Universiteit Leiden

Data Mining and Machine Learning for Online Prediction of Vessel Arrival Time Residuals

Maria Teodora Tudora (s3819779)

Supervisors:
Yingjie Fan & Robin van der Laag

Bachelor Thesis– Data Science and Artificial Intelligence

Leiden Institute of Advanced Computer Science (LIACS)
liacs.leidenuniv.nl July 1, 2026

Abstract

The goal of this thesis is to study the effect that integrating real-time meteorological features and dynamic port congestion indices has on vessel arrival time prediction accuracy. This thesis presents an online predictive pipeline that integrates real-time Automatic Identification System (AIS) telemetry with meteorological factors and Dynamic Congestion Indices (DCIs). By using a preprocessing registry, a memory-based tracking logic designed to mitigate temporal data sparsity by caching and forward-filling vessel characteristics, and employing ensemble machine learning models (Random Forest and XGBoost), this thesis evaluates the extent to which multi-modal data fusion can improve prediction accuracy. The primary prediction target is the arrival residual, defined as the temporal difference between the actual time of arrival (ATA) and the initial AIS-reported ETA. Using a dataset of 558 voyage samples and a standard 80/20 train-test split recovered through the Vibranium server, the results show that the human baseline achieves a Mean Absolute Error (MAE) of 214.34 hours. The ensemble machine learning models corrected these estimates, with the XGBoost model achieving an MAE of 178.77 hours, which represents a 16.6% improvement over the traditional baseline, with an R^2 score of 0.0998. This reduction outputs an average decrease of 35.57 hours per voyage in arrival uncertainty. Feature importance analysis reveals that physical characteristics, led by draught (21.2%), serve as the main predictors for the Random Forest model, while XGBoost prioritises factors such as the origin and wind speed, which account for 33.7%, respectively 14.9% of the weight of the model. These findings suggest that arrival residuals are linked to voyage trajectories and environmental conditions, demonstrating that ensemble models can reduce the uncertainty of human-reported data.

Contents

1 Introduction

1.1 The situation

The maritime industry functions as the backbone of the global economy, facilitating approximately 90% of international trade by volume [14]. Within this logistical framework, the precision of Estimated Time of Arrival (ETA) is not simply a convenience, but a vital requirement. For port authorities, an accurate ETA determines the efficiency of planning berth- and resource allocation, including the scheduling of maritime pilots, tugboats, and Ship-to-Shore cranes. For the supply chain, ETA precision allows for just-in-time synchronisation, reducing the carbon footprint of stationary vessels, and minimising the costs associated with port congestion.

The main motivation for improving ETA precision is the *Berth Allocation Problem* (BAP). Ports operate on very tight schedules where quay space and crane availability must be planned in advance. When a vessel's actual arrival deviates from its reported ETA, it leads to idle port resources or costly congestion in the anchorage zones. By integrating dynamic factors such as real-time weather and vessel draught, this research aims to provide a reliable prediction that mitigates the uncertainty that is inherent in traditional maritime scheduling.

Traditionally, standard ETA prediction models have relied on deterministic calculations [17]. In practice, these are represented by the AIS-reported ETA, which is generated by the onboard navigation systems using the great-circle distance between the current coordinates and the destination of a vessel, divided by the ship's current Speed Over Ground (SOG). However, these traditional methodologies are limited, since they treat the journey of a ship as a continuous path, unimpeded by external factors. In reality, when a vessel approaches a port, it enters an environment which is guided by stochastic variables. Traditional models fail to account for environmental factors, such as limited visibility or the wind speed, or aspects caused by port congestion, for instance queueing.

Theoretical research [17] suggests that vessel behaviour is a complex, non-linear system, influenced by both navigational intent and external constraints. The use of the Automatic Identification System (AIS) has improved the ability to model this behaviour. AIS data provides a high-resolution dynamic stream of telemetric data, such as speed and position, which is suitable for mining behavioural patterns through machine learning. Furthermore, recent empirical studies [3] have demonstrated that ensemble machine learning models, which combine the assets of multiple algorithms, can significantly reduce prediction error and enhance berth scheduling efficiency compared to single-algorithm models.

This research uses an asynchronous data pipeline to predict arrival residuals, defined as the temporal difference between the initial AIS-reported ETA and the Actual Time of Arrival (ATA). By focusing on residual rather than the raw arrival timestamp, the model acts as a correction layer to the crew's manual estimate. This is achieved by integrating two dynamic data streams into the predictive process: real-time meteorological data (wind speed, visibility), port-side operational metrics (congestion indices), and voyage-specific features (origin and destination ports). Unlike offline studies that use static and cleaned historical databases, the online nature of this thesis requires the model to react to live data snapshots. This approach aligns with the multiple stages strategy proposed by Gui [7], which suggests that the closer the vessel gets to a port, the more the model must focus more on operational and environmental factors instead of navigational data. By deploying this system on the LIACS Vibranium server, this thesis evaluates to what extent this multi-modal integration can improve the reliability of maritime logistics in an increasingly congested port network. This approach provides an alternative to deterministic models, capable of accounting for multi-dimensional noise that is encountered in global logistics.

This bachelor thesis is organised as follows: Chapter 2 contains the theoretical background on AIS telemetry and the impact of meteorological factors on maritime navigation. Chapter 3 discusses related work, reviewing current literature on machine learning applications in the maritime industry. Chapter 4 details the methodology, including data acquisition, archival, and stateful preprocessing pipeline. Chapter 5 describes the machine learning experiments and evaluates the performance of the model. Finally, chapter 6 concludes the thesis and suggests directions for future research.

1.2 Research Questions

To address the limitations of deterministic models currently used in the maritime industry, this thesis investigates the following primary research question:

To what extent can ensemble machine learning models reduce the residual error of AIS-reported ETAs by integrating physical, port-related, and meteorological features?

To answer this, the following sub-questions are investigated:

1. Which feature groups, specifically physical vessel characteristics, meteorological conditions, or voyage- and port-related factors, contribute the most to arrival residuals?
2. Does the inclusion of meteorological data and the dynamic congestion index features improve predictive performance compared to models using only AIS-based vessel features?
3. How does model performance vary across different port categories, specifically between tidal-restricted and open-sea ports?

2 Background

2.1 The Automated Identification System (AIS)

The Automatic Identification System (AIS) is an autonomous maritime broadcast system which operates in the Very High Frequency (VHF) band. It was originally mandated by the International Maritime Organization (IMO) under the International Convention for the Safety of Life at Sea (SOLAS) and its main purpose is collision avoidance and maritime security. AIS allows vessels to electronically exchange real-time data, such as identification, position, course, and speed, with other nearby ships, coastal stations, and satellite receivers [9].

Today, there are three different types of AIS products,classes A, B, and Receive Only Systems, the main ones being Classes A and B. Class A transceivers are required for all large commercial and SOLAS-regulated vessels. These units utilise a Self-Organized Time Division Multiple Access (SOTDMA) protocol, ensuring high reliability and frequent update rates. In contrast, Class B transceivers are intended for smaller, non-mandated vessels and have two separate IMO specifications. The first type utilises a Carrier-Sense Time-Division Multiple-Access (CSTDMA) system with lower priority, while the other type is similar to Class A transceivers and uses SOTDMA. For the purposes of this thesis, the focus is strictly on Class A commercial vessels, as they provide the high-density, standardised telemetry required for robust machine learning [12].

2.2 AIS Message Structure

From a data engineering point of view, the most vital aspect of AIS is its message-based architecture. AIS data is not a single unified record, but rather a stream of distinct message types. For arrival prediction, two categories are of essential [15]:

Dynamic Messages (Types 1, 2, and 3): Broadcast every 2 to 10 seconds. These provide the Speed Over Ground (SOG), the vessel's speed relative to the earth's surface, and the Course Over Ground (COG), the actual path followed relative to the ground.

Static and Voyage-Related Messages (Type 5): Broadcast every 6 minutes. These contain the Maritime Mobile Service Identity (MMSI), vessel dimensions, draught, and destination.

The difference in broadcast interval (seconds vs. minutes) creates temporal asynchronicity. This technical reality necessitates the stateful registry logic described in Chapter 4, where the system must remember the last known static state of a vessel to enhance the high-frequency dynamic telemetry.

2.3 Meteorological and Operational Impact on Navigation

Vessel transit time is not purely a function of engine output, but is heavily influenced by environmental friction. High wind speeds increase the aerodynamic drag on a vessel's superstructure, especially for high-sided vessels like container ships and tankers. Furthermore, reduced visibility due to fog or heavy precipitation necessitates safe speed manoeuvres under COLREGs (International Regulations for Preventing Collisions at Sea), leading to significant deviations from the reported ETA. From an operational point of view, as a vessel enters a port's authority, it is subject to pilotage requirements and traffic separation schemes (TSS). If a port is congested, vessels are often directed to anchorage zones, creating a non-linear waiting time that deterministic models based on instantaneous SOG cannot account for. [9]

2.4 The Deterministic Baseline

The traditional baseline for estimating arrival time is the deterministic kinematic model. This model assumes a linear progression, using the following equation:

$$\text{ETA}_{\text{baseline}} = T_{\text{current}} + \frac{d}{v}, \tag{1}$$

where v represents the SOG and d represents the great-circle distance to the destination, which represents the shortest path between 2 points on a surface of a sphere. This is obtained by applying the Haversine formula to the vessel's current coordinates and the coordinates of the destination port's centroid.

While mathematically straightforward, this model is limited by its reliance on instantaneous velocity. If a vessel is at anchor ($v = 0$), and the model cannot account for the waiting time caused by port congestion [7]. The maritime environment is stochastic, since factors such as currents, waves, and wind introduce variance that neither the deterministic model, nor the machine learning models can eliminate. However, while the deterministic model treats these factors as external noise, the machine learning models in this thesis attempt to capture their influence by including meteorological and operational variables as predictive features. In the experimental phase of this thesis, the AIS-reported ETA is used as the main representative of this deterministic approach.

2.5 Ensemble Machine Learning for Regression

To capture the complex, non-linear relationships between meteorological conditions, port congestion, and arrival residuals, this thesis uses ensemble machine learning. Ensemble methods are algorithms where predictions of multiple individual models, known as base learners, are combined to produce a final robust framework with a very good predictive performance. The collection of different models can counteract the individual errors of each of its components, leading to better generalisation on unknown data, compared to any single model of this collection. In the context of maritime technology, where AIS telemetry is frequently compromised by sensor noise and human errors, ensemble machine learning can provide mathematical stability [8].

The main advantage of using this approach is its ability to manage the bias-variance trade-off. The expected prediction error of a model contains bias, variance, and noise. Bias refers to the error introduced by making simplistic assumptions about the data. High-bias models, such as linear regression, tend to underfit the data and, in regards to this thesis, could fail to capture non-linear friction of port congestion. Variance refers to the model's tendency to fluctuate in response to small changes in the training set. High-variance models, such as unpruned decision trees, usually overfit the data by capturing outliers as if they were underlying patterns. Ensemble techniques are made to counteract these issues using two different strategies: bagging, which focuses on reducing variance, and boosting, which focuses on reducing bias. [4, 16, 5].

Both Random Forest [4] and XGBoost [5] use decision trees, specifically classification and regression trees, as their base learners. A regression tree splits the multi-dimensional feature space into a set of non-overlapping regions through the process of binary recursive splitting. At each node, the algorithm evaluates every available feature j and every possible threshold s to identify the optimal split that maximises the homogeneity of the resulting subsets. This is achieved by minimising the Sum of Squared errors (SSE) for the split, which is defined as:

$$\text{SSE}_{\text{split}} = \sum_{i \in R_L(j,s)} (y_i - \hat{y}_{R_L})^2 + \sum_{i \in R_R(j,s)} (y_i - \hat{y}_{R_R})^2 \tag{2}$$

where R_L and R_R represent the left and the right child nodes resulting from the split, y_i is the actual arrival residual for a given voyage, and $\hat{y}_R$ is the mean response of all samples within the region. By applying this formula iteratively, the tree identifies the most significant predictors, such as specific thresholds for vessel draught or wind speed, that correlate with delays in arrival.

In the context of maritime telemetry, where AIS signals are subject to sensor noise and human-entry errors, two specific ensemble models are evaluated against the baseline.

Random Forest (RF), as introduced by Breiman [4] in 2001, is an ensemble machine learning algorithm that builds multiple decision trees to produce a robust prediction. Breiman defines the model as a collection of tree-structured predictors $\{h(\mathbf{x}, \Theta_k), k = 1, \dots\}$, where random vectors $\{\Theta_k\}$ are independent and using the same distribution, ensuring that each tree is grown using a unique stochastic process. This is achieved using randomness: firstly, using bootstrap aggregating (bagging), where each tree is trained on a random subset of the training data sampled with replacement, and secondly, with feature randomness, where only a random subset of the features is used and evaluated at each node split. This process separates the individual base learners, allowing the model to identify patters across a diverse feature space that might be overlooked in one decision tree. Moreover, Breiman provides a mathematical proof using the Strong Law of Large Numbers that random forests do not overfit hence the number of trees increases, but instead converge to a limiting value of the generalisation error. To monitor this error during the training phase, the model uses out-of-bag (OOB) samples, which are the data points excluded from each bootstrap sample, which provide an unbiased estimate of the generalisation error during the training phase, ensuring robustness against telemetric noise [4].

In contrast to the parallel architecture of RF, XGBoost (Extreme Gradient Boosting), as introduced by Chen and Guestrin [5], is a scalable tree boosting system that implements a sequential learning strategy designed to minimise residuals through gradient descent [5]. This model builds on the gradient boosting machine (GBM) framework, which was initially proposed by Friedman [6]. This model generates a final prediction through K additive functions, where each successive tree is trained to predict the errors of the ensemble formed by all the previous trees. To control the complexity of the model and prevent overfitting, XGBoost minimises a regularised objective function:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \tag{3}$$

where $\Omega(f) = \gamma T + \frac{1}{2}\lambda \|w\|^2$ penalises the number oof leaves (T) and the leaf weights (w) [5]. The algorithm optimises this objective using a second-order approximation, leveraging first-order gradient statistics (g_i) and second-order gradient statistics (h_i) to identify optimal split points. A technical advantage of XGBoost is its sparsity-aware algorithm, which learns a default direction for missing values at each node based on the training data. This allows the model to handle datasets with missing entries, inherent in AIS static reports, in a unified way by exploiting the sparsity pattern to make computation complexity linear to the number of non-missing entries. By using a learning rate (η), the model scales newly added weights after each boosting step to prevent overfitting, allowing the ensemble to iteratively refine its correction of the human-reported ETA baseline. [6]

The interpretability of these ensemble ML models is achieved through feature importance analysis, which quantify the contribution of each temporal, physical, and environmental variable. On the one hand, for the RF model, Breiman [4] introduces a variable importance measure based on the permutation of the OOB samples. This process involves randomly permuting the values of a specific feature within the OOB data and measuring the resulting decrease in prediction accuracy. On the other hand, XGBoost evaluates feature importance through gain, which represents the fractional improvement in the regularised objective function brought by a specific feature during the recursive splitting process [5].

2.6 Evaluation Metrics

To measure the performance of the ensemble machine learning models, as well as the operational baseline, this thesis uses two main statistical metrics: Mean Absolute Error (MAE) and the Coefficient of Determination (R^2). These metrics provide a good view of model accuracy and explanatory power.

Mean Absolute Error (MAE): The MAE measures the average extent of errors in a prediction set, without considering their direction. It is the average over the test sample of the absolute differences between prediction and observation where individual differences have equal weight. [13] Mathematically, it is defined as:

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n} |y_i = \hat{y}_i| \tag{4}$$

In the context of this thesis, MAE is the most interpretable metric for port authorities, since it expresses rhe average rpedction error directly measured in hours. Unlike the Mean Squared Error (MSE), MAE does not square the residuals, which makes it more robust to outliers and typographical errors, which are common in manually-reported AIS data.

Coefficient of Determination (R^2): The R^2 score provides an indication of the goodness of fit and a measure of how well unseen samples are likely to be predicted by the chosen model. The R^2 score represents the proportion of variance in the dependent variable (the arrival residuals) that is predictable from the independent variables (the features) [13]. Mathematically, it is defined as:

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \tag{5}$$

where $\bar{y}$ is the mean of the observed data. A negative value of the R^2 score indicates that the model performs worse than a horizontal line which represents the mean residual. This highlights the stochasticity of global maritime trajectories.

3 Related Work

3.1 AIS Data Mining

The foundation of modern maritime informatics is the Automatic Identification System (AIS), which is a tracking system that has transformed maritime situational awareness. However, the transition from raw telemetric pings to high-fidelity predictive models is hindered by significant data quality challenges. Yang et al. categorise the AIS research landscape into a hierarchical framework, structured as a pyramid [15]. The base layer represents large-scale AIS data, requiring filtering and cleaning, the middle layer focuses on trajectory prediction, ad the top layer involves collision avoidance. This suggests that the reliability of high-level ETA predictions, dependent on the integrity of the underlying data processing layer.

A primary concern highlighted in the literature is the unreliability of raw AIS streams, which are often human-centred rather than simply technical. These issues are not only technical but often human-centred. Manually entered information, such as the vessel's current draught, destination, and crew-estimated ETA, is frequently subject to inconsistent formats, typos, or even a lack of updates. Furthermore, raw AIS data is characterised by blank segments, which are trajectory discontinuities. They are temporal windows where signal loss or intentional transponder deactivation lead to missing coordinates, making continuous tracking difficult. [15]

The asynchronous nature of AIS broadcasting is of particular technical importance to this thesis. While dynamic messages, which are types 1, 2, and 3 and provide coordinates and speed, are transmitted every few seconds, static and voyage-related messages of type 5 are broadcast only every six minutes or upon request. This creates a fragmentation problem where the majority of data points in a live stream lack essential physical characteristics such as ship dimensions or current loading status. Yang et al. note that many contemporary studies struggle with this untidy data, which can often result in datasets with significant blank segments or outliers. To mitigate this, researchers such as Zissis et al. suggest that more advanced data interpolation and resampling techniques are necessary. [17]

Furthermore, Yang et al. emphasise that traditional statistical methods, such as simple linear regressions, are increasingly incapable of handling the complexity of global AIS data. While global telemetry volumes are vast, this thesis focuses on a high-integrity subset of data to evaluate the feasibility of online prediction. They argue that the emergence of machine learning offers a promising approach to unlock the full potential of this data. However, they identify a significant gap in the current literature: the majority of research remains offline, using cleaned, historical datasets. This thesis addresses this by developing a targeted online pipeline that processes live snapshots of data. By being based on the methodology of Yang et al., this thesis treats AIS as a complex, multi-dimensional dataset that requires data engineering to be viable for real-time prediction.

3.2 Machine Learning in the Berth Allocation Problem

The operational and economic justification for high-precision ETA prediction is centred on the Berth Allocation Problem (BAP). The BAP is a combinatorial optimisation problem where port authorities must assign arriving vessels to specific berths and assign ship-to-shore(STS) cranes for cargo operations such as loading and unloading. Albloushi et al. highlight that this process is extremely sensitive to timing. When a vessel's actual arrival deviates from its reported ETA, it triggers operational disruptions, including vessel turnaround issues and suboptimal use of quay resources. Idle port resources, such as stationary cranes and waiting maritime pilots, represent opportunity costs for port authorities, while vessels forced to wait in anchorage zones contribute to increased fuel consumption and carbon emissions [3].

In their research, Albloushi et al. conducted a comparative analysis of various machine learning architectures to solve this uncertainty. The thesis utilised historical data from Norwegian base stations to train Gradient Boosting, K-Nearest Neighbours (KNN), and Random Forest (RF) Regression models. A key finding of their work, which informs the model selection for this thesis, is that Random Forest Regression consistently outperformed other algorithms, achieving an R^2 value of 0.704 and a Mean Absolute Percentage Error (MAPE) of 2.85%. Albloushi et al. attributes this performance to the ensemble nature of the Random Forest, which averages the results of multiple decision trees to capture the non-linear relationships between a vessel's physical particulars, such as length and draught, and its arrival time. This is particularly relevant in congested port environments where the relationship between speed and arrival is not a simple linear function of distance [3].

However, while the work of Albloushi et al. provides a robust framework for vessel-centric prediction, it reveals a limitation common in maritime literature: the exclusion of port-side operational context. Albloushi et al.'s models focus primarily on the characteristics of the ships, but do not explicitly account for the "queuing effect" or the real-time traffic density at the destination. This thesis extends Albloushi et al.'s methodology by introducing the DCI. This thesis argues that the BAP cannot be solved by looking at the vessel in a vacuum. By quantifying the ratio of anchored to moored vessels, this thesis provide the ensemble models with a proxy for port friction. This allows the research to move beyond the vessel-specific focus of Albloushi et al. and incorporate the state of the port as a critical predictive feature, thereby evaluating a multi-modal approach to the uncertainties inherent in maritime scheduling [3].

3.3 Data Sparsity and Quality

Data quality remains a challenge in the field of maritime artificial intelligence, a topic explored throughout the paper by Yang et al. [15]. They state that while the volume of AIS data is vast, its utility is often compromised by the heterogeneous nature of the telemetry. Specifically, manually entered information, such as a vessel's draught, dimensions, and destination, is prone to human error, inconsistent naming formats, and a lack of timely updates [15]. Furthermore, Rückert et al. [10] emphasise that scheduling uncertainty is often worsened by external factors such as sea currents and wind, which introduce stochastic noise that traditional deterministic models struggle to interpret [10]. In a data science context, this noise is a technical impediment that can increase measurement error and predictive variance. If a model is trained on incomplete or low-fidelity data, its ability to generalise to real-world port approach scenarios is diminished.

A specific technical challenge identified in this thesis is the temporal sparsity of AIS broadcasts.

While dynamic position reports are transmitted every few seconds, static characteristics such as vessel dimensions and draught are broadcast only every few minutes. In a live data stream, this results in a fragmented dataset where the majority of incoming position reports lack the physical features necessary for accurate regression. Traditional offline studies often avoid this issue by using historical databases where static data has been pre-joined or by simply dropping rows with missing values [3]. However, in an online environment, excessive data removal is not a viable strategy as it leads to a loss of real-time situational awareness. This thesis addresses this issue by using data interpolation concepts suggested in maritime literature to maintain a persistent vessel profile.

3.4 The Research Gap

While the reviewed literature covers a wide range of machine learning models and historical AIS analysis, there is limited exploration of real-time data fusion across different global geographic hubs. Many existing models in the maritime domain are offline, which means that they are trained and evaluated on static, pre-cleaned historical databases [3, 7]. While these studies are valuable for theoretical validation and algorithm benchmarking, they do not account for the real-time latency, data noise, and synchronisation challenges that are encountered in live port operations. This thesis addresses this gap by deploying a pipeline that performs inference on snaphopts of weather and port traffic. This approach aligns with the "Multiple Stages Strategy" proposed by Gui [7], which suggests that the variables influencing a ship's arrival change in importance as the vessel nears its destination.

Gui et al.'s [7] theory argues that in the open-sea phase of a voyage, kinematic features, such as SOG and Course Over Ground (COG), are the main arrival predictors. However, as the vessel enters the geofenced port area, defined in this thesis as the bounding box surrounding the port centre, environmental and operational factors become more influential. To evaluate this, this thesis has integrated real-time meteorological data, specifically wind speed and visibility, with the proposed DCI. As noted by Yang et al. (2024), the identification and integration of these multi-modal data sources pose a critical challenge, as researchers must align diverse streams based on their lowest temporal resolution [15]. This thesis solves this through a 30-minute temporal join, providing a framework that is operationally applicable for modern port authorities.

Furthermore, most existing studies are geographically limited, focusing on a single port or a specific region. By testing the system across sixteen different global ports 1, this thesis evaluates the generalisability of multi-modal fusion across different infrastructure profiles. This paper also includes a comparative evaluation between tidal-restricted river and open-sea ports to determine how geographic constraints influence the accuracy of the prediction. This thesis fills the gap between theoretical trajectory prediction and practical decision support by providing a model that accounts for port friction in real-time. Consequently, this thesis evaluates a predictive framework for the uncertainties of maritime scheduling that is currently missing from regional, offline studies.

4 Methodology

4.1 Data Acquisition and Management

The technical basis of this thesis is an asynchronous data pipeline hosted on the LIACS Vibranium server. Data is sourced in real-time from the *aisstream.io* WebSocket API [2], which provides a live stream of JSON-decoded AIS messages. The thesis period spans from February to July, capturing a progression from raw telemetric pings to a training set of nearly 558 voyage samples. The pipeline uses two message types: *PositionReport* (Types 1, 2, and 3) for kinematic data and *ShipStaticData* (Type 5) for vessel dimensions and draught. In this thesis, online refers to the real-time ingestion of data and the generation of predictions on live snapshots, while the model training process is conducted offline The source code for the data pipeline, preprocessing scripts, and machine learning models is available at the following repository: `https://github.com/teotudora/Vessel-Arrival-Time-Residual-Prediction.git`.

To reduce computational scope and filter-out open-sea noise, this thesis implements a geofencing strategy using sixteen discrete *bounding boxes*, as provided in Table 1. These boxes isolate traffic within a specific coordinate range of approximately 20-50 km of major port centroids, as detailed in Table 1. This ensures that the dataset is focused on the final part of the voyage, where environmental and operational friction is most common.

The following table details the sixteen global shipping hubs. Each port is defined by a rectangular geofence (bounding box) used to filter the live *aisstream.io* WebSocket data.

Port Name	**Latitude Range**	**Longitude Range**	**Port Type**
Rotterdam	[51.841, 52.157]	[3.773, 4.566]	Tidal / River
Antwerp	[51.148, 51.425]	[3.991, 4.629]	Tidal / River
Bremerhaven	[53.405, 53.742]	[8.319, 8.826]	Coastal / Estuary
Hamburg	[53.421, 53.766]	[9.420, 10.213]	Tidal / River
Felixstowe	[51.893, 52.003]	[1.204, 1.405]	Coastal
Singapore	[1.134, 1.478]	[103.590, 104.169]	Open Sea Hub
Hong Kong	[22.149, 22.435]	[113.816, 114.297]	Coastal Hub
Busan	[34.937, 35.314]	[128.745, 129.296]	Coastal Hub
Le Havre	[49.409, 49.543]	[0.006, 0.318]	Coastal
Los Angeles	[33.631, 33.804]	[-118.337, -118.165]	Coastal Hub
Genoa	[44.305, 44.437]	[8.747, 8.972]	Coastal
Piraeus	[37.913, 37.951]	[23.604, 23.646]	Coastal
New York	[40.471, 40.855]	[-74.155, -73.891]	Coastal / Estuary
Savannah	[31.901, 32.219]	[-81.212, -80.715]	Tidal / River
Vancouver	[49.029, 49.351]	[-123.476, -123.027]	Coastal / Fjord
Melbourne	[-38.572, -37.625]	[144.233, 145.334]	Coastal / Bay

Table 1: Target port network with precise bounding box coordinates and geographic classifications.

To ensure data integrity and reproducibility requirements, all incoming AIS packets are stored in their raw *JSONL* (JSON Lines) format. This strategy is chosen for its efficiency and ability to preserve the original UTC timestamps, navigational status codes, and nested voyage dictionaries, such as crew-reported ETA and destination. The system allows for rigorous offline validation and ensures that the data provided by the vessels remains unaltered during the ingestion phase, facilitating reproducibility. The raw telemetry is stored on the server's storage volumes for the entire duration of the thesis project. This facilitates iterative data processing and the extraction of features, such as the DCI, as the research requirements evolve. Access to the server and the JSONL files is restricted for the researcher. The data is accessed via secure SSH protocols, and file system permissions are configured to prevent unauthorised modification of data. These measures ensure the privacy of vessel data and maintain the integrity of the dataset used for model training.

4.2 Data Preprocessing and Voyage Reconstruction

The main technical challenge in this pipeline is the temporal asynchronicity of AIS messages. While dynamic position reports are transmitted every few seconds, static voyage characteristics are sent less frequently. To address this, a stateful registry logic was implemented. The system maintains a memory-resident dictionary matched to the vessel's `MMSI` (Maritime Mobile Service Identity). As new packets arrive, the registry updates the ship's state, pinning the most recent draught and destination to every subsequent position report. Data cleaning steps were also implemented to ensure model robustness:

To reconstruct the dataset for machine learning, the following definitions were applied:

- Voyage Sample: A voyage sample is defined as a unique sequence of telemetric signals representing a single vessel's transit from the boundary of a port's geofenced area to its final moored status at the ATA.
- Actual Time of Arrival (ATA): The ATA is the first recorded timestamp where the vessel's SOG reaches 0.0 knots within the geofenced area.
- Initial AIS-reported ETA: The temporal value extracted from the first ShipStaticData (Type 5) message that is received immediately following the ship's entry into the geofenced area.

To prevent data leakage, an 80/20 train-test split was performed at the voyage level rather tha the message level. This ensures that signals from the same journey do not appear in both the training and testing phases, which forces the models to generalise to unseen vessel approaches. data cleaning removed redundant messages, with invalid SOG, and vessels broadcasting destinations that are outside of the target network. Missing draught values were handled via the stateful registry. If no static report was available for a vessel, the median draught for that specific ship type was used.

4.3 Feature Engineering

To enhance the predictive accuracy of the models beyond crew-reported values, a multi-dimensional feature set was derived.

Physical Vessel Characteristics: Based on the *ShipStaticData* messages, features such as vessel length and current draught were extracted. The dataset was filtered to include only commercial cargo and tanker vessels (AIS ship types 70–89). According to the ITU-R M.1371-5 technical standard [1], these numerical identifiers represent all categories of cargo ships (70-79) and tankers (80-89). These categories were chosen because they follow standardised logistical routes and are subject to port friction and anchorage delays that this thesis aims to model.

Environmental Factors: Meteorological data was sourced from the OpenWeatherMap API [11] including wind speed *(m/s)* and visibility *(m)*, and merged with vessel telemetry via a 30-minute temporal join. A secondary feature set was established to track destination weather at 30-minute intervals from the vessel's departure time, recorded when a vessel entered the port's geofenced bounding box.

Operational Metrics: A **DCI** was calculated for each port hub. The index quantifies the ratio of vessels with a navigational status of "At Anchor" to the total number of vessels "Moored" within the geofenced area defined in table 1 and is defined as:

$$DCI = \begin{cases} \frac{V_{\text{Anchored}}}{V_{\text{Anchored}} + V_{\text{Moored}}} & \text{if } (V_{\text{Anchored}} + V_{\text{Moored}}) > 0 \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

Vessels underway or with restricted mobility are excluded to focus on the availability of berth resources. While AIS navigational status is manually set by the ship's crew, it remains a reliable real-time proxy for port-side latency. This feature allows the model to adjust predictions based on real-time traffic bottlenecks.

4.4 Baseline Selection and Predictive Modelling and Evaluation

To measure the predictive advantage of the ML models, a main benchmark needed to be established to represent current standards in the maritime industry. While a raw d/v deterministic calculation provides a theoretical baseline, this thesis uses the AIS-reported ETA as the operational deterministic baseline. This choice was made because the AIS-reported ETA represents the deterministic estimate which is currently used in maritime field. It includes both the vessel's internal calculations and the manual adjustments that are made by the ship's crew.

Two ensemble machine learning models were implemented:

1. Random Forest (RF): Confogured with 100 estimators to reduce variance through bagging.

2. XGBoost: Used a learning rate (η) of 0.05 and a maximum tree depth of 5 to prevent overfitting on the voyage samples.

By evaluating the ML models against this baseline, this thesis measures the arrival residual ($y = \text{ATA} - \text{ETA}_{\text{AIS}}$). Outperforming this baseline demonstrates the ability to account for stochastic friction that standard systems can ignore. the performance of the ensemble models is evaluated using the Mean Absolute Error (MAE) to measure the average extent of the prediction error, and the coefficient of determination (R^2) to quantify the proportion of variance in arrival residuals explained by the feature set.

5 Experiments and Results

This section presents the evaluation of the ensemble machine learning models. The predictive experiments were conducted using a dataset of 558 voyage samples. As defined in Section 4, the ATA was determined by the first instance of zero SOG within the bounding box. The main goal was to predict the *arrival residual*, although to ensure model robustness, voyages showing residuals over 720 hours were excluded in order to eliminate human errors. Voyage samples showing an arrival residual larger than 720 hours were excluded. This was selected to eliminate typographical errors in the manual AIS entries, such as incorrect day or month input by the ship crew, while preserving delays caused by meteorological events or congestion.

5.1 Experimental Setup

The dataset was divided using a 80/20 train-test split at the voyage level to prevent data leakage. The models were trained on a feature set containing nine variables:

- Physical Characteristics: Draught, Length, Width
- Meteorological Conditions: Wind Speed, visibility
- Operational Factors: Dynamic Congestion Index, Port_ID
- Temporal Features: Arrival Time

The ML models were implemented using the following configurations:

- RF: 100 estimators, using Mean Squared Error (MSE) as the function to measure the quality of the splits, with a fixed random state of 42.
- XGBoost: 100 estimators, a learning rate (η) of 0.05, and a maximum tree depth of 5. This depth was chosen to prevent overfitting considering the sample size of *N=550*.

The main prediction target (y) is the arrival residual:

$$y = \text{ATA} - \text{ETA}_{\text{AIS}} \tag{7}$$

By modelling the residual rather than the raw arrival timestamp, the ensemble models function as a dynamic correction layer to the human baseline. Two ensemble models were evaluated:*Random Forest* and *XGBoost*. Both models use the same nine features that include the physical characteristics of the ship, the temporal markers, and the environmental data.

5.2 Model Performance Comparison

The performance of the machine learning models was evaluated against the human baseline, which represents the raw error present in the manual AIS-reported ETAs. As established in Section 4.7, this baseline represents the standard estimate generated by onboard navigational systems. To calculate the predictive accuracy explained by each model, performance was measured using the Mean Absolute Error (MAE) and the Coefficient of Determination (R^2). The results are summarised in Table 2.

Model	**MAE (Hours)**	R^2 **Score**	**Improvement %**
AIS-reported ETA (Baseline)	214.34	0.0000	0.0%
Random Forest	180.16	0.0666	15.95%
XGBoost	**178.77**	**0.0998**	**16.59%**

Table 2: Comparison of model performance for predicting arrival residuals against the human baseline.

The results show that machine learning provides an advantage over the traditional estimate. XGBoost appeared as the superior model, achieving a 16.59% gain over the baseline and reducing uncertainty by 35.57 hours per journey. While the R^2 scores remain low, this is consistent with the stochasticity of global maritime voyages. However, the 35 hour reduction in MAE represents an improvement for port's resource planning as well as the mitigation of the BAP.

5.3 Ablation Study and Geographic Performance Analysis

To measure the impact of multi-modal data and address research question 3, an ablation analysis was performed by observing the impact of feature groups on the XGBoost model. The results show that while physical traits such as draught, length, and width provide a good foundation for the prediction, they are not sufficient to capture residuals in isolation. Including temporal factors and voyage-specific context provided the most important error reduction. The addition of meteorological data (wind) and the DCI provided the necessary correction layer to take into account the environmental and operational inefficiency, which results in a final MAE of 178.77% hours.

Moreover, the model's performance was disaggregated by port category. As defined in the methodology section 4, ports were classified as either tidal / river or coastal / open sea. The results of this comparison are summarised in the following table 3.

Category	Voyages (N)	Prediction MAE (hours)
Tidal / River	529	3.72
Coastal / Open Sea	74	3.23

Table 3: Model performance comparison based on port category.

The analysis shows that tidal-restricted river ports such as Rotterdam and Hamburg exhibit a higher MAE of 3.71 hours compared to coastal or open-sea ports, which show an MAE of 3.23

hours. This 14.8% increase in error for river-based ports confirms that geographical restrictions increase prediction difficulty. In tidal ports, ships are subject to narrow channel navigation and specific tidal windows, which introduce additional layers of operational bottlenecks which are less common in open-sea environments.

5.4 Feature Importance

A comparative analysis of feature importance reveals an important difference in the way each algorithm perceives maritime friction and corrects the human-reported baseline.

5.4.1 Random Forest: Physical Constraints

On the one hand, the *Random Forest* model prioritised the vessel's draught (21.2%) and the arrival hour (17.7%). This shows that the bagging-based model relies on stable physical and temporal factors to identify patterns in arrival residuals. In maritime logistics, a vessel's draught serves as a substitute for its displacement and cargo load. Deeper-draught vessels are prone to stricter navigational restrictions and slower pilotage procedures. By identifying these physical characteristics as the most reliable predictors of a vessel's deviation from its reported ETA, the RF model captures the physics of the voyage. This suggests that a small portion of the error in manual ETAs comes form a failure to account for the limitations of manoeuvring larger and heavier ships.

5.4.2 XGBoost: Operational and Environmental Dynamics

On the other hand, XGBoost shifted the focus to operational and dynamic variables. The most influential factor was the origin port (33.7%), followed by the the speed of the wind (14.9%) and port congestion (11.6%). This shift indicates that while physical characteristics provide a baseline for prediction, the actual residual error is determined by environmental factors and the ship's trajectory. The frequency of the origin identifier suggests that arrival residuals correlate to characteristics of the port of departure, which may introduce delays that vessel crews often miscalculate.

Moreover, XGBoost assigned a higher weight to environmental factors, with wind speed (14.9%), accounting for an important portion of the model's decision-making. This is almost triple the weight assigned by RF (5.7%). This demonstrates that the boosting process, which builds trees sequentially in order to minimise residuals, is better equipped to capture the non-linear relationship between atmospheric resistance and the resulting delays in the last parts of the journey. This finding validates the inclusion of meteorological factors in the predictive pipeline and confirms that environmental friction is an important factor of the uncertainty in manually-entered maritime estimates.

6 Conclusions and Further Research

6.1 Conclusion

This thesis evaluated the extent to which ensemble machine learning models can reduce the vessel arrival residual error of AIS-reported ETAs through multi-modal data streams. The main task was to develop an online data pipeline that is capable of integrating physical vessel characteristics, real-time meteorological factors, and port-related operational metrics in order to provide a correction to manual maritime estimates.

The experimental results show that XGBoost is the main model for this task, achieving an MAE of 178.77% hours. This model outputs a 16.59% improvement over the operational deterministic baseline, which effectively reduces arrival residuals by an average of 35.57% hours (*214.34% - 178.77% = 35.57%*) per ship journey. While the R^2 scores are relatively low, the reduction in MAE shows the usefulness of ML models in stochastic environments.

The analysis of the feature sets revealed that data regarding the context of voyages, mainly in the identifier of the origin (33.66% weight), is the main predictor of arrival time residuals. This suggests that the path of a ship is more important in determining ETA errors than the ship's dimensions are. While physical characteristics provided a good baseline for the RF model, they were less important in regards to the XGBoost model.

The merging of meteorological data and the DCI are significant factors in model prediction. Wind speed accounted for 14.9% of the predictive weight for XGboost, while DCI contributed 11.59%. These results confirm that environmental and operational bottlenecks are measurable factors of arrival time residuals that traditional baseline models do not capture.

Furthermore, the thesis confirms that the geographical aspects of ports also influence the predictive performance. Tidal-restricted ports showed a 14.86% higher MAE (3.71 hours) compared to open-sea ports (3.23 hours) in the final port-type analysis. This demonstrates that navigation in narrow channels and short tidal windows introduce even more stochasticity that make predictions even more difficult.

6.2 Limitations

The findings of this thesis are subject to several restrictions that provide context for model performance and reported R^2 scores.
Operational Limitations: The main operational limitation involves the temporal resolution of the meteorological data. This was identified by comparing the AIS telemetry with the 30-minute intervals of the weather API. Especially in narrow navigational channels, sudden shifts in visibility can influence decisions regarding pilotage within minutes. Because the model relies on 30-minute intervals, it may fail to account for these small events regarding the weather. Moreover, geographical data coverage was also identified as a limitation through coordinate analysis of the raw data, which revealed blind-spots in terrestrial AIS network, mainly in Eastern Asian ports. Finally, the human factor remains a main source of noise in data. Raw data inspection confirmed that manual input of the destination of a vessel is prone to typographical errors, which led to the exclusion of several hundred potential voyage samples.

Methodological Limitations: An important limitation is the sample size (*N=558*). While the dataset was enough to demonstrate a 16.59% improvement of the winning XGBoost model over the baseline, such a model requires thousands of samples to reach optimal convergence. This lack of data contributed to the low R^2 scores, as the model struggled to explore the variance of the maritime industry. In addition, this thesis assumes a linear relationship between certain features, such as wind speed and arrival residuals. In a real-life situation, environmental bottlenecks operate using thresholds. For example, wind speed might only impact manoeuvring once it exceeds a Beaufort scale threshold [17]. Finally, while the data pipeline facilitates online deduction, training the models remains an offline process, which prevents the system from adapting to sudden changes in maritime patterns.

6.3 Further Research

Future work could focus on implementing incremental online learning, where model weights are updated as soon as new data arrives. This would allow the predictive framework to remain relevant in changing logistical environments, such as during canal obstructions or strikes. Moreover, integrating higher-resolution hydrological data, including tidal heights and data regarding currents, could reduce the MAE especially in tidal ports.

Additionally, future research could also explore the use of natural language processing (NLP) to integrate radio communication logs or port authority scheduling data into the feature set. Finally, applying deep learning architectures might better capture the sequential dependencies in vessel trajectories that tree-based ensemble models, such as RF and XGBoost might overlook.

References

[1] e-navigation.nl. https://www.e-navigation.nl, 2026.

[2] AISStream. Aisstream.io: Real-time global ship tracking api. https://aisstream.io/, 2026.

[3] Anfal Albloushi, Ioannis Karamitsos, Andreas Kanavos, and Sanjay Modak. Predicting vessel arrival time using machine learning for enhanced port efficiency and optimal berth allocation. In *Artificial Intelligence Applications and Innovations*, pages 285–299, Cham, 2026. Springer Nature Switzerland.

[4] Leo Breiman. Random forests. *Machine Learning*, 45(1):5–32, October 2001.

[5] Tianqi Chen and Carlos Guestrin. Xgboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '16, page 785–794. ACM, August 2016.

[6] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[7] Xuhao Gui. Data-driven method for the prediction of estimated time of arrival (ETA). *Transportation Research Record*, 2675(12):1291–1305, 2021.

[8] Trevor Hastie, Robert Tibshirani, and Jerome Friedman. *The Elements of Statistical Learning: Data Mining, Inference and Prediction*. 2nd edition, 2009.

[9] International Maritime Organization. Convention on the International Regulations for Preventing Collisions at Sea, 1972 (COLREGs). https://www.imo.org, 2026.

[10] Lorenz Kolley, Nicolas Rückert, Marvin Kastner, Carlos Jahn, and Kathrin Fischer. Robust berth scheduling using machine learning for vessel arrival time prediction. *Flexible Services and Manufacturing Journal*, 35(1):29–69, March 2023.

[11] OpenWeather. Openweathermap api. https://openweathermap.org, 2026.

[12] Raymarine. "AIS Systems: Which kind is right for my boat?". https://www.raymarine.com/en-us/learning/online-guides/ais-systems-which-kind-is-right-for-my-boat, 2026.

[13] Abhishek V Tatachar. Comparative assessment of regression models based on model evaluation metrics. *International Research Journal of Engineering and Technology (IRJET)*, 8:850–854, 2021.

[14] UNCTAD. Review of maritime transport 2023. Technical report, United Nations Conference on Trade and Development, 2023.

[15] Ying Yang, Yang Liu, Guorong Li, Zekun Zhang, and Yanbin Liu. Harnessing the power of machine learning for AIS data-driven maritime research: A comprehensive review. *Transportation Research Part E: Logistics and Transportation Review*, 183:103426, 2024.

[16] Cha Zhang and Yunqian Ma. Ensemble machine learning: Methods and applications. 2012.

[17] Dimitrios Zissis, Elias K. Xidias, and Dimitrios Lekkas. Real-time vessel behavior prediction. *Evolving Systems*, 7(1):29–40, 2016.