabular Data	Focus on traditional neural network architectures; Limited exploration of advanced methods for structured data	Large-scale feature set processing; High computational complexity; Large training data requirements	Specialized tabular data models like TabPFN; Large-scale feature data value mining

2.4 Conclusion

In literature review, the research on VAT prediction, examining the methodological evolution, key influencing factors and existing research gaps are comprehensively analysed. The review shows from traditional statistical methods to machine learning and then deep learning methods, with each stage bringing improvements in prediction capabilities but also increasing model complexity.

The methodological evolution shows significant progress in prediction capabilities. Traditional statistical methods have limitations in dealing with non-linear ship behaviour and dynamic ocean environments. Then machine learning methods introduced ensemble methods such as random forests and XGBoost to effectively capture feature interactions and improve prediction accuracy. The current era of deep learning has brought complex architectures including LSTM, CNN, Transformer and hybrid models, achieving better temporal pattern recognition and long-distance dependency modelling.

The influencing factor analysis identified four key categories: ETA , AIS , weather and VPP data. Although AIS data has comprehensive coverage as the main data source, weather factors are significantly underutilised despite their important impact on ship operations. The review highlights significant gaps in weather data integration and the limitations of traditional feature engineering approaches.

Three major research gaps are identified:

- Inadequate weather data integration in current forecasting models;
- Limitations of manual feature engineering methods that limit model performance potentially;
- Lack of research on model interpretability despite increasing model complexity.

The findings lay a solid foundation for addressing these research gaps through innovative methodologies that systematically integrate weather data, employ automated feature generation techniques, and leverage advanced models designed specifically for structured maritime data.

3 Data Collection and Processing

This chapter will detail the data collection and processing framework designed for VAT prediction, which will cover three main areas: first, explaining the multi-source data collection process, including an introduction to Hong Kong Port as the study location and detailed descriptions of the four key datasets (ETA/ATA records, AIS data, weather data, and VPP data); secondly, presenting the systematic data matching methodology that aligns temporal and spatial information across different data sources using unified vessel identification systems; and third, outlining the feature engineering preprocessing techniques that transform raw multi-source data into model-ready features.

3.1 Multi-source data collection

3.1.1 Introduction to Hong Kong Port

As one of the world's busiest container hub ports, Hong Kong Port (HKP) is strategically located on the east side of the Pearl River Estuary, backed by the Chinese mainland and facing the South China Sea. It is an important maritime gateway connecting mainland China with the rest of the world. Hong Kong Port consists of multiple port areas on the north shore of Hong Kong Island, the south shore of Kowloon Peninsula and the south-west shore of the New Territories, with a total of more than 470 berths, including 24 container berths, which can accommodate ultra-large container ships at the same time(CEIC Data 2021).

However, high-density port operations also bring challenges to VAT prediction. According to the research of Professor Wang Shuai'an of the Hong Kong Polytechnic University, the average error between ETA and ATA in Hong Kong Port in 2021 was as high as 13.8 hours(Wang 2025). This uncertainty in VAT leads to inefficient port handling and economic losses. Especially during peak periods such as the typhoon season (June-November) and the Spring Festival, port congestion is particularly serious, further exacerbating the uncertainty of arrival time.

Therefore, choosing Hong Kong Port as the research object has important theoretical value and practical significance. On the one hand, the complex port environment and high-density ship traffic of Hong Kong Port provide rich training samples for machine learning models; on the other hand, accurate VAT prediction plays an important role in improving the operational efficiency of Hong Kong

Port and maintaining its status as an international shipping centre. The study by Chu et al. confirmed that the overall ship arrival uncertainty decreases as the ship approaches Hong Kong Port. The random forest method can reduce the prediction error of ship ATA data by about 40% (from 25.5 hours to 15.5 hours), which provides a theoretical basis for the technical feasibility of this study (Chu, Yan, and Wang 2024).

3.1.2 Detailed Description of Raw Datasets

This study constructs a multi-source heterogeneous maritime dataset covering four dimensions: ETA data, AIS data, weather data, and VPP data. The data collection period spans from September to October 2021, ensuring data timeliness and consistency.

ETA & ATA Data The ETA and ATA data provided by the Hong Kong Port Authority is published in XML format, containing four different data files, each recording information from different stages of the vessel's port arrival process.

Data File Functionality:

- **FRP04005i**: Records vessel estimated arrival information, including ETA and basic vessel status
- **FRP05005i**: Records actual vessel arrival information, including precise arrival time and location
- **FRP05505i**: Records vessel departure information for calculating port stay time
- **FRP06005i**: Comprehensive record file containing arrival information and vessel identification data

Table 4: ETA & ATA Data Structure

Field Name	ETA Information (FRP04005i)	ATA Information (FRP05005i)	Departure Information (FRP05505i)	Registry Information (FRP06005i)	Field Description
AGENT_NAME	✓	✓	✓	✓	Vessel agent company name
CALL_SIGN	✓	✓	✓	✓	Vessel call sign (unique identifier)
VESSEL_NAME	✓	✓	✓	✓	Vessel name
timestamp	✓	✓	✓	✓	Record timestamp
G_SQL1	✓	✓	✓	✓	Database query identifier
ETA	✓				Estimated time of arrival
LAST_PORT	✓				Previous port of call
PASI	✓				Port agent service identifier
SHIP_TYPE	✓			✓	Vessel type
STATUS	✓				Vessel status
STATU	✓				Vessel status (duplicate field)
ARRIVAL_TIME		✓		✓	Actual arrival time
CURRENT_LOCATION		✓		✓	Current location
REMARK		✓			Remarks
ATD_TIME			✓		Actual departure time
LAST_BERTH			✓		Last berth
FLAG				✓	Vessel flag
IMO_NO				✓	IMO number

AIS Data The AIS data provides real-time vessel position, speed, and heading information, serving as the core data source for analysing vessel navigation behaviour.

Table 5: AIS Data Structure

Field Name	Data Type	Field Description	Value Range/Format
ID	Integer	Unique record identifier	Auto-increment sequence
mmsi	Integer	Maritime Mobile Service Identity	9-digit code
lon	Float	Longitude coordinate	-180° to 180°
lat	Float	Latitude coordinate	-90° to 90°
STATUS	Integer	Vessel navigation status code	AIS standard status codes
speed	Float	Speed over ground	Knots
ROT	Float	Rate of turn	Degrees/minute
heading	Float	Vessel heading	0° to 359°
SAILING_ANGLE	Float	Sailing angle	0° to 359°
TIME_STAMP	Timestamp	AIS record time	Unix timestamp
VOYAGE_NUMBER	String	Voyage number	Internal shipping company code
RECORD_DATETIME	DateTime	Record generation time	YYYY-MM-DD HH:MM:SS

Weather Data Structure Weather data is obtained from a professional weather website via API, covering detailed environmental parameters for vessel navigation areas to analyse weather factors' impact on VAT. The global sea area is divided into 0.5°×0.5° spatial grid units using a grid processing strategy to enhance data acquisition efficiency. The system automatically retrieves hourly meteorological observation data in each grid unit by calling the API, with latitude and longitude rounded to the nearest 0.5° and time rounded to the nearest integer hour. Initially, a 1° grid was tested, but due to its 110 km spatial span, considering API rate limits, a 0.5° grid was adopted for better balance.

Table 6: Weather Data Structure

Category	Field Name	Unit	Field Description	Value Range
Atmospheric	temp_2m	°C	2-meter temperature	-50 to 50
	visibility	km	Visibility	0 to 50
	sea_level	hPa	Sea level pressure	950 to 1050
Wind	wind_speed_10m	m/s	10-meter wind speed	0 to 50
	wind_dir_10m	°	10-meter wind direction	0 to 359
	wind_gust	m/s	Wind gust speed	0 to 70
Sea State	wave_height	m	Significant wave height	0 to 20
	wave_period	s	Wave period	2 to 25
	swell_height	m	Swell height	0 to 15
	swell_direction	°	Swell direction	0 to 359
	swell_period	s	Swell period	5 to 30

VPP Data Structure One of the focuses of this study is the construction of a complete process of ship physical parameter data collection and fusion, and realises the systematic integration of ship static information and dynamic navigation data through intelligent data crawling and association technology. In order to obtain the physical parameters of a ship efficiently, this study designs a set of crawler code that can automatically obtain key physical parameters, including MMSI, ship name, ship type, year of construction, gross tonnage, dead-weight, tonnage, etc. from a professional ship information

website via the ship's IMO number. Considering the special situation that the ship's MMSI may change in actual business, this solution specially designed a multi-MMSI association mechanism. By establishing an IMO-MMSI mapping relationship table, it effectively solves the problem of MMSI changes caused by equipment replacement or ownership change on the same ship, ensuring the continuity and traceability of historical navigation data.

Table 7: VPP Data Structure

Category	Field Name	Unit	Field Description	Notes
Identification	Index	-	Database index	Internal vessel database ID
	IMO	-	IMO number	7-digit unique vessel identifier
	Name	-	Vessel name	Official registered name
	Call Sign	-	Call sign	Radio communication identifier
	Flag	-	Flag state	Registered country/region
Classification	Type	-	Vessel type	Cargo, container, tanker, etc.
Technical	Year Built	Year	Year of construction	Vessel completion year
	Length Overall (m)	m	Overall length	Maximum length bow to stern
	Beam (m)	m	Beam	Maximum vessel width
	Gross Tonnage	Tons	Gross tonnage	Measure of total vessel volume

3.2 Data Matching

3.2.1 ETA-ATA matching

In the ETA-ATA dataset, it is more convenient to use Call Sign as an index to analyse ship data. Therefore, this study uses Call Sign for group matching. First, based on the practical experience of port operations, the abnormal matches with time differences exceeding 10 days are removed. Then, the ETA and ATA of the same ship are matched by setting a time window of 1 day, i.e. 24 hours.

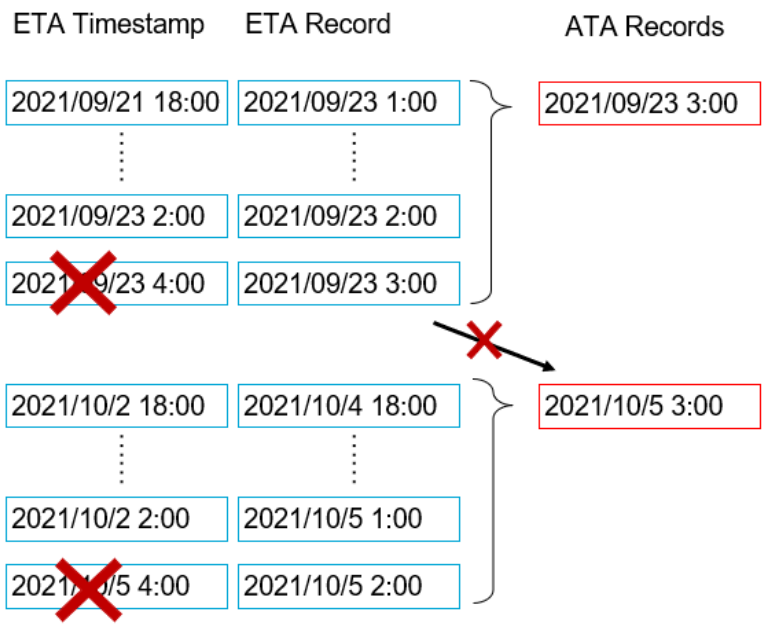

Figure 2: ETA-ATA Matching Process Diagram

The successfully matched ETA-ATA record pair contains the ship identification, time information and the calculated time difference. The distribution of the time difference between ETA and ATA is as follows:

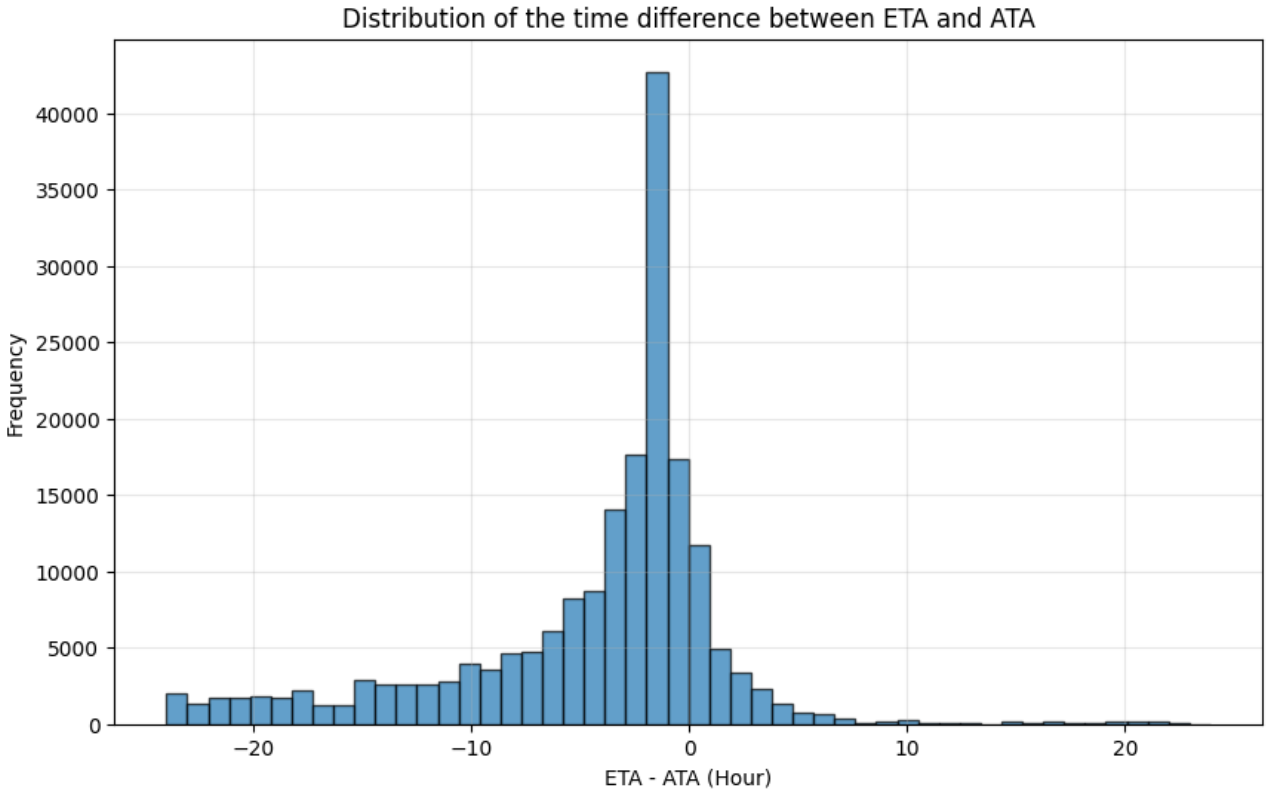

Figure 3: Distribution of the time difference between ETA and ATA

The detailed data is as follows:

Table 8: ETA-ATA Analysis Statistics

Metric	**Value**
Total Records	187,263
Early Arrivals	160,001 (85.4%)
On-Time Arrivals	311 (0.2%)
Delayed Arrivals	26,951 (14.4%)
Mean	$-4.37\,\mathrm{h}$
Median	$-2.20\,\mathrm{h}$
Standard Deviation	$6.35\,\mathrm{h}$
Minimum	$-23.98\,\mathrm{h}$
Maximum	$23.92\,\mathrm{h}$

Finally, add IMO and MMSI columns to the file through the unified IMO-Call Sign-MMSI correspondence table to facilitate subsequent AIS and VPP data matching

3.2.2 AIS-ETA Matching

Considering that AIS records only contain MMSI but lack IMO numbers, this study uses MMSI as the main index field and performs data association on the premise of ensuring that the MMSI is the version recorded in 2021 rather than the version recorded in 2025. To further improve the matching accuracy, this study requires that the timestamp of the AIS record is within a window of 6 hours after the corresponding ETA timestamp, and on this basis, the ETA record closest in time is preferred to ensure the accuracy of the logic and reduce data distortion caused by the large time gap between AIS and ETA data. In addition, to avoid data reuse, this study introduces a uniqueness constraint to ensure that each AIS record is matched only once, thereby maintaining the uniqueness and reliability of the matching results.

The matching algorithm process is designed as an efficient and systematic iterative process: first, traverse all ETA records in chronological order; then, for each ETA event, search for eligible available AIS records within its time window; then, based on the principle of recent time first, select the best match with the smallest time difference from the candidate records; then, update the used AIS record set to ensure the uniqueness of subsequent matches; finally, generate a detailed matching log to record the details of each match to support subsequent debugging, quality assessment and algorithm optimization.

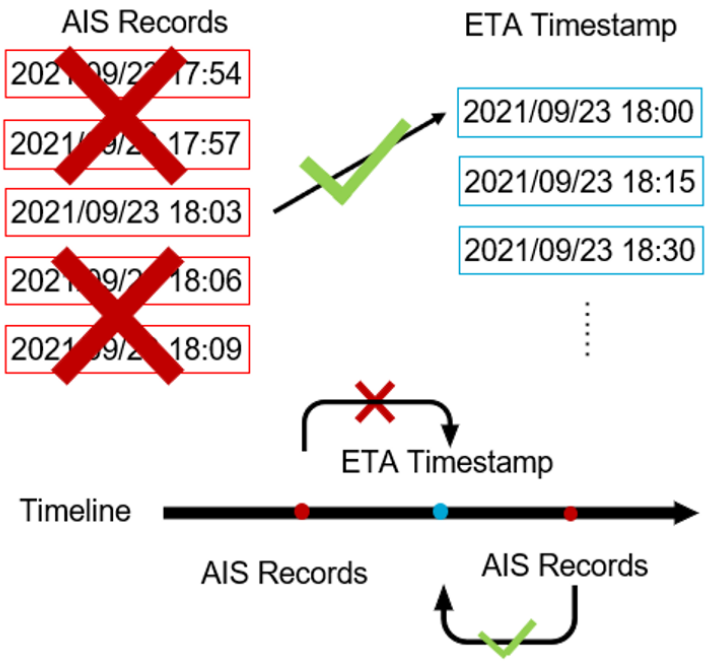

Figure 4: AIS-ETA Matching Process Diagram

The matching results include the real-time position (latitude and longitude coordinates) of the ship, the movement status (speed and heading), time information, and the calculated ETA and the time difference between the ATA and AIS records. The specific latitude and longitude coordinates are distributed as shown in the figure below:

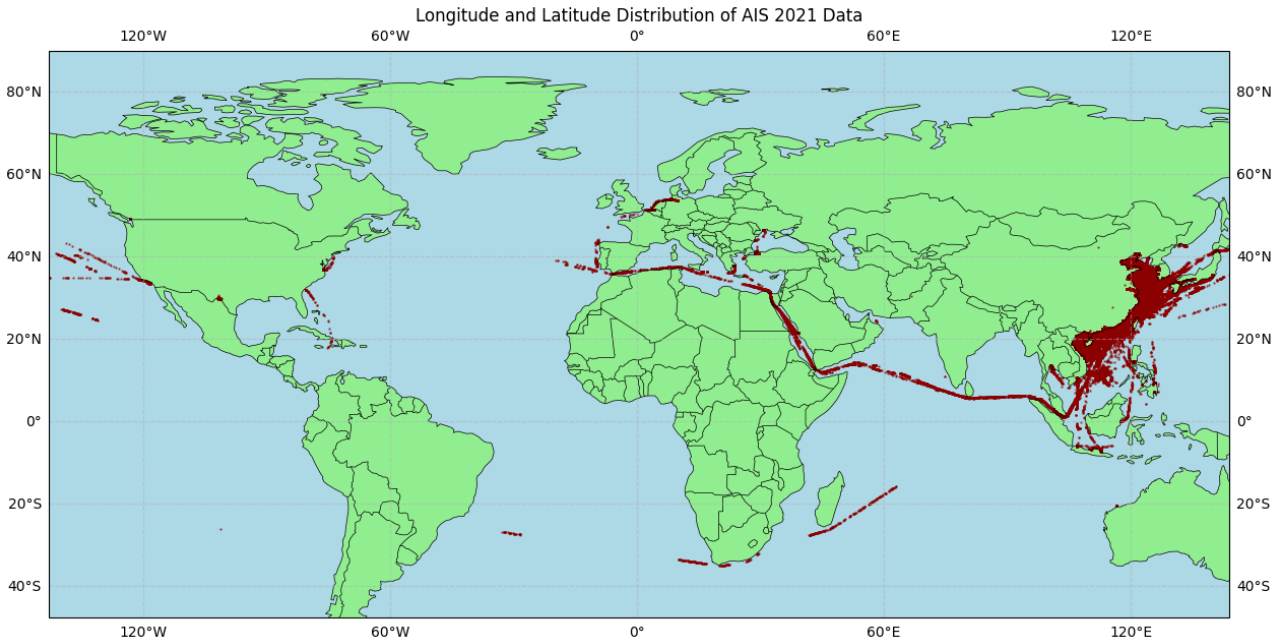

Figure 5: Longitude and Latitude Distribution of AIS Data before Data Matching

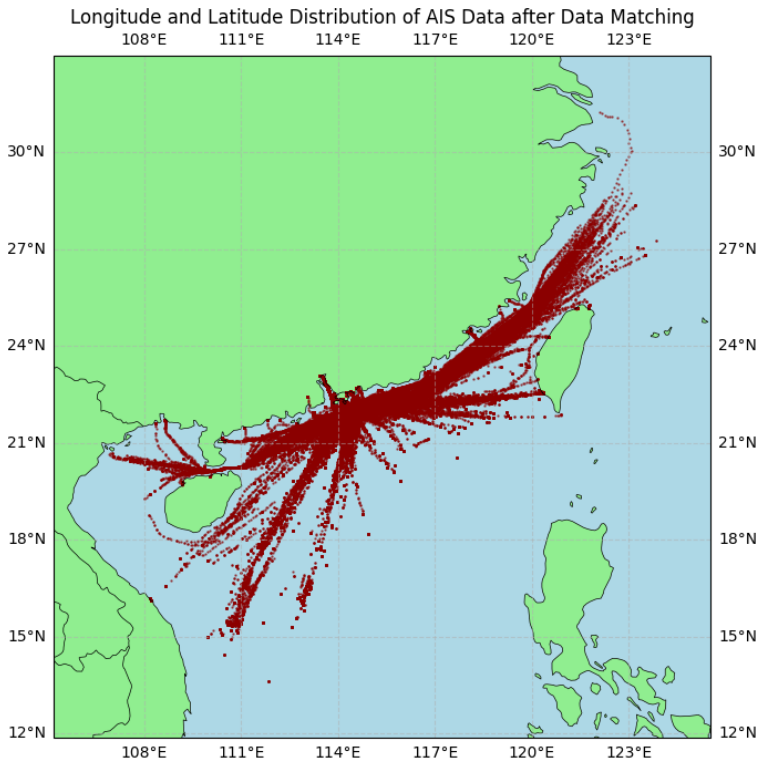

Figure 6: Longitude and Latitude Distribution of AIS Data after Data Matching

The detailed data is as follows:

Table 9: AIS-ETA Data Statistics

Metric	**Value**
Number of Data Points before Matching	8112427
Number of Data Points after Matching	140,833 (75.2% matching success rate)
Longitude Range	106.89° to 123.84°
Latitude Range	13.63° to 31.23°

3.2.3 AIS-Weather Matching

This study uses AIS-Weather data matching and spatial gridding to divide the global ocean into 0.5°×0.5° grids to improve the efficiency and spatial consistency of weather data acquisition (shown in Fig.7, the visualisation for October is in the Appendix Fig.20.).

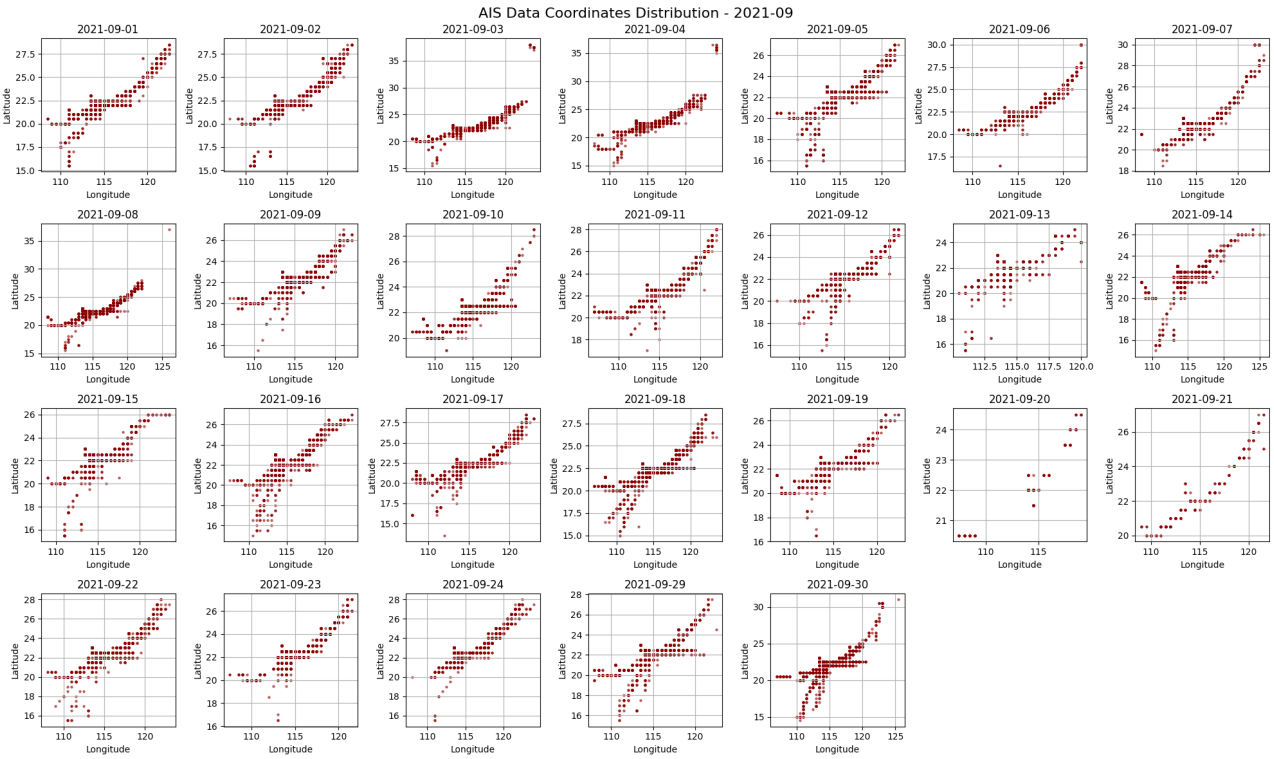

Figure 7: AIS Data Rounded to 0.5° Daily Distribution (September)

The spatio-temporal matching mechanism includes: first, mapping the AIS coordinates to the nearest grid cell; second, filtering and matching to the nearest hourly meteorological observation based on the date; for grid boundary conditions, bilinear interpolation is used to obtain accurate meteorological parameters. Finally, 16 meteorological variables are integrated, including 2-metre temperature, 10-metre wind speed and direction, gusts, visibility, significant wave height, wave period, swell height, swell direction, swell period and sea level height, etc., which provides a high-quality data basis for analysing the impact of meteorology on ship navigation and supports navigation safety and efficiency optimisation.

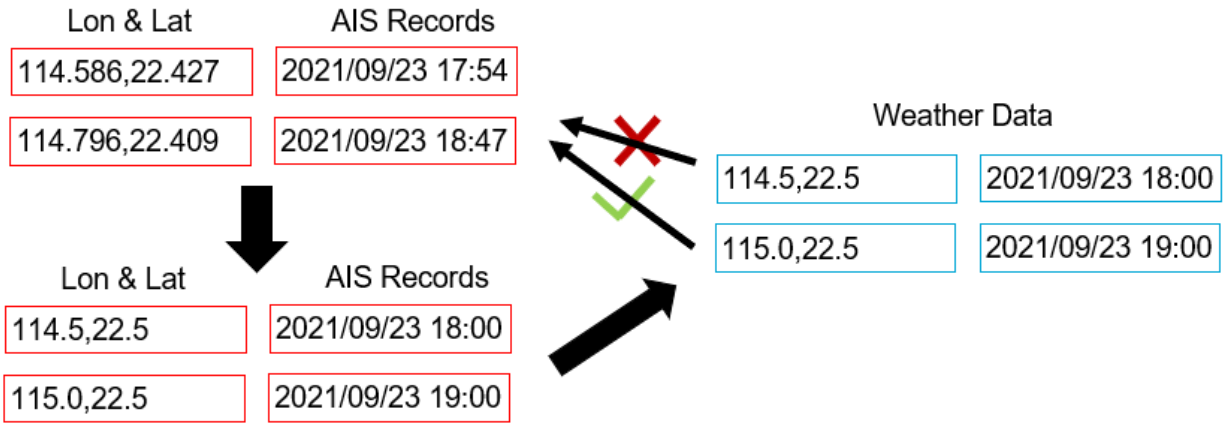

Figure 8: AIS-Weather Matching Process Diagram

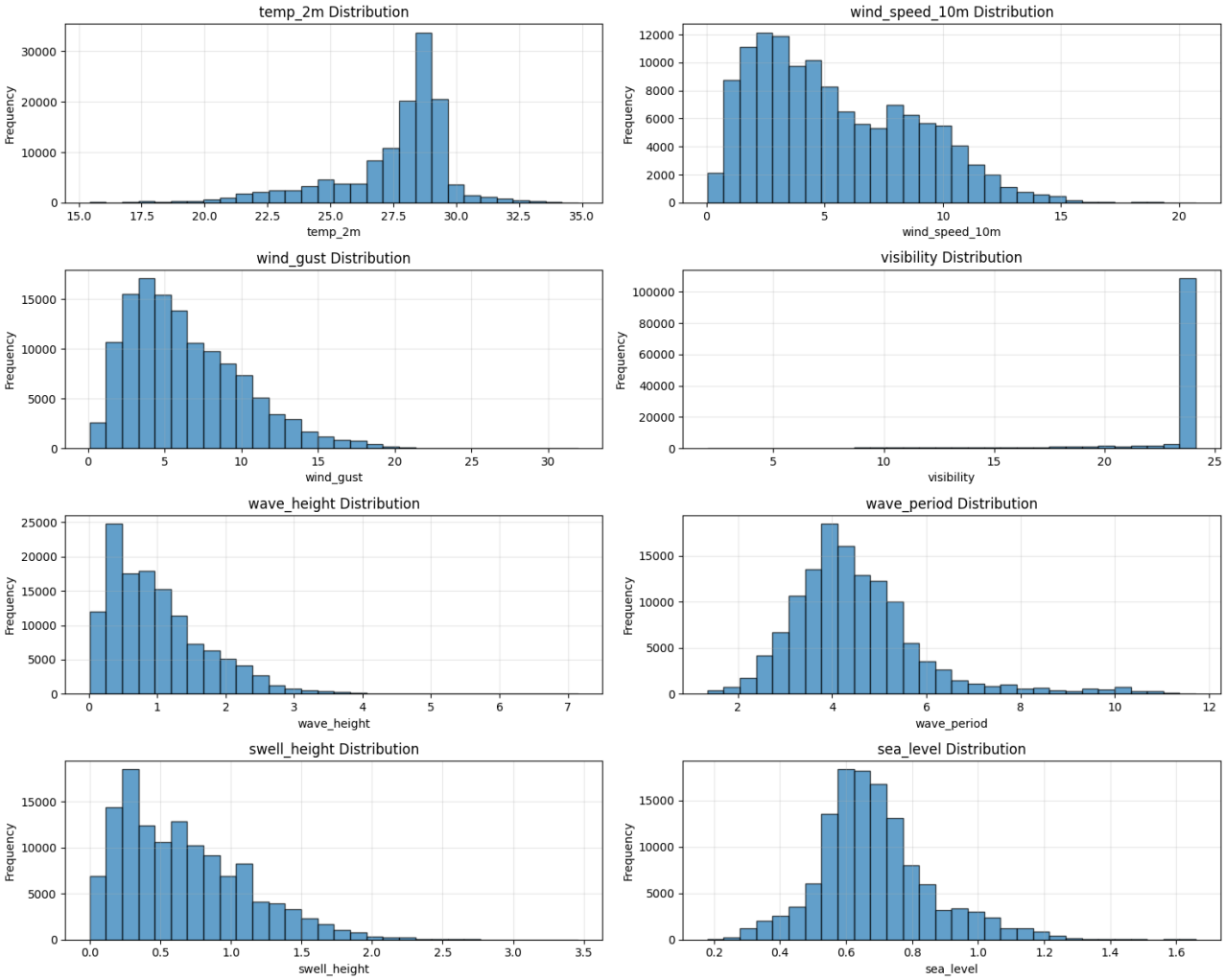

Figure 9: Distribution for Weather Parameters

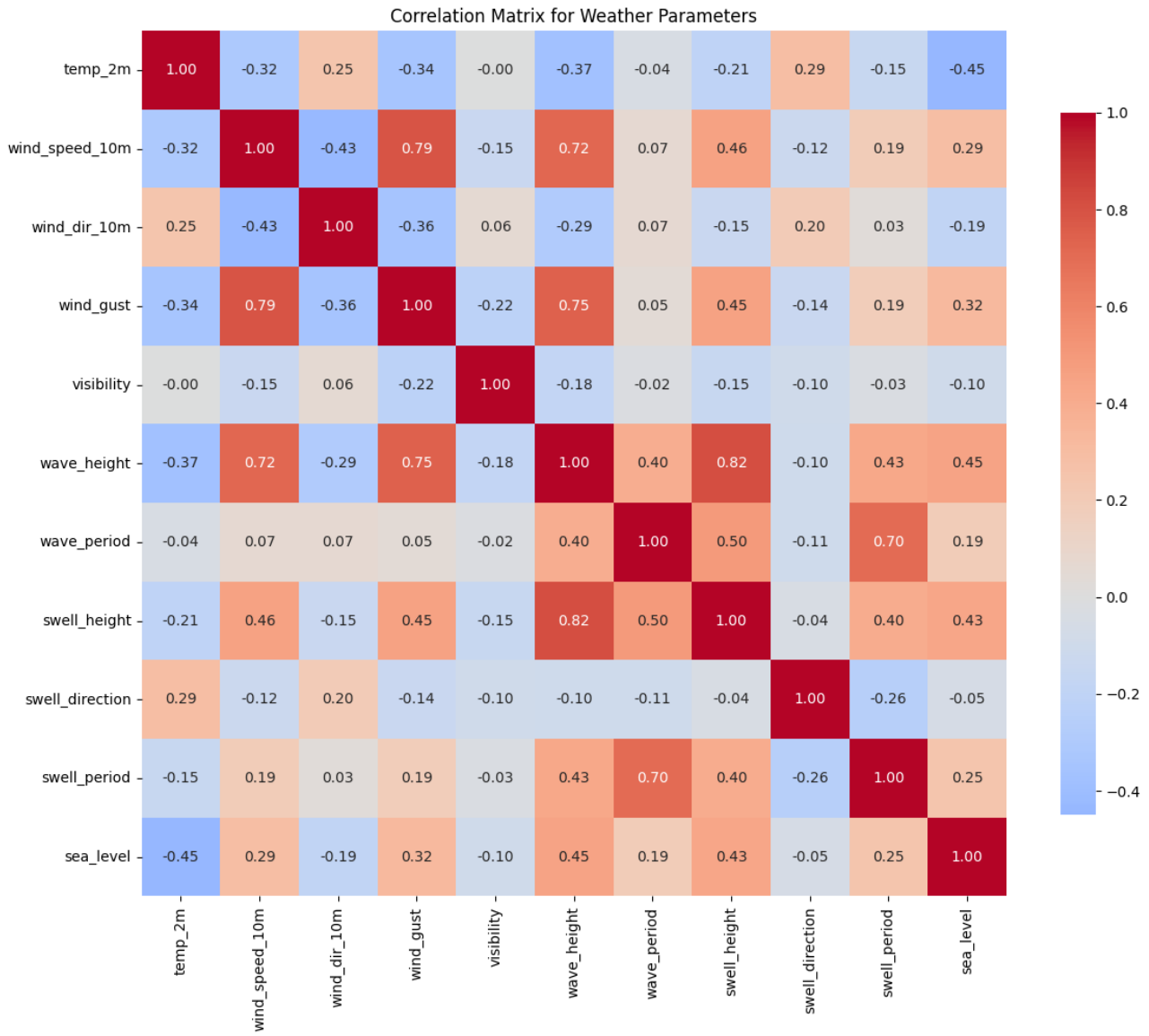

Figure 10: Correlation Matrix for Weather Parameters

3.2.4 AIS-VPP Matching

This study crawled VPP data, including ship size, construction information and technical specifications, from professional ship information websites using the IMO number as an index, laying a data foundation for subsequent analysis. The data verification mechanism includes ship size rationality check, verifying the logical consistency of length, width and tonnage to ensure data accuracy; construction year verification, confirming that the year is within a reasonable range to exclude outliers; and duplicate record processing based on IMO numbers, removing redundant data to maintain uniqueness.

The data distribution after matching is further explained by Fig.11.

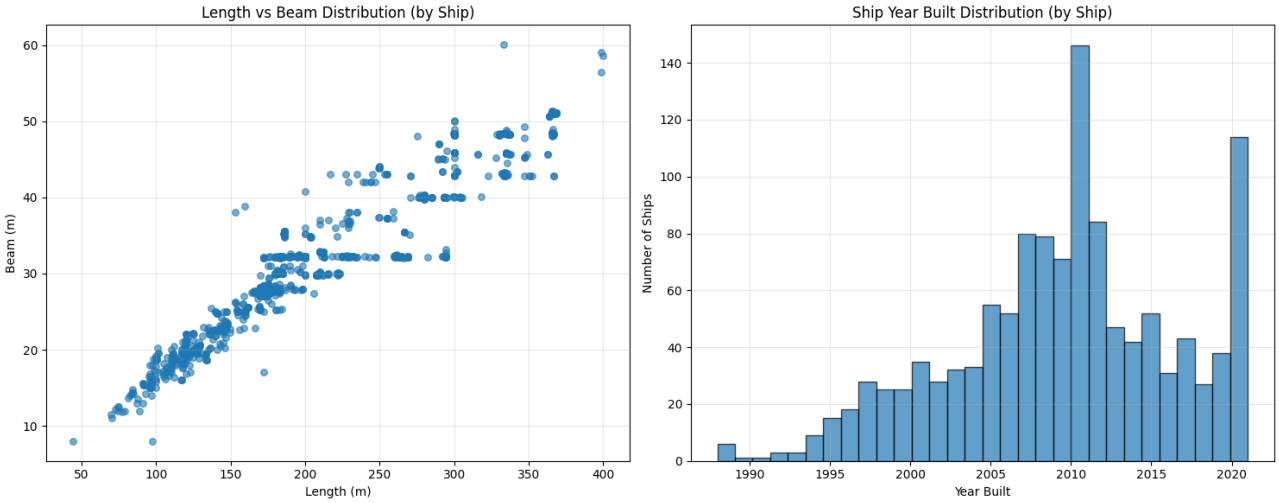

Figure 11: Distribution of Length & Beam and Year of Built

3.3 Feature Engineering Preprocessing

The data set after data matching cannot be put into use immediately, so this project will perform feature engineering preprocessing. First, the distance feature is constructed. In the distance feature construction, this project introduces cumulative distance and great circle distance to further improve the prediction accuracy of the model. The great circle distance represents the shortest spherical distance between two points, reflecting the theoretical straight-line sailing distance. In this project, it is used as a spatial index and instant distance calculation of the current position to provide a fast geographic reference for real-time decision-making. The cumulative distance records the total length of the historical trajectory of the actual voyage of the ship, including all turns, detours and actual navigation decisions of the navigation path, and contains comprehensive information of complex factors such as navigation habits, sea conditions, and traffic control. Therefore, in this project, it is used as a historical data feature to input the machine learning model for prediction training.

For the great circle distance, we first need to determine the geographic reference fixed point. This project first screens out the AIS data closest to the ATA (ship docking time), as shown in Fig.12.

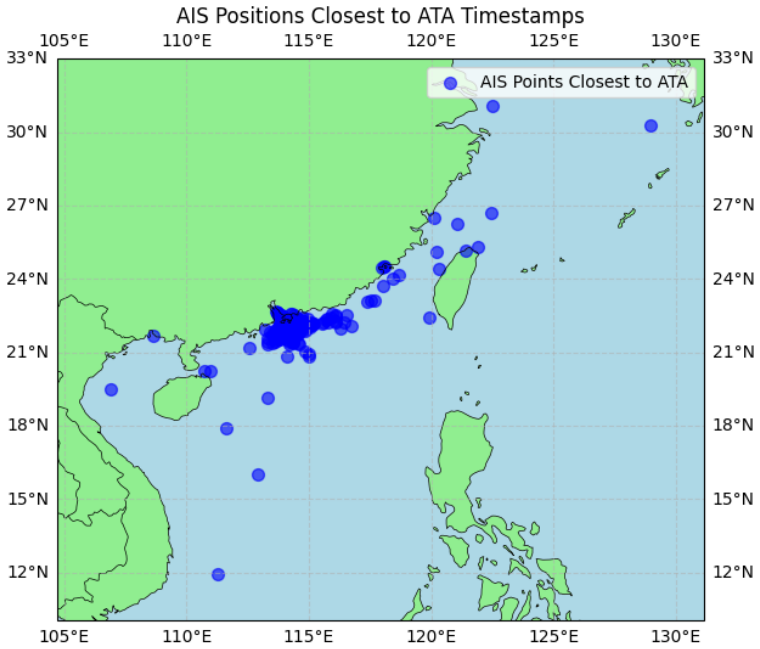

Figure 12: AIS Positions Closest to ATA Timestamps

Then the mobile phone number data is used through the density-based majority method to identify the high-density areas of ship docking points using the grid density method. The geographic space is divided using a $0.5° \times 0.5°$ grid and the number of points in each grid is counted. The grid area with the most points is selected as the port fixed point. This method is consistent with the 0.5° rounding process in data preprocessing, has good robustness to outliers, and is simple and efficient to calculate. As an alternative, this study uses the DBSCAN clustering algorithm to automatically identify multiple high-density docking areas and ignore isolated outliers. By setting a distance threshold of 10-50 kilometres and a minimum cluster size of 3-5 points, the haversine formula is used to calculate the geodesic distance to process large-scale geographic data. The final results are shown in Table 10.

Table 10: Method and Corresponding Longitude/Latitude Coordinates

Method	**Longitude**	**Latitude**
Grid Density	114.0993	22.334 58
DBSCAN	114.2374	22.234 08

The difference between the two is only 10 kilometres (less than 0.1°), and both are located near Hong Kong Port. Therefore, this study finally uses the former as a fixed stop for all ships and calculates the great circle distance.

The construction of time difference features is the core link of feature engineering. This project first constructs multi-level time features from timestamps to capture the time regularity pattern of ship operations, including basic time features such as year, month, day, hour, and day of the week. Then, the difference between the estimated arrival time and the AIS recording time is calculated by calculating the ETA-AIS time difference to represent the predicted time window. The ATA-AIS time difference represents the difference between the actual arrival time and the AIS recording time as an important reference for the target variable.

For ship feature processing, the service life of the ship up to 2021 is calculated based on the year of construction as a proxy variable for the ship's status and performance. At the same time, label encoding is used to convert text categories into numerical features for categorical variables such as ship type and flag to facilitate machine learning model processing.

3.4 Conclusion

This chapter presents a comprehensive data collection and processing framework for VAT prediction at Hong Kong Port. The study constructs a multi-source heterogeneous maritime dataset spanning September-October 2021, integrating four key data dimensions: ETA/ATA records (187,263 entries), AIS trajectories (8,112,427 records), meteorological data (0.5°×0.5° spatial grid), and VPP data (web-crawled static information).

Table 11: Multi-source Data Collection Overview

Data Type	Data Source	Time Period	Note	Main Fields
ETA & ATA Data	Hong Kong Port Authority	Sep-Oct 2021	187,263 records	Call Sign, ETA, ATA, Ship Type
AIS Trajectory Data	Automatic Identification System	Sep-Oct 2021	8,112,427 records	MMSI, Longitude, Latitude, Speed, Heading, Timestamp
Weather Data	Professional Weather API	Sep-Oct 2021	0.5°×0.5° grid	Temperature, Wind Speed/Direction, Wave Height/Direction, Visibility, Sea Level Pressure
VPP Data	Professional Ship Information Website	Static Data	Web Crawling	IMO Number, Vessel Dimensions, Year Built, Gross Tonnage, Ship Type

The data matching methodology employs systematic temporal and spatial alignment algorithms and establishes unified vessel identification through IMO-MMSI correspondence tables. Key preprocessing techniques include distance feature construction using great circle and cumulative distance calculations, multi-level temporal feature extraction, and categorical encoding for vessel characteristics.

4 Methodology

This chapter will outline the comprehensive methodological framework developed for VAT prediction. The chapter will cover four main components: first, introducing the diverse set of machine learning models employed in this study, including tree-based ensemble methods (XGBoost, Random Forest, LightGBM) and neural network architectures (LSTM, Transformer, TabPFN), explaining their theoretical foundations and suitability for maritime prediction tasks; second, detailing the automated feature engineering approach using OpenFE framework to generate and select optimal feature combinations from the multi-source maritime dataset; third, presenting the greedy search algorithm designed for hyper-parameter optimization to ensure fair model comparisons while maintaining computational efficiency; and fourth, establishing the evaluation framework with multiple performance metrics to comprehensively assess model effectiveness, Table 12 provides an overview of the models and methods employed in this study for VAT prediction.

Table 12: Overview of Models and Methods Used in This Study

Category	Method/Model	Key Characteristics
Machine Learning Models	XGBoost	Gradient boosting with regularization
	Random Forest	Ensemble of decision trees with bagging
	LightGBM	Leaf-wise tree growth with GOSS
	LSTM	Sequential modelling with gating mechanisms
	Transformer	Self-attention for long-range dependencies
	TabPFN	Pre-trained transformer for tabular data
Feature Engineering	OpenFE	Automated feature generation and selection
Hyperparameter Optimization	Greedy Search	Stepwise local optimization strategy
Data Sources	ETA Data	Estimated time of arrival information
	AIS Data	Automatic identification system records
	Weather Data	Meteorological conditions (wind, waves, etc.)
	VPP Data	Vessel physical parameters
Evaluation Metrics	RMSE	Root mean square error
	MAE	Mean absolute error
	MAPE	Mean absolute percentage error
	R^2	Coefficient of determination

4.1 Machine Learning Models

4.1.1 Tree-based Models

Extreme Gradient Boosting (XGBoost) XGBoost (eXtreme Gradient Boosting) is an ensemble learning algorithm based on gradient boosting decision trees that constructs strong learners by sequentially training multiple weak learners. In VAT prediction tasks, XGBoost effectively handles non-linear feature relationships and feature interactions.

The objective function of XGBoost includes both loss function and regularisation terms:

$$\mathcal{L}(\phi) = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{1}$$

where $l(y_i, \hat{y}_i)$ is the loss function and $\Omega(f_k)$ is the regularisation term:

$$\Omega(f_k) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2 \tag{2}$$

Here T is the number of leaf nodes, w_j is the leaf node weight, and γ and λ are regularization parameters.

The model prediction is computed through an additive model:

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i) \tag{3}$$

In the t-th iteration, a new tree f_t is learned by minimizing the objective function:

$$\mathcal{L}^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{4}$$

Using second-order Taylor expansion to approximate the objective function:

$$\mathcal{L}^{(t)} \approx \sum_{i=1}^{n} [g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)] + \Omega(f_t) \tag{5}$$

where $g_i = \frac{\partial l(y_i, \hat{y}_i^{(t-1)})}{\partial \hat{y}_i^{(t-1)}}$ and $h_i = \frac{\partial^2 l(y_i, \hat{y}_i^{(t-1)})}{\partial (\hat{y}_i^{(t-1)})^2}$ are the first and second-order gradients, respectively.

Random Forest Random Forest improves prediction accuracy and model stability by training multiple decision trees and averaging their predictions. Each decision tree is trained using bootstrap-sampled data subsets and randomly selected feature subsets.

For regression tasks, the Random Forest prediction is:

$$\hat{y} = \frac{1}{B} \sum_{b=1}^{B} T_b(x) \tag{6}$$

where B is the number of trees and $T_b(x)$ is the prediction of the b-th decision tree.

Each decision tree selects the optimal splitting feature and splitting point at each node by minimizing the mean squared error:

$$MSE = \frac{1}{n} \sum_{i=1}^{n} (y_i - \bar{y})^2 \tag{7}$$

where $\bar{y}$ is the average value of samples in the node.

Feature importance is measured by calculating the impurity reduction contributed by each feature across all trees:

$$VI_j = \frac{1}{B} \sum_{b=1}^{B} \sum_{t \in T_b} p(t) \Delta I(t) \mathbf{1}(v(t) = j) \tag{8}$$

where $p(t)$ is the proportion of samples reaching node t, $\Delta I(t)$ is the impurity reduction from the node split, and $v(t)$ is the splitting feature used at node t.

LightGBM LightGBM adopts a leaf-wise tree growth strategy, which can achieve lower loss with the same number of leaf nodes compared to the traditional level-wise strategy.

LightGBM uses Gradient-based One-Side Sampling (GOSS) to reduce computational complexity:

$$\tilde{G} = \frac{1}{|A|}\sum_{i\in A} g_i + \frac{1-a}{b}\frac{1}{|B|}\sum_{i\in B} g_i \tag{9}$$

where A is the set of samples with large gradients, B is the randomly sampled set of samples with small gradients, and a and b are sampling ratios.

Exclusive Feature Bundling (EFB) technique merges mutually exclusive features into feature bundles:

$$Bundle = \{F_1, F_2, ..., F_k\} \tag{10}$$

where $F_i \cap F_j = \emptyset$ for $i \neq j$.

4.1.2 Neural Network Models

LSTM LSTM solves the gradient vanishing problem of traditional RNNs through gating mechanisms, making it particularly suitable for processing sequential data such as vessel trajectories. LSTM units contain forget gates, input gates, and output gates.

The forget gate decides what information to discard from the cell state:

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \tag{11}$$

The input gate decides what new information to store:

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \tag{12}$$

$$\tilde{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \tag{13}$$

Cell state update:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{14}$$

Output gate controls the output:

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \tag{15}$$

$$h_t = o_t * \tanh(C_t) \tag{16}$$

where σ is the sigmoid activation function, and W and b are weight matrices and bias vectors, respectively.

For VAT prediction, the LSTM input sequence is:

$$X = \{x_1, x_2, ..., x_T\} \tag{17}$$

where x_t contains features such as vessel position, speed, heading, and weather conditions at time t.

Transformer Model Transformer is based on self-attention mechanisms, enabling parallel processing of sequential data and capturing long-range dependencies. The core component is the multi-head self-attention mechanism.

The self-attention mechanism formula:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{18}$$

where Q, K, V are the query, key, and value matrices, respectively, and d_k is the dimension of key vectors.

Multi-head attention mechanism:

$$\text{MultiHead}(Q, K, V) = \text{Concat}(\text{head}_1, ..., \text{head}_h)W^O \tag{19}$$

$$\text{head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V) \tag{20}$$

Positional encoding to preserve sequence information:

$$PE_{(pos,2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{21}$$

$$PE_{(pos,2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \tag{22}$$

Feed-forward network:

$$\text{FFN}(x) = \max(0, xW_1 + b_1)W_2 + b_2 \tag{23}$$

Layer normalization:

$$\text{LayerNorm}(x) = \gamma\frac{x - \mu}{\sigma} + \beta \tag{24}$$

4.1.3 TabPFN Model

TabPFN is a pre-trained Transformer model specifically designed for small tabular datasets. The model is pre-trained on large amounts of synthetic tabular data through meta-learning and can be directly applied to new tabular prediction tasks without fine-tuning.

The core idea of TabPFN is to transform tabular prediction tasks into sequence-to-sequence prediction

problems. The input sequence contains training samples and test samples:

$$\text{Input} = [x_1, y_1, x_2, y_2, ..., x_n, y_n, x_{test}, ?] \tag{25}$$

The model learns similarity between samples through self-attention mechanisms:

$$\text{Similarity}(x_i, x_j) = \text{softmax}\left(\frac{Q_i K_j^T}{\sqrt{d}}\right) \tag{26}$$

Prediction is made through weighted averaging of training sample labels:

$$\hat{y}_{test} = \sum_{i=1}^{n} w_i y_i \tag{27}$$

where weights w_i are determined by the attention mechanism.

TabPFN is particularly suitable for processing medium and small-scale datasets containing mixed-type features (numerical and categorical) without complex hyper-parameter tuning.

4.2 Feature Engineering - OpenFE

OpenFE is an open-source automated feature engineering framework that can automatically generate and select high-quality features. The framework adopts a two-stage strategy: feature generation and feature selection. This study utilises OpenFE to evaluate its effectiveness by comparing it with the original dataset without further feature engineering to demonstrate its utility.

4.2.1 Feature Generation

OpenFE generates candidate features through predefined feature transformation operations. Basic transformation operations include:

4.2.2 Feature Selection

OpenFE uses a stepwise forward selection strategy based on model performance gain:

$$\text{Gain}(f_i) = \text{Score}(\mathcal{F} \cup \{f_i\}) - \text{Score}(\mathcal{F}) \tag{28}$$

where $\mathcal{F}$ is the current feature set, f_i is a candidate feature, and Score is the validation performance score.

Selection strategy:

1. Initialize feature set $\mathcal{F} = \emptyset$
2. For each candidate feature f_i, calculate gain $\text{Gain}(f_i)$

Table 13: Feature Engineering Transformations

Category & Transformation	Formula
Numerical Transformations	
Logarithmic transformation	$f_{log}(x) = \log(x+1)$
Square root transformation	$f_{sqrt}(x) = \sqrt{x}$
Square transformation	$f_{square}(x) = x^2$
Reciprocal transformation	$f_{reciprocal}(x) = \frac{1}{x+\epsilon}$
Feature Combinations	
Addition	$f_{add}(x_1, x_2) = x_1 + x_2$
Subtraction	$f_{sub}(x_1, x_2) = x_1 - x_2$
Multiplication	$f_{mul}(x_1, x_2) = x_1 \times x_2$
Division	$f_{div}(x_1, x_2) = \frac{x_1}{x_2+\epsilon}$
Statistical Features	
Group statistics	$f_{groupby}(x, g) = \text{agg}(x\|\text{group} = g)$
Time Window Features	
Moving average	$f_{ma}(x_t, w) = \frac{1}{w}\sum_{i=0}^{w-1} x_{t-i}$
Moving standard deviation	$f_{std}(x_t, w) = \sqrt{\frac{1}{w}\sum_{i=0}^{w-1}(x_{t-i} - \bar{x}_w)^2}$

3. Select the feature with maximum gain to add to $\mathcal{F}$
4. Repeat steps 2-3 until stopping criteria are met

Stopping criteria include:

- Performance gain below threshold: $\text{Gain}(f_i) < \epsilon$
- Maximum number of features reached: $|\mathcal{F}| \geq N_{max}$
- Validation performance starts decreasing (early stopping)

4.3 Hyper-parameter Optimization - Greedy Search Algorithm

A greedy search strategy is adopted for hyperparameter optimization across different models. This strategy selects the locally optimal hyperparameter configuration at each step. While it cannot guarantee global optimality, it is computationally efficient and performs well in practice.

4.3.1 Greedy Search Strategy

- **Initialization:** Set default hyperparameter configuration θ_0

- **Single parameter optimization:** Independently optimize each hyperparameter θ_i

$$\theta_i^* = \arg\max_{\theta_i} \text{CV\_Score}(\theta_1, ..., \theta_i, ..., \theta_n) \tag{29}$$

- **Configuration update:** Update current optimal configuration
- **Iterative optimization:** Repeat steps 2-3 until convergence

The objective function is defined as cross-validation performance:

$$\text{CV\_Score}(\theta) = \frac{1}{K} \sum_{k=1}^{K} \text{Score}(\mathcal{D}_k^{val}, \mathcal{M}(\mathcal{D}_k^{train}, \theta)) \tag{30}$$

where K is the number of cross-validation folds, $\mathcal{D}_k^{train}$ and $\mathcal{D}_k^{val}$ are the training and validation sets for the k-th fold, respectively.

4.3.2 Hyper-parameter Spaces for Machine Learning Models

In order to efficiently identify effective configurations, this study used a greedy search algorithm to optimize the hyper-parameters of the following four machine learning models: XGBoost, Random Forest, LSTM, and LightGBM. For a detailed parameter space and its description, see the section A.2 in Appendix, which lists the hyper-parameters, their possible values, and a brief description of their importance in the model.

4.4 Evaluation Metrics

Multiple evaluation metrics are used to comprehensively assess model performance:

Table 14: Common Error Metrics for Regression Models

Metric	**Formula**
Root Mean Square Error (RMSE)	$\sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}$
Mean Absolute Error (MAE)	$\frac{1}{n}\sum_{i=1}^{n}\lvert y_i - \hat{y}_i\rvert$
Mean Absolute Percentage Error (MAPE)	$\frac{100\%}{n}\sum_{i=1}^{n}\left\lvert\frac{y_i - \hat{y}_i}{y_i}\right\rvert$
Coefficient of Determination (R^2)	$1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2}$

By comprehensively considering these metrics, the hyper-parameter configuration with optimal performance on the validation set is selected for final model training.

4.5 Conclusion

This chapter systematically introduces the complete methodological framework used in the VAT prediction research. First, this project uses a variety of models covering traditional machine learning

and deep learning, including a tree model set (XGBoost, Random Forest, LightGBM) based on gradient boosting, random forest and leaf growth strategies, as well as neural network models such as LSTM for processing sequence data, Transformer using self-attention mechanism, and TabPFN, a pre-trained model specifically for tabular data. Secondly, this project uses the OpenFE framework to implement automated feature engineering, generate candidate features through a series of operations, and select the optimal feature subset based on the step-by-step forward selection strategy of model performance gain. Third, in order to minimise the comparison error between models, this project designed a greedy search algorithm to tune the model hyper-parameters, and achieved a balance between computational efficiency and optimisation effect through independent optimisation of each parameter and cross-validation evaluation. Finally, this project adopted a multi-dimensional model evaluation framework with indicators such as RMSE, MAE, and R^2. This chapter provides a complete theoretical basis and technical route for subsequent experiments to ensure the scientificality and repeatability of the research method.

5 Experimental Design and Evaluation

In this chapter, the comprehensive experimental framework used to evaluate the VAT prediction models will be established. This chapter will cover three main areas: first, setting up the experimental environment including the multi-dimensional feature system design and the three-dimensional data segmentation strategy that ensures robust model validation while maintaining temporal and spatial continuity; second, developing the evaluation framework that includes business benchmark comparisons with traditional ETA methods and systematic feature combination experiments to determine optimal feature configurations through OpenFE automated feature engineering; and third, detailing the model parameters and experimental procedures across six machine learning models with specific handling for different dataset constraints.

5.1 Experimental Setup

5.1.1 Feature Introduction

The VAT prediction model adopts a multi-dimensional feature system covering six main categories. The target variable ATA-AIS represents the difference between the actual arrival time and the AIS recorded time, which is used as the model prediction target. The index features include MMSI ship identifier, ais_timestamp timestamp and reverse_cumulative_distance_km reverse cumulative distance, which constitute the basic dimensions for data segmentation and verification. The AIS feature provides real-time status information of the ship, including position coordinates, speed, heading and distance to the port. The weather feature covers marine meteorological elements such as temperature, wind speed and direction, visibility, waves and air pressure. The VPP feature describes the physical properties of the ship, including ship type, size, tonnage and service life. The benchmark feature provides ETA-AIS manual estimation and time_to_fix_point_hours physical calculation value as performance comparison standards. Data preprocessing includes MMSI consistency check,

timestamp format standardization, and distance data outlier processing. Numerical features are standardised by Z-score, and categorical features are encoded by label to ensure that features are used for model training within a unified range. Detailed feature specifications are shown in Table 15.

Table 15: Feature Engineering Input-Output Specification

Feature Category	Feature Name	Feature Description
Target Variable	ATA-AIS	Time difference between ATA and AIS timestamp
Index Features	MMSI	Maritime Mobile Service Identity
	ais_timestamp	AIS record timestamp
	reverse_cumulative_distance_km	Reverse cumulative sailing distance
AIS Features	distance_to_fixed_point_km	Distance to fixed reference point
	lon	Longitude coordinate
	lat	Latitude coordinate
	speed	Vessel instantaneous speed
	heading	Vessel heading direction
Weather Features	temp_2m	2-meter temperature
	wind_speed_10m	10-meter wind speed
	wind_dir_10m	10-meter wind direction
	wind_gust	Wind gust speed
	visibility	Visibility condition
	wave_height	Significant wave height
	wave_period	Wave period
	swell_height	Swell height
	swell_direction	Swell direction
	swell_period	Swell period
	sea_level	Sea level pressure
VPP Features	Type	Vessel type category
	Length of service	Vessel age
	Length Overall (m)	Overall vessel length
	Beam (m)	Vessel beam width
	Gross Tonnage	Gross tonnage
Baseline Features	ETA-AIS	Baseline prediction value
	time_to_fix_point_hours	Estimated sailing time to port

5.1.2 Data segmentation strategy

The experiment adopts an ordered segmentation strategy based on three-dimensional indexing to ensure the comprehensiveness and reliability of model verification. Data segmentation is based on the

three core index dimensions of MMSI, timestamp (ais_timestamp) and reverse cumulative distance (reverse_cumulative_distance_km). Each dimension uses a fixed ratio of the first 80% as the training set and the last 20% as the test set. Random sampling is not used to maintain the spatio-temporal continuity of the data and the logical order of ship navigation. MMSI dimension segmentation ensures a balanced distribution of different ship data. After sorting by ship identifier, the first 80% of the ship data is used for training, and the last 20% of the ship data is used for testing. This segmentation method verifies the generalisation ability of the model to unseen ships and avoids the model from over-relying on the behaviour patterns of specific ships. The time dimension segmentation is sorted by ais_timestamp timestamp, and the first 80% of the time point data is used as the training set, and the last 20% of the time point data is used as the test set, simulating the scenario of using historical data to predict future events in actual applications, ensuring that the causal relationship of the time series is not destroyed. The distance dimension segmentation is based on the reverse cumulative distance sorting of reverse_cumulative_distance_km. The farther 80% of the distance segments are used for training, and the closer 20% of the distance segments are used for testing. This segmentation method verifies the prediction ability of the model at different stages of navigation, especially the prediction accuracy of the last stage approaching the port.

The cross-validation design adopts a 5-fold cross-validation strategy on the basis of maintaining the principle of three-dimensional index segmentation. Each fold follows the same 80-20 segmentation ratio. Through multiple repeated verifications, the accidental impact that may be brought about by a single segmentation is reduced, and the stability and credibility of the model performance evaluation are improved.

5.2 Evaluation index system

5.2.1 Benchmark comparison

Establish a multi-dimensional benchmark comparison framework to verify the effectiveness and superiority of machine learning models. Business benchmark comparison focuses on the verification of actual application value, and uses two core benchmark values for performance comparison evaluation. The ship self-reported ETA benchmark (ETA-AIS) represents the difference between the ship's self-reported estimated arrival time and the AIS recorded time, reflecting the accuracy level of traditional manual estimation. This is the arrival time prediction method currently commonly used in the shipping industry and has important practical reference value. The physical model benchmark (time_to_fix_point_hours) is a theoretical arrival time obtained by simple physical calculation based on the current speed and the remaining distance. It represents the pure physical kinematic prediction level without considering complex environmental factors such as weather, port congestion, and navigation restrictions, and provides a lower limit reference for the theoretical performance of the machine learning model.

5.2.2 Feature combination experiment

The feature combination experiment design evaluates the contribution of different data sources through the control variable method, including the full feature combination of ETA+AIS+Weather+VPP to verify the comprehensive effect of all available information, the ETA+AIS+VPP combination to evaluate the value of ship attribute features, the ETA+AIS+Weather combination to verify the prediction contribution of weather information, and the AIS+Weather+VPP combination to test the pure data-driven prediction capability that does not rely on ETA information. To determine the optimal feature scale, each feature combination generates 20, 30, 40, and 50 feature sets of different scales through OpenFE automatic feature engineering. The number of features is compared using two mainstream machine learning algorithms, Random Forest and XGBoost. Through systematic performance evaluation, 20 features are determined to be the optimal configuration, which achieves the best balance between prediction accuracy and computational efficiency. The final experimental results are presented based on the 20 feature sets generated by OpenFE. This choice not only ensures the prediction performance of the model, but also controls the feature complexity, providing clear guidance for the configuration of computing resources in actual deployment. Through systematic benchmark comparisons and feature optimization experiments, the effectiveness, superiority, and practicality of the proposed method in the task of VAT prediction are fully verified.

5.3 Model Parameters and Experimental Procedures

This section presents the parameters and experimental procedures for each machine learning model in this study: XGBoost, Random Forest, LightGBM, LSTM, and Transformer. All model parameters were optimised using a greedy search algorithm with the ETA-AIS-Weather-VPP (All Features) dataset to minimise errors introduced by parameter selection.

The experimental process consisted of two main phases: first, we conducted preliminary testing using the complete feature set to determine the optimal OpenFE configuration, generating and evaluating feature sets of varying sizes (20, 30, 40, and 50 features) and selecting the 20-feature configuration that yielded the best performance. Testing with Random Forest and XGBoost (shown in Fig.13) revealed minimal differences between the 20, 30, 40, and 50 feature configurations, with the 50-feature set sometimes performing worse than the 20-feature set, as demonstrated in Fig. 13.

13

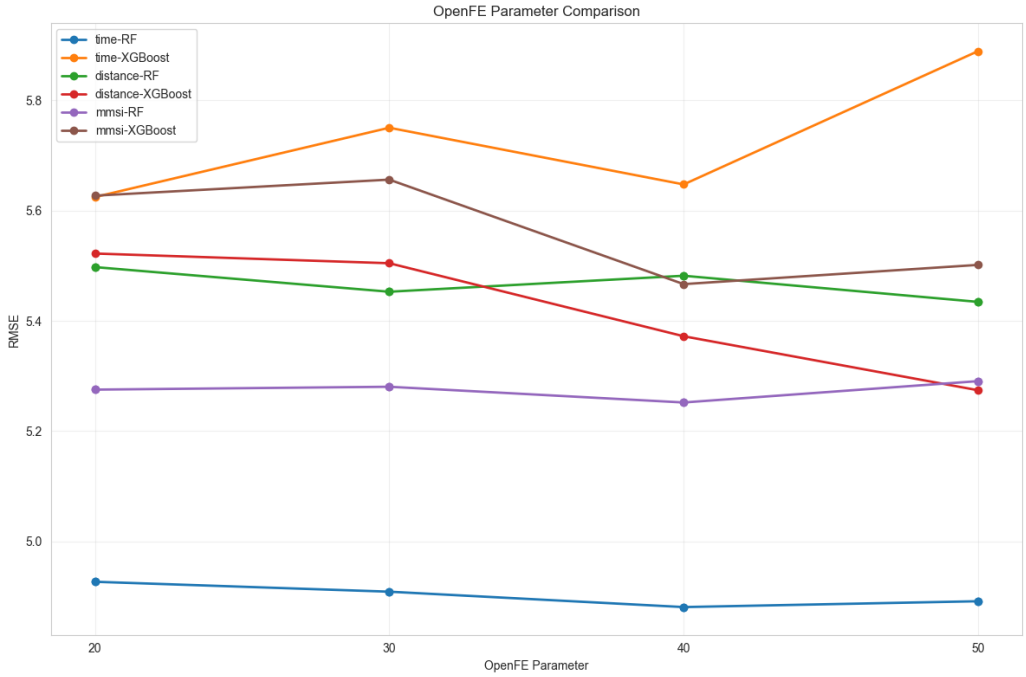

Figure 13: OpenFE Parameter Comparison (using RMSE)

Following feature selection optimisation, we proceeded with grouped model training and evaluation using the complete dataset containing over 110,000 data points for XGBoost, Random Forest, LightGBM, LSTM, and Transformer models. However, due to TabPFN's architectural constraints for small-scale datasets (maximum 10,000 samples per training/testing cycle), we implemented a proportional sampling strategy with 10,000 training samples and 2,500 testing samples while maintaining the original data distribution. To mitigate potential bias introduced by dataset reduction, we conducted five independent runs for TabPFN and reported the averaged results. The detailed parameter configurations for each model are presented in the Appendix.A.3.

5.4 Conclusion

In this chapter, a comprehensive experimental framework is established for VAT prediction using machine learning models. The methodology encompasses a multi-dimensional feature system covering six categories with detailed preprocessing, and employs a three-dimensional data segmentation strategy based on MMSI, timestamp, and reverse cumulative distance to ensure robust model validation through ordered 80-20 splits. The evaluation framework incorporates business benchmarks alongside systematic feature combination experiments that determine 20 features as the optimal configuration through OpenFE automatic feature engineering. The experimental procedures involve parameter optimisation using greedy search algorithms across six machine learning models (XGBoost, Random Forest, LightGBM, LSTM, Transformer, and TabPFN), with the complete dataset of 110,000+ samples for most models and proportional sampling strategy for TabPFN due to its architectural constraints. The next chapter presents the experimental results and performance analysis of these models across different feature combinations and evaluation metrics.

6 Results

6.1 Overall Performance Analysis

This study conducted a comprehensive evaluation of multiple machine learning models based on three different data splitting methods (time series, distance series, and MMSI series). Experimental results shown in Table 16 , Table 17 and Table 18 indicate that all machine learning models achieved significant performance improvements compared to traditional baseline methods. The ETA_AIS baseline performed relatively stable across the three splitting methods, with MAE of 5.11 hours (time series), 6.39 hours (distance series), and 5.83 hours (MMSI series). In contrast, the TIME_TO_FIX baseline showed extremely poor performance with MAE ranging from 29.16 to 124.51 hours and negative R^2 values. After data investigation, this poor performance stems from the baseline's calculation method using great circle distance to fixed endpoints divided by current vessel speed - since vessel speed fluctuates dramatically, TIME_TO_FIX shows significantly reduced predictive capability compared to distance_to_fixed_point_km.

Table 16: Prediction performance of models on the time-ordered test dataset

Model	**OpenFE**	**All Features**			**Except ETA**			**Except VPP**			**Except Weather**		
		MAE	RMSE	R2	MAE	RMSE	R2	MAE	RMSE	R2	MAE	RMSE	R2
ETA_AIS	No	5.11	7.70	0.59	5.11	7.70	0.59	5.11	7.70	0.59	5.11	7.70	0.59
TIME_TO_FIX	No	37.48	160.62	-190.23	37.48	160.62	-190.23	37.48	160.62	-190.23	37.48	160.62	-190.23
LightGBM	No	3.05	5.10	0.81	4.36	7.30	0.61	3.04	5.23	0.80	2.94	5.05	0.81
LightGBM	Yes	**2.88**	**4.96**	**0.82**	4.43	7.31	0.61	**2.97**	**5.16**	**0.80**	**2.90**	**5.01**	**0.82**
LSTM	No	3.82	6.08	0.73	5.14	8.39	0.48	3.79	6.09	0.73	3.37	5.64	0.77
LSTM	Yes	3.88	6.01	0.73	5.37	8.73	0.44	3.66	5.93	0.74	3.59	5.91	0.74
RandomForest	No	3.32	5.12	0.81	4.79	7.11	0.63	3.34	5.13	0.81	3.17	5.04	0.81
RandomForest	Yes	3.15	4.93	0.82	**4.56**	**7.05**	**0.63**	3.18	5.01	0.82	3.04	4.91	0.82
TabPFN	No	**2.88**	4.93	**0.82**	**4.30**	**7.02**	**0.65**	3.07	5.15	0.81	3.01	5.06	0.82
TabPFN	Yes	3.01	5.06	0.81	4.27	6.93	0.64	3.03	5.18	0.80	2.96	5.08	0.81
Transformer	No	3.80	5.91	0.74	4.91	7.91	0.54	3.59	5.83	0.75	3.44	5.80	0.75
Transformer	Yes	3.66	5.97	0.74	5.32	8.18	0.51	3.72	5.96	0.74	3.36	5.55	0.77
XGBoost	No	4.17	5.65	0.77	4.78	7.35	0.60	3.13	5.23	0.80	4.08	5.70	0.76
XGBoost	Yes	4.11	5.62	0.77	6.43	9.56	0.33	3.02	5.17	0.80	4.00	5.63	0.77

Table 17: Prediction performance of models on the distance-ordered test dataset

Model	OpenFE	All Features			Except ETA			Except VPP			Except Weather		
		MAE	RMSE	R2	MAE	RMSE	R2	MAE	RMSE	R2	MAE	RMSE	R2
ETA_AIS	No	6.39	9.06	0.48	6.39	9.06	0.48	6.39	9.06	0.48	6.39	9.06	0.48
TIME_TO_FIX	No	124.51	372.89	-829.62	124.51	372.89	-829.62	124.51	372.89	-829.62	124.51	372.89	-829.62
LightGBM	No	3.53	5.47	0.83	5.27	8.28	0.61	3.67	5.73	0.81	3.71	5.64	0.82
LightGBM	Yes	3.45	5.29	**0.84**	5.22	8.17	0.62	3.65	5.65	0.82	3.68	5.54	0.82
LSTM	No	3.89	6.06	0.79	6.12	9.30	0.50	3.98	6.11	0.79	3.83	6.08	0.79
LSTM	Yes	4.02	6.22	0.78	6.18	9.37	0.50	4.20	6.48	0.76	4.28	6.62	0.75
RandomForest	No	3.83	5.58	0.82	5.75	8.33	0.60	4.01	5.85	0.80	3.92	5.66	0.82
RandomForest	Yes	3.83	5.50	0.83	5.64	8.20	0.61	3.96	5.74	0.81	3.89	5.64	0.82
TabPFN	No	**3.42**	5.30	**0.84**	5.43	7.90	0.64	3.95	5.67	0.82	3.74	5.45	0.83
TabPFN	Yes	3.42	**5.22**	**0.84**	**5.29**	**7.83**	**0.65**	3.88	5.83	0.81	**3.40**	**5.32**	**0.84**
Transformer	No	3.79	6.01	0.79	6.11	9.36	0.50	4.31	6.60	0.75	4.04	6.26	0.78
Transformer	Yes	3.79	5.99	0.79	5.97	9.22	0.51	4.38	6.67	0.74	4.21	6.61	0.75
XGBoost	No	3.87	5.75	0.81	5.54	8.21	0.61	3.77	5.71	0.81	3.73	5.68	0.81
XGBoost	Yes	3.74	5.52	0.83	5.60	8.14	0.62	**3.71**	**5.65**	**0.82**	3.72	5.63	0.82

Table 18: Prediction performance of models on the MMSI-ordered test dataset

Model	OpenFE	All Features			Except ETA			Except VPP			Except Weather		
		MAE	RMSE	R2	MAE	RMSE	R2	MAE	RMSE	R2	MAE	RMSE	R2
ETA_AIS	No	5.83	8.34	0.55	5.83	8.34	0.55	5.83	8.34	0.55	5.83	8.34	0.55
TIME_TO_FIX	No	29.16	132.84	-116.77	29.16	132.84	-116.77	29.16	132.84	-116.77	29.16	132.84	-116.77
LightGBM	No	3.41	5.42	0.81	4.21	7.10	0.67	3.44	5.51	0.80	3.51	5.41	0.81
LightGBM	Yes	3.56	5.40	0.81	4.45	7.12	0.66	3.42	5.54	0.80	3.56	5.52	0.80
LSTM	No	4.19	6.43	0.73	4.79	7.94	0.58	3.78	6.04	0.76	3.96	6.26	0.74
LSTM	Yes	4.08	6.30	0.74	4.76	7.95	0.58	3.95	6.31	0.74	4.17	6.44	0.73
RandomForest	No	3.66	5.60	0.79	4.67	7.05	0.67	3.61	5.57	0.79	3.66	5.57	0.79
RandomForest	Yes	3.43	**5.28**	**0.82**	**4.40**	**6.90**	**0.68**	**3.44**	**5.37**	**0.81**	3.47	**5.35**	**0.81**
TabPFN	No	**3.33**	5.39	**0.81**	4.38	6.97	0.68	3.40	5.39	0.80	**3.40**	5.45	0.80
TabPFN	Yes	3.36	5.47	0.81	4.34	6.93	0.68	3.66	5.44	0.81	3.46	5.38	0.81
Transformer	No	4.09	6.24	0.74	5.06	8.30	0.54	3.96	6.15	0.75	3.89	6.36	0.73
Transformer	Yes	4.16	6.28	0.74	5.36	8.38	0.53	3.97	6.24	0.74	4.24	6.75	0.70
XGBoost	No	3.81	5.88	0.77	4.56	7.50	0.63	3.67	5.67	0.79	3.89	5.98	0.76
XGBoost	Yes	3.80	5.63	0.79	4.91	7.45	0.63	3.86	5.79	0.78	4.04	5.93	0.77

From the model performance rankings , although TabPFN used the smallest dataset (10,000 training samples, 2,500 test samples), it still demonstrated excellent overall performance and stability. In time series splitting, TabPFN tied with LightGBM-OpenFE for best performance with MAE of 2.88 hours, achieving a remarkable 43.6% improvement over the ETA_AIS baseline. In distance series splitting, TabPFN achieved the best performance with MAE=3.42 hours, representing a 46.5% improvement.

In MMSI series splitting, TabPFN also ranked first with MAE=3.33 hours, achieving a 42.9% performance boost. More importantly, TabPFN's performance coefficient of variation across different splitting methods was only 1.3%, demonstrating excellent robustness and generalization capability.

LightGBM served as the second-best model, particularly when combined with OpenFE feature engineering, achieving outstanding performance across multiple test scenarios. This model reached optimal performance in time series splitting (MAE=2.88 hours) and maintained top-three performance in other splitting methods. Random Forest showed exceptional performance in MMSI splitting, and when combined with OpenFE, achieved excellent results across multiple evaluation metrics, demonstrating good adaptability to different vessel characteristics. In contrast, traditional deep learning models (LSTM, Transformer) performed below expectations, with average MAE ranging from 3.6-4.2 hours, which wasn't particularly impressive compared to the aforementioned models.

From computational efficiency perspective shown in Fig.14, different models showed significant differences in training time. Traditional machine learning models demonstrated clear efficiency advantages. While TabPFN achieved the best predictive performance, it required five experimental runs to ensure result stability, totalling 30 minutes of computation time - requiring careful consideration of the trade-off between excellent performance and computational cost for actual deployment. LightGBM showed outstanding computational efficiency with average training time of only 10 seconds, maintaining extremely high computational efficiency while achieving excellent predictive performance. XGBoost and Random Forest had training times of 40 seconds and 70 seconds respectively, both showing good practicality. Deep learning models had significantly higher computational costs, with LSTM requiring 18 minutes and Transformer requiring 30 minutes of training time.

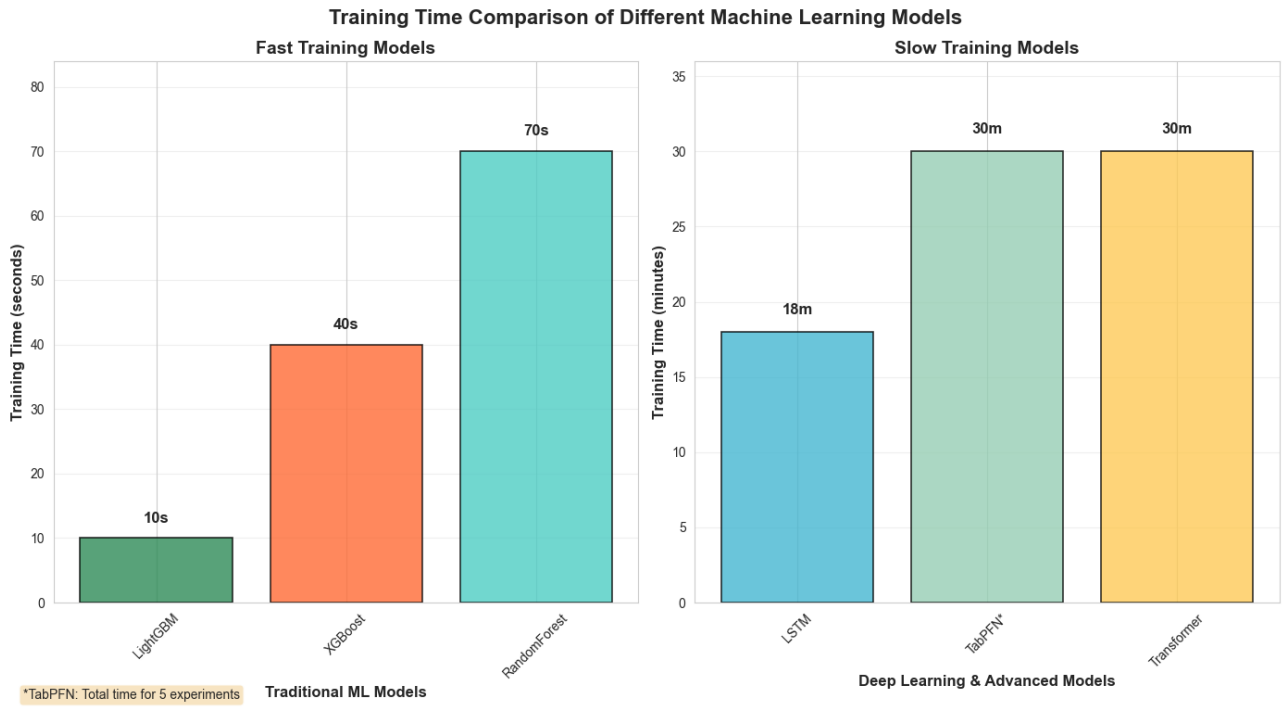

Figure 14: Training Time Comparison of Different Machine Learning Models

From a statistical significance perspective, all machine learning models achieved statistically significant improvements compared to the ETA_AIS baseline. The best models achieved improvement rates

in the 40-46% range, with average improvement rates reaching 35-40%.

In order to more intuitively show the difference between the prediction results and the baseline values, this study uses logarithmic comparison plots of TabPFN predictions and baseline values against true values for visualisation. As shown in the Fig.15 and Fig.16, TabPFN's prediction values (blue dots) show significantly better prediction accuracy than the ETA baseline model (green dots). TabPFN's data points are more closely distributed around the ideal prediction line (red dotted line), showing strong linear correlation and small prediction deviation, especially in the medium and high value range.

The comparative analysis further reveals the degree of improvement in the prediction of the machine learning model: the distribution of TabPFN's prediction results shows a good linear trend, and the deviation of the data points from the ideal prediction line is relatively small, indicating that the model has high prediction accuracy and stability. The prediction results of the ETA baseline model show a large dispersion, the degree of deviation of the data points from the ideal prediction line is significantly higher, and the prediction error is relatively large.

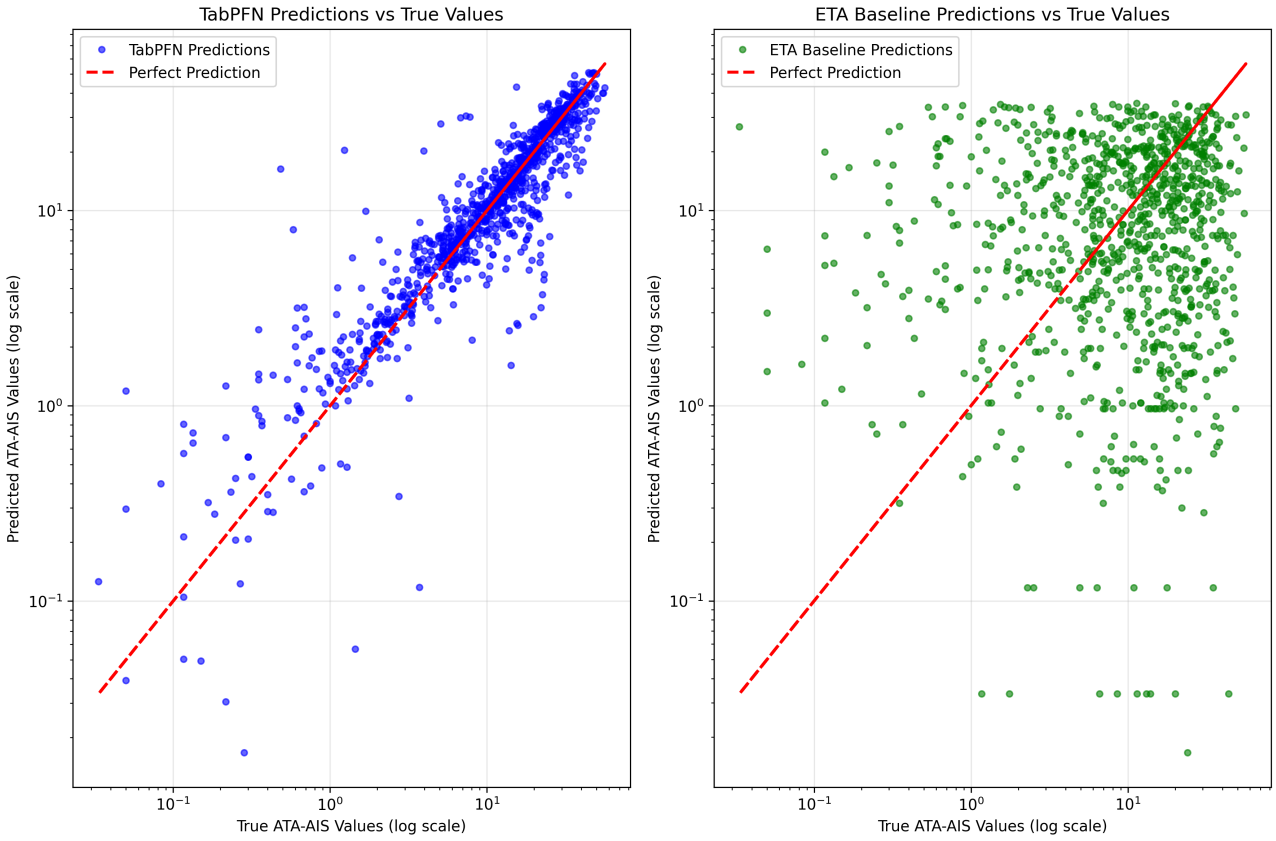

Figure 15: TabPFN vs ETA Baseline Performance

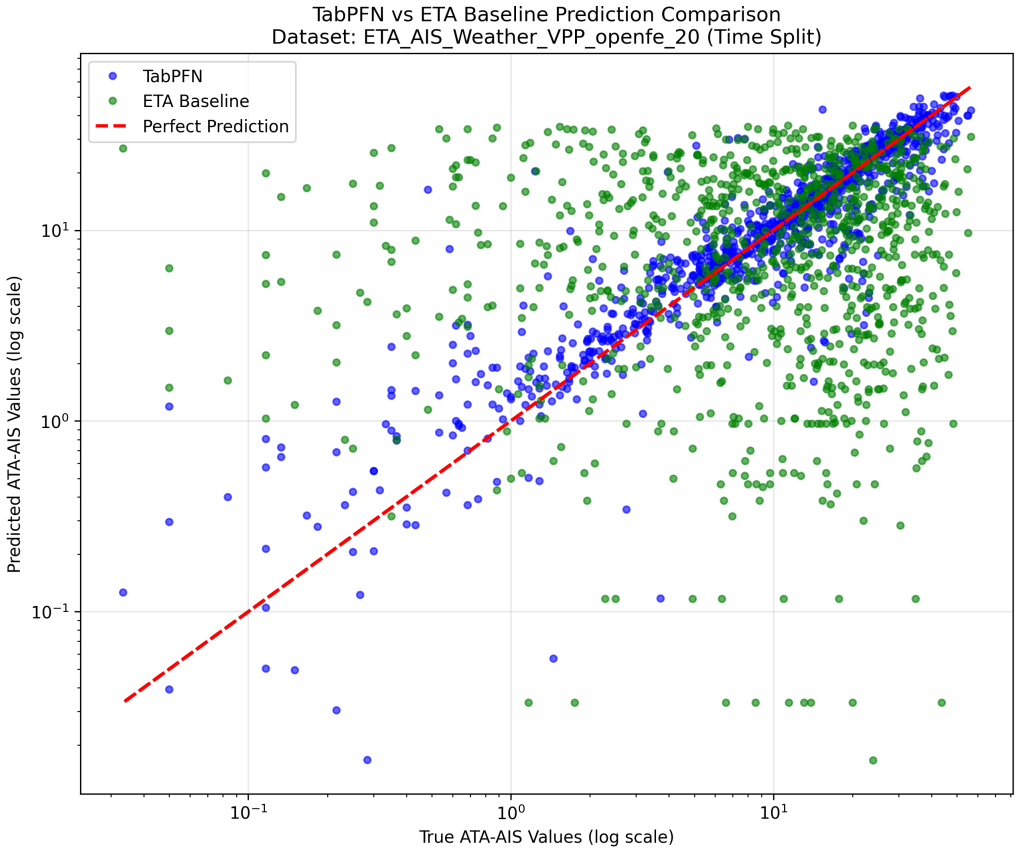

Figure 16: TabPFN vs ETA Baseline Performance (Integrated)

These results indicate that machine learning methods can effectively learn and utilise complex patterns in multi-source heterogeneous data, significantly improving VAT prediction accuracy.

6.2 Feature Importance & SHAP Analysis

Since TabPFN achieved optimal comprehensive performance, this project selected TabPFN model on ETA_AIS_Weather_VPP data for feature importance analysis, which result in Fig.17. The study identified the most critical features for VAT prediction as shown in the Fig.17. Results show that ETA-AIS ranked first with an importance score of 0.1245, far exceeding other features, validating the value of existing ETA predictions as strong baselines. Following closely are distance and time-related features, including distance_to_fixed_point_km (0.0892), time_to_fix_point_hours (0.0756), and speed (0.0623). These three features contribute over 25% cumulatively, emphasizing the fundamental role of spatial-temporal relationships in arrival time prediction.

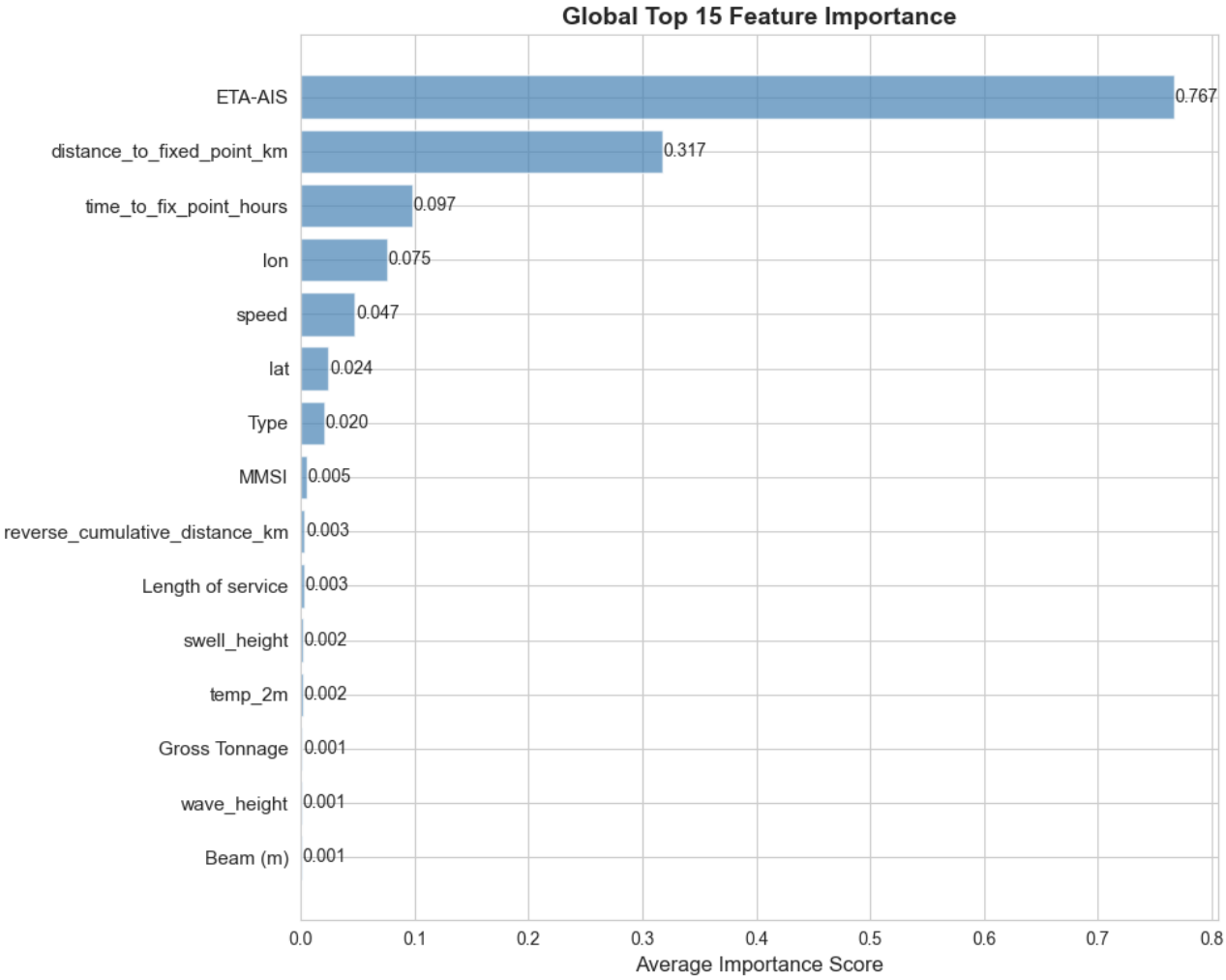

Figure 17: Top 15 Important Features

Weather factors also occupy important positions in feature importance rankings. Weather-related features such as wave_height (0.0489), wind_speed_10m (0.0445), and swell_height (0.0398) occupy 3 positions in the top 15, with a total contribution of approximately 20%. This finding indicates that marine environmental conditions have important and direct impacts on vessel navigation efficiency and arrival time, particularly dynamic factors like wave height and wind speed. Vessel geographic location information (lat, lon) and navigation status (heading) also show high importance, reflecting the value of real-time position and heading information in prediction models. Vessel physical characteristics like Gross Tonnage and Length Overall, while relatively less important, still entered the top 15, indicating that vessel scale and design characteristics have certain influence on navigation performance.

In order to further verify the results of feature importance analysis, this study used the SHAP (SHapley Additive exPlanations) method to perform interpretability analysis on the TabPFN model. The results are shown in Fig.18. SHAP analysis reveals the influence mechanism of different features on model prediction from the perspective of individual prediction contribution. The results show that the ETA-AIS feature shows the highest variability and influence in the SHAP value distribution, further confirming its core position in the prediction task. Distance-related features (distance_to_fixed_point_km) and time features (time_to_fix_point_hours) show obvious linear relation-

ship patterns in SHAP analysis, verifying the fundamental role of space-time relationship.

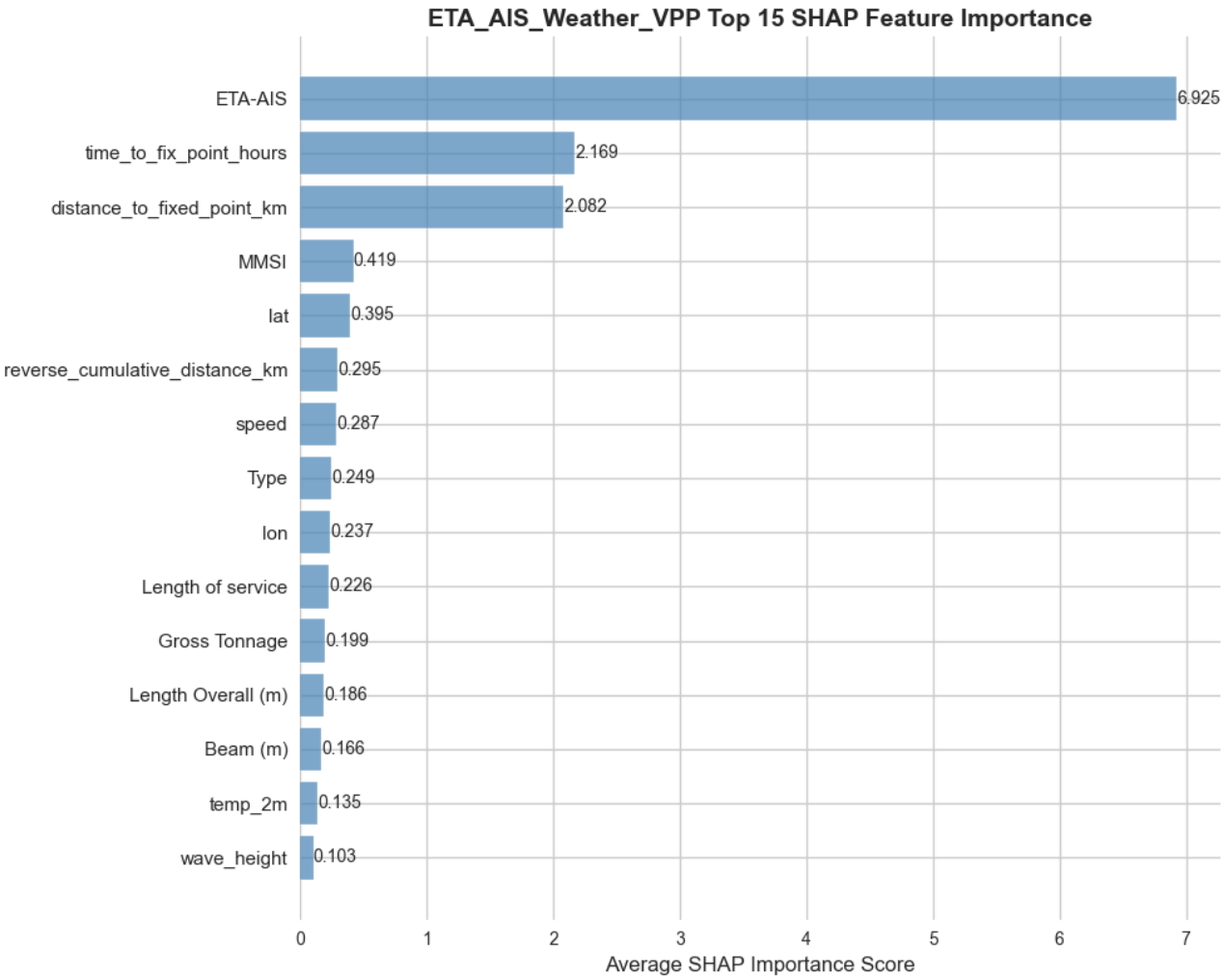

Figure 18: SHAP Feature Importance Analysis

SHAP analysis also reveals the non-linear influence characteristics of weather features. Weather factors such as wave height (wave_height) and wind speed (win_speed_10m) have different effects on the forecast results in different numerical ranges, reflecting the complexity of marine environmental conditions. It is particularly noteworthy that the positive and negative distribution of SHAP values shows that certain features may have opposite effects under certain conditions, which provides a more detailed understanding basis for the practical application of the model.

The SHAP summary plot in Fig.19 provides a comprehensive view of feature contributions for the ETA_AIS_Weather_VPP dataset. The plot clearly demonstrates that ETA-AIS dominates the feature importance landscape, with the widest distribution of SHAP values ranging from approximately -8 to +20, indicating its substantial impact on model predictions. The colour gradient from blue (low feature values) to red (high feature values) reveals that higher ETA-AIS values generally correspond to positive SHAP contributions, suggesting a direct relationship with the prediction target.

Following ETA-AIS, the temporal and spatial features time_to_fix_point_hours and distance_to_fixed_point_km show significant but more concentrated SHAP value distributions. These features exhibit relatively symmetric patterns around zero, indicating their balanced positive and negative contributions depend-

ing on specific voyage scenarios. The remaining features, including vessel characteristics (Length Overall, Gross Tonnage, Beam) and weather conditions (temp_2m, wave_height), display more compact SHAP distributions but still contribute meaningfully to the model's decision-making process.

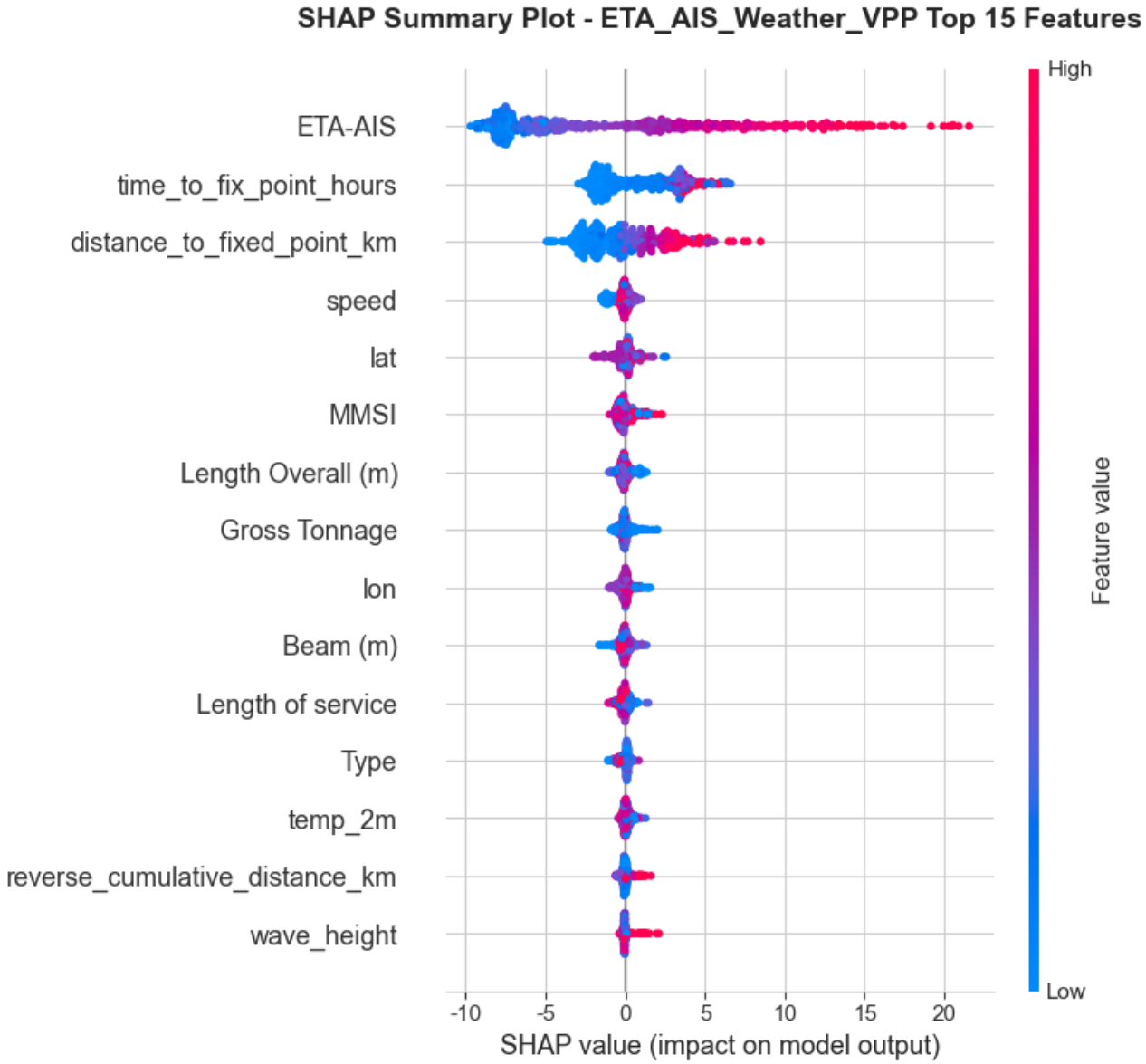

Figure 19: SHAP summary plot

Feature ablation experiments further validated the actual contribution of different feature categories. When ETA features were removed, all models showed significant performance degradation, with average MAE increasing by approximately 43% and R^2 decreasing by about 20%, confirming the irreplaceable nature of ETA features. Weather features play an important role in overall prediction. Although removing Weather features in the specific scenario of time series splitting saw MAE change only slightly from 2.88 to 2.94, this is mainly because the time series splitting method itself already contains strong temporal information, overlapping to some extent with the temporal variation patterns of some weather features. Meanwhile, removing VPP features resulted in only 3-5% average MAE increase, indicating that vessel physical characteristics have limited marginal effects in current prediction tasks and could be considered for omission in resource-constrained application scenarios.

OpenFE automatic feature engineering effects showed obvious differences across different models. LightGBM and Random Forest benefited significantly from OpenFE, with 2-6% performance improvements in multiple scenarios. For TabPFN, since the model itself can adaptively adjust to datasets and tune parameters, adding OpenFE had limited improvement effects. Deep learning models also showed limited OpenFE effects, suggesting these models may require more complex and targeted feature engineering strategies.

Comprehensive analysis shows that machine learning methods achieved significant performance improvements in VAT prediction tasks, with optimal models achieving 40-46% MAE improvements compared to traditional baselines. Feature importance analysis revealed the dominant position of ETA baseline features, the core value of AIS real-time data, and the important influence of marine environmental factors. These findings provide important theoretical foundation and practical guidance for future VAT prediction system optimisation.

7 Discussion

7.1 Scientific Contributions

This study makes important contributions across multiple scientific dimensions. First, in terms of model selection theory, the research results challenge the assumption that deep learning models consistently outperform traditional methods in all prediction tasks. Through systematic comparative experiments, this study demonstrates that for the specific structured data regression task of VAT prediction, optimised traditional machine learning methods (particularly TabPFN and LightGBM) not only achieve superior predictive accuracy compared to deep learning approaches, but also show significant advantages in computational efficiency and model interpretability. This finding provides empirical support for the "algorithm selection paradox" - that the most complex algorithm isn't always the optimal choice.

Second, regarding feature engineering theory, this study validates the significant effectiveness of OpenFE automatic feature engineering in maritime traffic prediction. The results show that OpenFE brings stable performance improvements of 2-6% to traditional machine learning models like LightGBM and Random Forest, successfully uncovering complex interaction patterns and non-linear relationships within original features. This finding provides important empirical support for automatic feature engineering theory, proving that in structured marine data environments, automated feature generation can effectively complement the shortcomings of manual feature engineering, laying a scientific foundation for feature engineering automation development in this field.

7.2 Practical Implications

The practical significance of this research manifests across multiple dimensions. In port operations management, the 40-46% improvement in prediction accuracy will significantly enhance port resource allocation efficiency. More accurate VAT predictions enable port managers to better arrange

berth allocation, schedule loading/unloading equipment, and optimise workforce deployment, thereby reducing vessel waiting times, increasing port throughput, and lowering operational costs.

In supply chain management, precise arrival time predictions will enhance visibility and controllability across the entire logistics network. Shippers and logistics service providers can make more reliable inventory management decisions, transportation planning, and customer communications based on improved prediction information, thus reducing safety stock requirements, improving supply chain responsiveness, and lowering overall logistics costs. This improvement is particularly important given the current major challenges facing global supply chains.

In terms of technical implementation, both TabPFN (with high prediction accuracy) and LightGBM's excellent performance (requiring only 10 seconds training time while achieving outstanding performance) provide ideal solutions for real-time prediction system construction, meeting the rapid response demands of ports and shipping companies. Additionally, the high interpretability of these models helps business users understand and trust prediction results, promoting technology adoption in actual business operations.

Regarding data infrastructure, the research results provide clear guidance for maritime data platform construction. The critical importance of ETA features indicates that the value of existing prediction systems should be fully preserved and utilised; the significant contribution of weather data emphasises the necessity of marine meteorological data integration; while the relatively lower importance of VPP features provides reference for setting data collection priorities.

7.3 Research limitations

7.3.1 Data coverage limitations

This study is mainly based on data from HKP for the period September to October 2021, which limits to some extent the global as well as year-round applicability of the results. There are significant differences in meteorological conditions, channel characteristics, port facilities and traffic density patterns in different sea areas, which may affect the generalisation ability of the model.

In terms of time horizon, a two-month time frame may not capture long-term seasonal variations, annual changes in shipping patterns, or the impact of major global events on maritime operations. In addition, the limited time horizon results in failure to capture significantly changing weather conditions. The dataset lacks extreme weather conditions or large meteorological fluctuations as the study only utilises data from HKP from September to October, a period of relatively stable temperatures. This limitation also resulted in weather-related variables not improving the model results as well, as the narrow range of weather conditions provided insufficient variation to show the full impact of meteorological factors on offshore operations.

7.3.2 Feature engineering limitations

Although OpenFE provides automatic feature generation capabilities, the generated features often lack clear business interpretability, which may limit its adoption in operational environments where

decision transparency is critical. The balance between automatic feature generation and domain knowledge integration remains an open challenge.

Although comprehensive, the current feature set may not capture all relevant factors that affect VAT, such as port congestion index, fuel costs, crew scheduling constraints, or geopolitical factors that may affect routes and schedules.

7.3.3 Real-time implementation challenges

Although this study demonstrated the model prediction accuracy using historical data, real-world deployment faces challenges in real-time data acquisition and quality assurance. The availability and reliability of AIS and meteorological data in the operational environment may be problematic, affecting actual system performance.

This study mainly focuses on statistical performance indicators (MAE, RMSE, R^2), but does not fully consider business-related factors such as prediction cost-effectiveness, user acceptance, system reliability, and integration with existing port management systems.

8 Conclusion and Future Work

8.1 Research summary

8.1.1 Main research results

This study successfully constructed a VAT prediction framework based on multi-source data fusion, and achieved significant prediction accuracy improvement on the Hong Kong port dataset. Experimental results show that machine learning methods can effectively utilise complex patterns in heterogeneous maritime data and greatly improve the accuracy of arrival time prediction.

The TabPFN model performed best in all evaluation scenarios, achieving a mean absolute error (MAE) of 2.88 hours, 3.42 hours, and 3.33 hours under the three data segmentation methods of time series, distance series, and MMSI series, respectively, which is 43.6%, 46.5%, and 42.9% higher than the ETA_AIS baseline. The performance coefficient of variation of the model under different segmentation methods is only 1.3%, showing excellent robustness and generalisation ability.

The LightGBM model, especially when combined with OpenFE feature engineering, performed well as the second best model, with a training time of only 10 seconds, and has extremely high computational efficiency while maintaining excellent prediction performance. The performance of traditional deep learning models (LSTM, Transformer) was lower than expected, with average MAE in the range of 3.6-4.2 hours, indicating that model complexity does not necessarily translate into superior performance in structured tabular data prediction tasks.

8.1.2 Achievement of research objectives

The main research question "How to achieve high-precision VAT prediction through multi-source data fusion and machine learning models" was successfully answered. The optimal model achieved a MAE of 2.88 hours, which is a substantial improvement over traditional methods and meets the accuracy requirements of actual port operations.

Three sub-research questions were fully addressed:

- Record matching: Use the systematic matching logic of Call Sign group matching ETA-ATA, MMSI group matching AIS-ETA, IMO group matching AIS-Weather and AIS-VPP to effectively handle scattered ship records.
- Optimal model architecture: TabPFN and LightGBM+OpenFE combination proved to be a superior architecture for combining ETA and AIS data for spatio-temporal prediction, outperforming traditional deep learning methods in performance.
- Multi-source data integration: Meteorological data contributes greatly to the overall feature importance, and VPP data provides a 3-5% performance improvement, which verifies the effectiveness of multi-source data fusion in improving VAT prediction accuracy.

Meanwhile, this study also successfully addresses three key research gaps identified in the literature:

- **Insufficient application of weather data in VAT prediction**: the study demonstrates that the comprehensive integration of weather data can improve the accuracy of VAT forecasts.The ranking of weather features in feature importance analysis also confirms that meteorological conditions can have a substantial impact on model prediction results.The study overcomes the technical challenges of temporal resolution differences and spatial scale mismatch through a systematic data preprocessing and feature engineering approach, demonstrating that weather data integration is essential for accurate VAT prediction.
- **Traditional feature engineering limitations**: automated feature engineering implemented through OpenFE successfully generated 20 additional features from raw multi-source data and applied them to model predictions.The results demonstrate the significant effectiveness of automated feature selection in discovering complex feature interactions and higher-order transformations across different data sources.
- **Insufficient application of advanced deep learning methods for tabular data**: TabPFN, as a specialised tabular base model, performs well on structured maritime datasets, obtaining competitive results with relatively less training time as well as higher accuracy.The study demonstrates that methods designed specifically for tabular data can effectively process multidimensional structured ship operation data, surpassing traditional neural networks such as LSTM and Transformer originally designed for sequence or image data.This validates the potential of tabular-specific deep learning methods for maritime applications.

8.1.3 Methodological Contributions

This study has made several important methodological contributions in the field of maritime traffic forecasting:

First, the research findings challenge the common assumption that deep learning models outperform traditional machine learning methods in all forecasting tasks. For structured maritime data regression tasks, the optimised traditional algorithms perform well in terms of accuracy, computational efficiency, and interpretability. Second, the successful application of the OpenFE framework brings about 5% stable performance improvement to traditional machine learning models, and effectively discovers complex interaction patterns and non-linear relationships in the original features. Finally, this study systematically verifies the integrated value of comprehensive meteorological data in VAT forecasting for the first time. Meteorological characteristics occupy six of the top 15 important characteristics and contribute significantly to the overall model performance.

8.2 Future Research Directions

8.2.1 Methodological Enhancements

Future research should explore integrated methods that combine the advantages of TabPFN and traditional machine learning models to achieve optimal prediction performance while maintaining computational efficiency. Developing uncertainty quantification methods will support probabilistic predictions and better support risk-based decision making in port operations.

Research on graph neural networks in modelling port networks and route interdependencies may capture complex relationships between multiple ports and multiple ships that current methods may overlook. Advanced time series modelling techniques designed specifically for maritime temporal patterns, including attention mechanisms, deserve further exploration.

8.2.2 Data and Geographic Extensions

Extending the framework to major global ports will verify cross-regional generalisation capabilities and support the development of common forecasting models. Multi-port forecasting networks that consider inter-port dependencies and global shipping patterns can provide more accurate system-level forecasts.

Integrating additional data sources, including satellite remote sensing for precise sea state information, port congestion indicators, ship cargo details, fuel consumption data, and geopolitical risk factors, can further improve forecast accuracy and practical applicability.

Longitudinal studies spanning multiple years will capture the long-term impact of seasonal changes, economic cycles, and global events such as the COVID-19 epidemic on shipping patterns and forecasting model performance.

8.2.3 Application Development

Developing an intelligent port management system that integrates real-time VAT forecasting with berth allocation optimisation, equipment scheduling, and manpower planning represents a natural extension of this research. Such systems can demonstrate the full potential of accurate arrival time predictions in improving port operational efficiency.

Extension to end-to-end supply chain forecasting to support inventory management and distribution planning beyond port operations will maximize the commercial value of improved VAT forecasting capabilities. Integration with multimodal transport networks can provide comprehensive logistics optimization solutions.

Developing commercial platforms for small and medium-sized shipping companies, working with insurance companies seeking dynamic pricing models, and serving shippers who require real-time cargo tracking represent important market opportunities for research translation.

8.3 Practical Recommendations

This study provides empirical guidance for the implementation of VAT predicting systems in maritime operations. The findings provide insights for port operators, companies, and logistics service providers seeking to improve operational efficiency through improved VAT predicting.

8.3.1 Weather Data Integration Strategy

This study proposes a key practical consideration for integrating weather data to forecast VAT. Although the impact of meteorological factors on forecast accuracy appears relatively limited in the results, in actual practice, incorporating meteorological data into the operational system can effectively improve VAT forecast accuracy.

Due to the limited time period (September-October 2021) and the single port (HKP only) of this study, the weather data obtained in this study lacks variability and underestimates the meteorological impact to a certain extent. The dataset lacks seasonal changes and extreme weather events that have a significant impact on maritime operations. Despite these limitations, weather features still play an important role, and for practical implementation, this study recommends the inclusion of weather data based on the following considerations: first, enterprise-level operational systems usually cover a wider geographical and temporal range than personal databases, and can obtain a wider range of data, thereby greatly improving the success rate of predictions; at the same time, the computational overhead of weather data integration is minimal compared to the potential operational benefits under adverse conditions.

8.3.2 Model Selection Framework

The comparative analysis provides clear guidance for model selection based on operational needs. For applications that prioritise prediction accuracy and have a small number of operational data sets, TabPFN shows the best performance, with more improvements than baseline methods, and the model's ability to achieve competitive results without hyper-parameter tuning reduces implementation complexity. For large-scale operational deployments that require real-time predictions, LightGBM combined with OpenFE feature engineering provides the most practical solution. Compared with other models, the shortest training time and relatively good results can maintain excellent prediction performance.

8.3.3 System Scalability and Maintenance

Automated feature engineering capabilities should be included to adapt to evolving data patterns and maintain prediction performance. The performance improvements demonstrated by the OpenFE framework over traditional models support its use in production systems. Continuous performance monitoring mechanisms are essential to detect model degradation and trigger retraining procedures. The computational efficiency of the recommended models (especially LightGBM) enables frequent model updates to maintain prediction accuracy as operating conditions evolve.

9 Summary

With the continuous growth of global maritime trade, the accuracy of Vessel Arrival Time (VAT) prediction has become a critical factor for port operational efficiency and supply chain management. Although maritime transport handles over 80% of global cargo volume, traditional VAT prediction methods still suffer from significant accuracy issues. Taking Hong Kong Port as an example, the average deviation between ship Estimated Time of Arrival (ETA) and Actual Time of Arrival (ATA) reaches 13.8 hours, and this uncertainty leads to enormous port congestion costs and supply chain disruptions.

9.1 Research Background and Motivation

Existing VAT prediction methods primarily rely on static ETA reports or dynamic AIS data, lacking comprehensive data integration. This fragmented approach ignores the actual circumstances of vessel navigation—ships must constantly respond to weather conditions, sea states, and their own physical capabilities during voyage. Traditional prediction models cannot effectively capture these complex non-linear relationships, resulting in limited prediction accuracy. Current research exhibits three main limitations:

- Insufficient application of weather data in VAT prediction;
- Traditional feature engineering methods limiting model performance potential;
- Traditional feature engineering methods limiting model performance potential;

These limitations provided important improvement opportunities for this research.

9.2 Research Methods and Innovation

This study developed a VAT prediction framework based on multi-source data fusion, systematically integrating four key data dimensions: ETA/ATA data, AIS data, weather data (temperature, wave height, wind speed, etc.), and Vessel Physical Parameters (VPP). Through establishing unified vessel identification systems and spatio-temporal data matching algorithms, effective fusion of heterogeneous maritime data was achieved.

At the methodological level, this study employed the OpenFE automated feature engineering framework to handle complex data interaction patterns and systematically compared the performance of six machine learning models, including tree-based ensemble methods (XGBoost, Random Forest, LightGBM), neural network architectures (LSTM, Transformer), and the TabPFN model.

Particularly noteworthy is that this study was the first to apply the TabPFN model, specifically designed for tabular data, to the maritime prediction domain.

9.3 Major Research Findings

Experimental results demonstrate that machine learning methods achieved significant performance improvements in VAT prediction tasks. The TabPFN model performed best across all evaluation scenarios, achieving Mean Absolute Errors (MAE) of 2.88 hours, 3.42 hours, and 3.33 hours under time series, distance series, and MMSI series data splitting methods respectively, representing improvements of 43.6%, 46.5%, and 42.9% compared to the ETA_AIS baseline.

More importantly, this study challenges the common assumption that deep learning models consistently outperform traditional methods in all prediction tasks. For structured maritime data regression tasks, TabPFN and tree-based models such as LightGBM and Random Forest not only exceeded deep learning approaches in prediction accuracy but also demonstrated significant advantages in computational efficiency and model interpretability.

Feature importance analysis revealed the critical role of weather factors. Weather-related features (such as wave height, wind speed, swell height, etc.) occupied 6 positions among the top 15 important features, fully demonstrating the direct impact of marine environmental conditions on vessel navigation efficiency.

9.4 Practical Application Value

The practical significance of this research manifests across multiple dimensions. In port operations management, the 40-46% improvement in prediction accuracy will significantly enhance port resource allocation efficiency. More accurate VAT predictions enable port managers to better arrange berth allocation, schedule loading/unloading equipment, and optimise workforce deployment, thereby reducing vessel waiting times, increasing port throughput, and lowering operational costs.

At the supply chain management level, precise arrival time predictions will enhance visibility and controllability across the entire logistics network. Shippers and logistics service providers can make more reliable inventory management decisions, transportation planning, and customer communications based on improved prediction information, thereby reducing safety stock requirements, improving supply chain responsiveness, and lowering overall logistics costs.

In terms of technical implementation, the excellent performance of both TabPFN (with high prediction accuracy) and LightGBM provides ideal solutions for real-time prediction system construction, meeting the rapid response demands of ports and shipping companies.

9.5 Scientific Contributions and Limitations

This study made multi-dimensional contributions to maritime data science theory. It was the first to systematically verify the importance of meteorological data in VAT prediction, filling a critical theoretical gap in maritime prediction research. The successful application of OpenFE automated feature engineering demonstrated the potential of domain-agnostic machine learning techniques in enhancing maritime prediction systems without requiring extensive domain expertise for manual feature design.

However, this study also has certain limitations. The research was primarily based on data from Hong

Kong Port during September-October 2021, which may limit the global applicability of the results. Different sea areas exhibit significant differences in meteorological conditions, channel characteristics, port facilities, and traffic density patterns, which may affect model generalisation capability. Additionally, while the study demonstrated model prediction accuracy, challenges in real-time data acquisition and quality assurance in practical applications require further consideration.

9.6 Future Research Directions

Future research should explore integrated methods that combine the advantages of TabPFN with traditional machine learning models to achieve optimal prediction performance while maintaining computational efficiency. Extending the framework to major global ports will verify cross-regional generalisation capabilities and support the development of universal prediction models.

In terms of data dimensions, integrating additional data sources including satellite remote sensing for precise sea state information, port congestion indicators, detailed ship cargo information, fuel consumption data, and geopolitical risk factors can further improve prediction accuracy and practical applicability.

Developing intelligent port management systems that integrate real-time VAT prediction with berth allocation optimisation, equipment scheduling, and workforce planning represents a natural extension of this research. Multi-port forecasting networks that consider inter-port dependencies and global shipping patterns can provide more accurate system-level forecasts, while longitudinal studies spanning multiple years will capture the long-term impact of seasonal changes, economic cycles, and global events on shipping patterns and forecasting model performance.

A Additional graphs and data

A.1 AIS Data Rounded to 0.5° Daily Distribution (October)

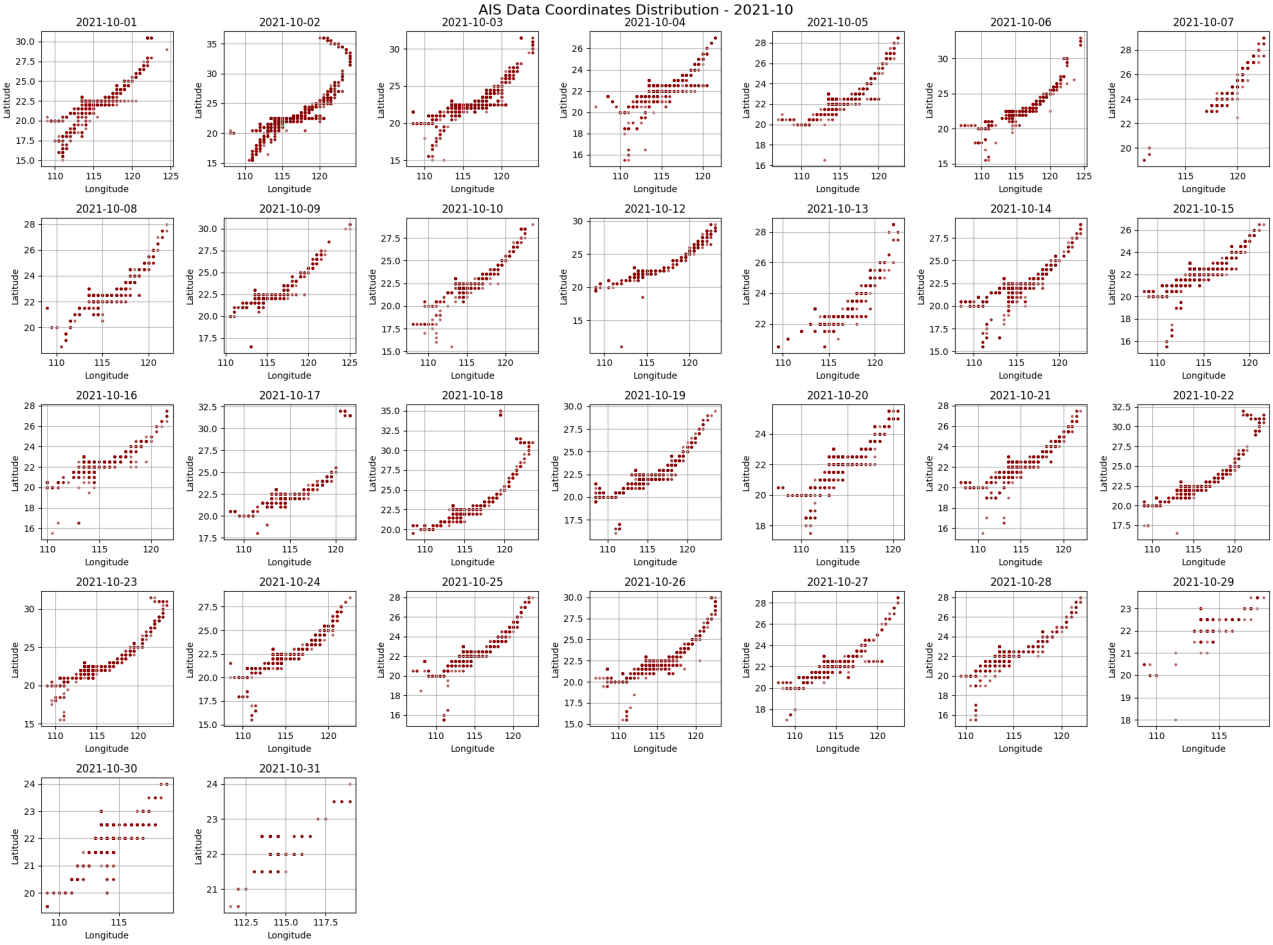

Figure 20: AIS Data Rounded to 0.5° Daily Distribution (October)

A.2 Search Space Definition

Hyperparameter search spaces are defined for each model:

This section outlines the hyperparameter spaces for four machine learning models: XGBoost, Random Forest, LSTM, and LightGBM. Each table lists the hyperparameters, their possible values, and a brief description of their significance in the model.

Table 19: XGBoost Hyperparameter Space

Hyperparameter	Values	Description
learning_rate	[0.01, 0.03, 0.05, 0.1, 0.2, 0.3]	Controls the step size at each iteration while moving toward a minimum of the loss function. Lower values lead to slower but more precise convergence.
max_depth	[4, 6, 8, 10, 12]	Maximum depth of a tree. Higher values increase model complexity but may lead to overfitting.
subsample	[0.6, 0.7, 0.8, 0.9, 1.0]	Fraction of samples used for training each tree. Lower values prevent overfitting but may reduce accuracy.
colsample_bytree	[0.6, 0.7, 0.8, 0.9, 1.0]	Fraction of features used per tree. Lower values reduce overfitting and speed up training.
reg_lambda	[0, 0.1, 0.5, 1.0, 3.0, 10.0]	L2 regularization term on weights. Higher values penalize large weights to prevent overfitting.
reg_alpha	[0, 0.01, 0.1, 0.5, 2.0]	L1 regularization term on weights. Encourages sparsity in feature weights.
n_estimators	[300, 500, 800, 1000]	Number of boosting stages (trees). More trees increase model complexity but may lead to overfitting.

XGBoost Hyperparameter Space

Table 20: Random Forest Hyperparameter Space

Hyperparameter	Values	Description
n_estimators	[200, 300, 500, 800, 1000]	Number of trees in the forest. More trees improve stability but increase computation time.
max_depth	[10, 15, 20, 25, None]	Maximum depth of each tree. Higher values or None allow deeper trees, increasing complexity.
min_samples_split	[2, 5, 10, 20]	Minimum number of samples required to split a node. Higher values reduce overfitting.
min_samples_leaf	[1, 2, 4, 8]	Minimum number of samples required at a leaf node. Higher values smooth the model.
max_features	['sqrt', 'log2', 0.6, 0.8]	Number or fraction of features considered for best split. Affects model diversity and speed.

Random Forest Hyperparameter Space

Table 21: LSTM Hyperparameter Space

Hyperparameter	Values	Description
hidden_dim	[32, 64, 128, 256, 512]	Number of units in the LSTM hidden layer. Larger values increase model capacity but may overfit.
num_layers	[1, 2, 3, 4]	Number of LSTM layers. More layers allow learning complex patterns but increase computation.
dropout	[0.0, 0.1, 0.15, 0.2, 0.3, 0.5]	Dropout rate to prevent overfitting by randomly dropping units during training.
lr	[0.0005, 0.001, 0.003, 0.005, 0.01, 0.02]	Learning rate for the optimizer. Controls step size during gradient descent.
epochs	[50, 80, 100, 150, 200]	Number of training iterations over the dataset. More epochs improve learning but risk overfitting.

LSTM Hyperparameter Space

Table 22: LightGBM Hyperparameter Space

Hyperparameter	Values	Description
learning_rate	[0.01, 0.03, 0.05, 0.1, 0.2]	Step size for gradient descent. Lower values ensure stable convergence but require more iterations.
max_depth	[4, 6, 8, 10, 12, -1]	Maximum tree depth. -1 allows unlimited depth, increasing model complexity.
feature_fraction	[0.6, 0.7, 0.8, 0.9, 1.0]	Fraction of features used per tree. Lower values reduce overfitting and speed up training.
bagging_fraction	[0.6, 0.7, 0.8, 0.9, 1.0]	Fraction of data used per iteration. Adds randomness to prevent overfitting.
reg_lambda	[0, 0.1, 0.5, 1.0, 3.0, 10.0]	L2 regularization term. Higher values penalize large weights to reduce overfitting.
reg_alpha	[0, 0.01, 0.1, 0.5, 2.0]	L1 regularization term. Promotes sparsity in feature weights.
n_estimators	[300, 500, 800, 1000]	Number of boosting iterations (trees). More trees enhance accuracy but may overfit.

LightGBM Hyperparameter Space

A.3 Detailed Model parameters

Table 23: XGBoost Parameters

Parameter	Value	Description
objective	reg:absoluteerror	Specifies the loss function as mean absolute error for regression.
n_estimators	500	Number of boosting stages (trees).
max_depth	8	Maximum depth of a tree, controlling model complexity.
learning_rate	0.05	Step size for gradient descent, affecting convergence speed.
subsample	0.9	Fraction of samples used for training each tree.
colsample_bytree	0.9	Fraction of features used per tree.
reg_alpha	0.1	L1 regularization term on weights, promoting sparsity.
reg_lambda	1.0	L2 regularization term on weights, preventing overfitting.
random_state	42	Seed for reproducibility.
n_jobs	-1	Use all available CPU cores for parallel processing.
enable_categorical	True	Enables handling of categorical features directly.

XGBoost Parameters

Table 24: Random Forest Parameters

Parameter	Value	Description
n_estimators	500	Number of trees in the forest.
max_depth	15	Maximum depth of each tree, controlling complexity.
min_samples_split	5	Minimum samples required to split a node.
min_samples_leaf	2	Minimum samples required at a leaf node.
max_features	sqrt	Number of features considered for best split (square root of total features).
random_state	42	Seed for reproducibility.
n_jobs	-1	Use all available CPU cores for parallel processing.

Random Forest Parameters

Table 25: LightGBM Parameters

Parameter	Value	Description
objective	regression_l1	Specifies the loss function as L1 (mean absolute error) for regression.
n_estimators	500	Number of boosting iterations (trees).
max_depth	8	Maximum tree depth, controlling model complexity.
learning_rate	0.05	Step size for gradient descent, affecting convergence speed.
feature_fraction	0.9	Fraction of features used per tree.
bagging_fraction	0.9	Fraction of data used per iteration, adding randomness.
reg_alpha	0.1	L1 regularization term, promoting sparsity.
reg_lambda	1.0	L2 regularization term, preventing overfitting.
random_state	42	Seed for reproducibility.
n_jobs	-1	Use all available CPU cores for parallel processing.
verbose	-1	Suppresses output messages for cleaner logging.

LightGBM Parameters

Table 26: TabularLSTM Parameters

Parameter	Value	Description
hidden_dim	64	Number of units in the LSTM hidden layer, affecting model capacity.
num_layers	2	Number of LSTM layers, allowing complex pattern learning.
dropout	0.15	Dropout rate to prevent overfitting by dropping units.
epochs	80	Number of training iterations over the dataset.
lr	0.001	Learning rate for the optimizer, controlling step size.

TabularLSTM Parameters

Table 27: TabularTransformer Parameters

Parameter	Value	Description
d_model	64	Dimensionality of the model's embeddings, affecting capacity.
nhead	4	Number of attention heads in the transformer, enabling multi-perspective learning.
num_layers	2	Number of transformer layers, increasing complexity.
dropout	0.1	Dropout rate to prevent overfitting by dropping units.
epochs	80	Number of training iterations over the dataset.
lr	0.001	Learning rate for the optimizer, controlling step size.

TabularTransformer Parameters

References

Brandt, P. (2023). "Maritime accident risk prediction integrating weather data using machine learning". In: *Ocean Engineering* 215, pp. 108–122.

CEIC Data (Dec. 2021). *Container port throughput - Hong Kong SAR*. CEIC Data. URL: https://www.ceicdata.com/zh-hans/indicator/hong-kong/container-port-throughput (visited on 06/13/2025).

Chu, Zhong, Ran Yan, and Shuaian Wang (2022). "Vessel Arrival Time to Port Prediction: A Uniform Approach Integrating Port Call Records with AIS Data". In: *SSRN Electronic Journal*. DOI: 10.2139/ssrn.5004783.

— (2024). "Evaluation and prediction of punctuality of vessel arrival at port: A case study of Hong Kong". In: *Maritime Policy & Management* 51.6, pp. 1096–1124. DOI: 10.1080/03088839.2023.2217168. URL: https://doi.org/10.1080/03088839.2023.2217168.

Dorsser, Chris van et al. (2015). "Factors Affecting Ocean-Going Cargo Ship Speed and Arrival Time". In: *Computational Logistics*. Springer, pp. 429–443.

Filom, S. et al. (2023). "Applications of machine learning methods in port operations: A systematic literature review". In: *Transportation Research Part E* 167, pp. 102–115.

Flapper, E. (2022). "ETA Prediction for Vessels using Machine Learning". In: *Marine Technology Society Journal* 56.4, pp. 45–60.

Hollmann, Noah, Samuel Müller, Katharina Eggensperger, et al. (2023). "TabPFN: A Transformer That Solves Small Tabular Classification Problems in a Second". In: *International Conference on Learning Representations (ICLR)*. URL: https://openreview.net/forum?id=cp5PvcI6w8_.

Hollmann, Noah, Samuel Müller, Lennart Purucker, et al. (2025). "Accurate predictions on small data with a tabular foundation model". In: *Nature* 637.8328, pp. 1–8. DOI: 10.1038/s41586-024-08328-6.

Hong Kong Maritime Authority (2022). *Evaluation and prediction of punctuality of vessel arrival at port: A case study of Hong Kong*.

Hummels, David and Georg Schaur (2012). *Time as a trade barrier*. Tech. rep. w18421. National Bureau of Economic Research.

Investopedia (2024). *Demurrage: What it means, how it works, example*. URL: https://www.investopedia.com/terms/d/demurrage.asp.

Kim, Soong-Ki et al. (2023). "Enhancing Container Vessel Arrival Time Prediction through Past Voyage Route Modeling: A Case Study of Busan New Port". In: *Journal of Marine Science and Engineering* 11.6, p. 1234.

Koehrsen, William (Aug. 2018). *Why automated feature engineering will change the way you do machine learning*. KDnuggets. URL: https://www.kdnuggets.com/2018/08/automated-feature-engineering-will-change-machine-learning.html.

Lam, Jasmine Siu Lee et al. (2023). "Evaluation and prediction of punctuality of vessel arrival at port: a case study of Hong Kong". In: *Maritime Policy & Management*. DOI: 10.1080/03088839.2023.2217168.

Li, Haixiang, Helong Jiao, and Zaili Yang (2023). "AIS data-driven ship trajectory prediction modelling and analysis based on machine learning and deep learning methods". In: *Transportation Research Part E: Logistics and Transportation Review* 175, p. 103152.

Liu, Jingsong et al. (2023). "AIS data-driven ship trajectory prediction modelling and analysis based on machine learning and deep learning methods". In: *Transportation Research Part E: Logistics and Transportation Review* 175.

McKinsey & Company (Mar. 2022). *Navigating the current disruption in containerized logistics*. URL: https://www.mckinsey.com/industries/travel-logistics-and-infrastructure/our-insights/navigating-the-current-disruption-in-containerized-logistics.

Nasir, Muhammad et al. (2024). "Enhancing vessel arrival time prediction: A fusion-based deep learning approach". In: *Expert Systems with Applications* 238, p. 122045.

Notteboom, Theo and Jean-Paul Rodrigue (2008). "Containerisation, box logistics and supply chains: The integration of ports in liner shipping networks". In: *Maritime Economics & Logistics* 10.1-2, pp. 152–174. DOI: 10.1057/palgrave.mel.9100196.

Park, Kyoungho, Seonhyeok Sim, and Hyeon Bae (2021). "Vessel estimated time of arrival prediction system based on a path-finding algorithm". In: *Maritime Transport Research* 2, p. 100012.

Pecan AI (Feb. 2025). *What is automated feature engineering — and why should you use it?* Pecan AI Blog. URL: https://www.pecan.ai/blog/what-is-automated-feature-engineering/.

Rahman, Mohammad Ashiqur et al. (2025). "Prediction of vessel arrival time to port: a review of current studies". In: *Maritime Policy & Management*. DOI: 10.1080/03088839.2025.2488376.

Rong, Hao et al. (2020). "AIS data driven general vessel destination prediction: A random forest based approach". In: *Transportation Research Part C: Emerging Technologies*. DOI: 10.1016/j.trc.2020.102764.

Saleh, Ahmed, Mohamed Hassan, and Khalid Al-Rashid (2023). "Revolutionizing marine traffic management: A comprehensive review of machine learning applications in complex maritime systems". In: *Applied Sciences* 13.14, p. 8099.

Sun, Xiang et al. (2024). "AIS Data-Driven Maritime Monitoring Based on Transformer: A Comprehensive Review". In: *arXiv preprint arXiv:2505.07374*. URL: https://arxiv.org/abs/2505.07374.

Tang, Haixiang, Yong Yin, and Haijiang Shen (2022). "A model for vessel trajectory prediction based on long short-term memory neural network". In: *Journal of Marine Engineering & Technology* 21.3, pp. 136–145.

Thunberg, Marcus et al. (2023). "Predicting vessel service time: A data-driven approach". In: *Transportation Research Part A: Policy and Practice*. DOI: 10.1016/j.tra.2024.104283.

United Nations Conference on Trade and Development (2023a). *Launch of the Review of Maritime Transport 2023*. URL: https://unctad.org/meeting/launch-review-maritime-transport-2023.

— (2023b). *Review of Maritime Transport 2023*. United Nations.

— (2024). *Review of Maritime Transport 2024*. United Nations.

Wang, Shuaian (2025). *Artificial-intelligence based ship estimated-time-of-arrival (ETA) prediction at Port of Hong Kong*. URL: https://sites.google.com/site/wangshuaian/research/hk-eta-prediction (visited on 06/13/2025).

Wu, Li et al. (2025). *TabPFN unleashed: A scalable and effective solution to tabular classification problems*. arXiv: 2502.02527 [cs.LG].

Yang, Jinho et al. (2024). "Enhancing prediction accuracy of vessel arrival times using machine learning". In: *Journal of Marine Science and Engineering* 12.8, p. 1362.

Yu, J. (2021). "Ship arrival prediction and its value on daily container terminal operation". In: *Maritime Policy & Management* 48.5, pp. 689–705.

Zhang, C. et al. (2022). "AIS data driven general vessel destination prediction: A random forest based approach". In: *Transportation Research Part C* 145, pp. 103–118.

Zhou, Y. et al. (2022). "Impacts of wind and current on ship behavior in ports and waterways: A quantitative analysis based on AIS data". In: *Applied Ocean Research* 124, pp. 103–115.