Journal of *Marine Science and Engineering*

MDPI

Article

Automatic Identification System-Based Prediction of Tanker and Cargo Estimated Time of Arrival in Narrow Waterways

Homayoon Arbabkhah, Atefe Sedaghat, Masood Jafari Kang and Maryam Hamidi *

Department of Industrial and Systems Engineering, Lamar University, Beaumont, TX 77710, USA; harbabkhah@lamar.edu (H.A.); asedaghat@lamar.edu (A.S.); mjafarikang@lamar.edu (M.J.K.)
* Correspondence: mhamidi@lamar.edu

Abstract: In maritime logistics, accurately predicting the Estimated Time of Arrival (ETA) of vessels is pivotal for optimizing port operations and the global supply chain. This study proposes a machine learning method for predicting ETA, drawing on historical Automatic Identification System (AIS) data spanning 2018 to 2020. The proposed framework includes a preprocessing module for extracting, transforming, and applying feature engineering to raw AIS data, alongside a modeling module that employs an XGBoost model to accurately estimate vessel travel times. The framework's efficacy was validated using AIS data from the Port of Houston, and the results indicate that the model can estimate travel times with a Mean Absolute Percentage Error (MAPE) of just 5%. Moreover, the model retains consistent accuracy in a simplified form, pointing towards the potential for reduced complexity and increased generalizability in maritime ETA predictions.

Keywords: travel time; ETA prediction; AIS data; XGBoost

Citation: Arbabkhah, H.; Sedaghat, A.; Jafari Kang, M.; Hamidi, M. Automatic Identification System-Based Prediction of Tanker and Cargo Estimated Time of Arrival in Narrow Waterways. *J. Mar. Sci. Eng.* **2024**, *12*, 215. https://doi.org/10.3390/jmse12020215

Academic Editors: Ryan Wen Liu, Dongfang Ma and Xinqiang Chen

Received: 16 December 2023
Revised: 15 January 2024
Accepted: 22 January 2024
Published: 25 January 2024

1. Introduction

Shipping plays a pivotal role in the intricate web of global trade, serving as the lifeblood of international commerce. It facilitates the movement of goods across vast oceans as well as connecting markets and economies in a seamless exchange of products and resources. With over 80% of the world's trade volume being carried by ships, the maritime industry is a cornerstone of the global economy [1]. Consequently, the efficiency of the shipping system, particularly in ports, yields a profound influence on global trade and supply chains. Port congestion, which is attributed to 93.6% of delays, primarily stems from congestion issues, and it underscores the critical need for effective port operational planning [2]. For ports to function smoothly, ships must adhere to arrival schedules. Research indicates that shipping delays significantly impact port operations. Consequently, the estimated time of arrival (ETA) assumes a pivotal role in port operational planning and management. ETA represents the anticipated date and time of a shipment's arrival at a specified destination. An uncertain ETA hampers the ability of ports to formulate efficient logistics plans, emphasizing the crucial role of accurate arrival time predictions.

In maritime terms, a "narrow waterway" designates a constricted waterway characterized by limited breadth and depth, posing navigational challenges. Contrasted with the ample maneuvering space in open seas or larger water bodies, these confined areas demand precise navigation. Skillful handling and often the aid of local pilots are essential for navigating these constrained spaces [3]. Challenges include a restricted turning radius where big vessels have to sail extra miles to make a U-turn [4], possible strong currents or tides, and close quarters with other vessels or the boundaries of the waterway itself [5]. Deep-draft vessel navigation in main waterways is managed by channel pilots, tasked with ensuring safe, orderly bidirectional traffic. The process begins as arriving deep-draft vessels request pilotage. A pilot is assigned to a vessel when two key criteria are met: the availability of the intended dock and the accessibility of the channel. Channel unavailability

can arise from various factors, such as fog or the transit of large vessels. Since halting in the main channel's centerline is prohibited, vessels await at sea buoys or designated anchorages until these conditions are fulfilled [3].

The Estimated Time of Arrival (ETA) in narrow waterways is crucial and significantly impacts maritime operations. When a vessel completes its ocean journey and enters shallow waterways, terminal operations need to be meticulously managed. This involves guiding vessels through the waterway to their destination terminal for loading/unloading. Predicting the ETA to the destination greatly affects terminal efficiency. The ETA is vital for various operational decisions, such as scheduling at the terminal, assigning pilots to vessels, controlling traffic at waterway–highway intersections such as bascule bridges, and decision-making regarding dredging and temporary channel closures to manage maritime traffic in channels and narrow waterways. This study focused on the critical role of travel time estimation in narrow waterways for its significant implications for maritime transportation.

While the estimation of travel time on highways and roadways has been extensively explored in the literature, there is a noticeable gap in research concerning the Estimated Time of Arrival (ETA) for shipping systems. Traditional traffic estimation methods typically rely on aggregated data such as traffic flow, average speed, and congestion distribution. Additionally, statistical modeling of travel time has been applied in city planning. More recently, machine learning techniques have gained prominence in studying vehicle travel times within urban routes. These methods encompass established approaches like random forest and decision trees, as well as sophisticated deep learning architectures. In contemporary city planning, diverse deep learning structures, including Recurrent Neural Networks (RNN) with Long-Short Term Memory (LSTM) [6] and even Graph Neural Networks (GNN) [7], are employed for travel time estimation. However, the majority of studies focus on travel time estimation within city routes. This research uniquely employs a machine learning approach to estimating the time of arrival for ship movement in waterways, extending from a specific origin to their destined locations.

To tackle the challenge of estimating travel time, this study introduces a machine learning framework that leverages AIS data for predicting arrival times between two points along sea routes. This framework is designed to not only process AIS data but also incorporate additional spatial information about vessel trajectories, such as path weight and segment features, as supplementary data inputs. The proposed framework involves the initial preprocessing of AIS data, followed by inputting the prepared data into an XGBoost (v 2.0.1) model and determining the optimal parameters for the model. Subsequently, the trained model undergoes testing using historical AIS data to assess its performance. The experimentation in this study is conducted on historical AIS data of Houston Ports in the United States. The results demonstrate the efficacy of the model, evaluated through five different metrics, namely, mean absolute error (MAE), root mean square error (RMSE), R squared, mean absolute percentage error (MAPE), and root mean square logarithmic error (RMSLE).

The subsequent sections of this paper are structured as follows: Section 2 provides a review of related work in the field of estimating time of arrival. Section 3 outlines the AIS data, defines the problem, and extracts important information from the dataset. Section 4 details the proposed model framework and its constituent modules. Finally, Section 5 presents our conclusions and outlines potential future research.

2. Related Works

Considerable research has been conducted on the precise and timely prediction of Estimated Time of Arrival (ETA) to enhance decision-making across diverse application domains. ETA prediction plays a vital role in air traffic control, impacting arrival sequencing, scheduling, methods for assigning airport gates, and flight arrival time [8–13]. In the realm of road transportation, studies have been undertaken to forecast vehicle management [14], as well as the ETAs of buses [15,16], emergency ambulance services [17,18], and cargo. The current literature lacks studies on estimating vessel arrival times in ports

using historical tracking data. Vessel ETA from AIS messages are often unreliable due to manual input. Despite the requirement for a 72 h advance notice, accurate predictions are challenging, as authorities must verify notifications, the lead time is sometimes too long, and there is a need for insight into approaching vessel volumes for optimal port operations. Therefore, investigating the theoretical foundations and practical applications of machine learning in a business context, research has delved into the performance of various algorithms, including Random Forests, Neural Network architectures, and Linear Regression, on different datasets related to maritime transportation [19]. These machine learning algorithms are applied to provide a qualitative estimation of vessel ETA, aiming to alleviate the consequences of inconsistent arrivals at ports [20].

2.1. Path Finding/Other Methods

In recent years, the Dijkstra algorithm [21], which is often used to find the shortest path in problems, and its derivative, the A* algorithm [22], have been used in studies to calculate routes considering weather conditions. Alessandrini et al. [23] discussed a novel data-driven method for estimating vessel arrival times in port areas, leveraging the abundant data available from ship reporting systems like (AIS) and Long-Range Identification and Tracking (LRIT). The approach utilizes historical maritime traffic data from these systems, focusing on a specific area of interest. It employs an optimized data-driven path-finding algorithm to process these data. Chen et al. employed maritime image sequences for predicting the trajectories of ships [24]. Park et al. [2] presented an ETA prediction system based on a path-finding algorithm. With increasing container volumes and vessel sizes, efficient port operations are crucial. The proposed methodology utilizes AIS data-driven techniques, including data mining and reinforcement learning, to identify possible vessel trajectories. Additionally, the Markov Chain property and Bayesian Sampling are introduced to estimate the vessel's speed over ground (SOG). Wu et al. [25] introduced an AIS-data-based model for the precise estimation and distribution of vessels' travel time and trip numbers in narrow channels, crucial for efficient traffic control. The model involves identifying a vessel's destination dock, arrival/departure times, and estimating travel time between specific points. Additionally, the model addresses the separation of a vessel's trips, contributing to a comprehensive understanding of its journey. Applied to the Houston Ship Channel, the model reveals that travel times are lognormally distributed and influenced by vessel characteristics. Interestingly, trip numbers and travel times exhibit a correlation, providing valuable insights for channel management.

Wu et al. [26,27] examined the transit patterns of tankers and cargos through the Sabine–Neches Waterway (SNWW). Unlike highway traffic, vessel travel time at the SNWW and the entrance of Galveston Ship Channel showed independence from traffic density [28]. Kang et al. [29] studied 15 legs in the Singapore Strait, observing a correlation between vessel travel time and traffic density. The disparity in findings may be attributed to differing vessel volumes, with the Singapore Strait experiencing higher traffic than SNWW and Galveston Ship Channel.

2.2. Machine Learning

Artificial intelligence is about building systems capable of understanding and solving real-world problems by acquiring knowledge from experience. Machine learning is a subfield of artificial intelligence that refers to the ability of extracting insights from data. Machine learning models include deep learning, which is an advanced class of machine learning, inspired by the human brain function and based on Artificial Neural Networks (ANN) and representation learning. Deep learning has networks with the capability to learn complicated concepts even with unstructured data.

2.2.1. Road Application

Several studies have used machine learning (ML) approaches to predict travel times based on GPS traces from vehicles [30,31] or the so-called live Automatic Vehicle Locations

(AVL) data [32,33]. Larsen et al. [34] employed an NN to predict the travel times of buses using open real-time data derived from the Sao Paulo City bus fleet location, real-time traffic data, and traffic forecast from Google Maps. Alam et al. [35] used a Recurrent NN (RNN) architecture to predict the ETA irregularities by exploring live AVL data from buses, provided by the Toronto Transit Commission, along with schedules retrieved from GTFS and weather data. Chondrodima et al. [36] addressed the challenge of predicting public transport ETA using General Transit Feed Specification (GTFS) data. The proposed approach employs a novel combination of Particle Swarm Optimization (PSO) and Radial Basis Function (RBF) neural networks, incorporating a modified PSO-NSFM algorithm for training. A unique pre-processing pipeline, CR-GTFS, is introduced for cleansing and reconstructing GTFS data.

2.2.2. Waterway Application

ANNs have been applied to solve some issues in shipping, for instance, container flow forecasting [37], container dwell time [38], navigational behavior prediction [39], and detecting navigable area for autonomous navigation [40]. To minimize the unpredictability of ship arrivals, recent studies have turned to data mining methods for arrival prediction. In [41], for instance, the author utilized a Neural Network (NN) model to forecast the time intervals between ship arrivals. Subsequently, the predicted interval times are integrated into a model that optimizes the allocation of human resources, leading to successful outcomes that offer valuable support to planners. Pani et al. [42,43] utilized both k-means and Ward's method to cluster daily records from the Cagliari International Container Terminal. This clustering aims to categorize arrival delays into three levels. Subsequently, Classification and Regression Trees (CART), Random Forest (RF), and Naive Bayes (NB) were employed to estimate the delay level. Notably, RF demonstrated superior predictive performance, boasting a relative absolute error of 29% when compared to CART and NB. In a different approach. Pallotta et al. [44] introduced the unsupervised method Traffic Route Extraction and Anomaly Detection (TREAD) to learn a statistical model from AIS data for maritime traffic at the Cagliari International Container Terminal. Pani et al. [20] employed Logistic Regression (LR), CART, and RF to estimate arrival ship deviations at both the Cagliari International Container Terminal and the PSA-Antwerp terminal. Additionally, Parolas et al. [45] applied Support Vector Machines (SVM) and NN to predict ETA for container ships arriving at the Port of Rotterdam. The results showed that both SVM and NN outperformed ETA predictions based on ship agent estimations, with SVM models surpassing NN in Mean Absolute Error. Collectively, these studies hold significance and provide valuable references for the application of data mining in predicting ship arrival times at specific ports. Noman et al. [46] investigated the use of Gradient Boosting Decision Trees (GBDT), Multi-Layer Perceptron Neural Networks (MLP), and Gated Recurrent Unit Neural Networks (GRU) for predicting vessel ETA in inland waterways. It used historical AIS data for training and compared the accuracy of these methods. The GRU algorithm outperformed the others. Yu et al. [47] focused on ship arrival prediction and its impact on the daily operations of Gangji (Yining) Container Terminal (GYCT) in China. Utilizing data mining methods such as Back-Propagation network (BP), CART, and RF, the study aims to enhance the accuracy of predicting ship arrival delays or advances. The results indicate that RF outperforms BP and CART, with ETA month and ship length identified as crucial factors influencing arrivals at GYCT.

3. Problem Definition

AIS Data

The Automatic Identification System (AIS), first introduced in 1990, is used in maritime traffic to record the historical trajectory of vessels. Its main objectives are to enhance maritime safety, improve situational awareness, and facilitate efficient maritime operations. In 2004, the International Maritime Organization (IMO) required all ships exceeding 300 gross tonnages to record and broadcast AIS data [48].

The frequency of AIS message transmission from vessels varies from 2 to 30 s, depending on their speed. The range at which these messages can be received is influenced by various factors such as signal propagation conditions, sea state, and the height and strength of the transmitting and receiving antennas. Reception ranges can vary from 20 nautical miles to up to 350 nautical miles under optimal conditions. Typically, an AIS receiver network is expected to achieve an average reception radius of around 40 nautical miles.

AIS operates by acquiring position and movement data from the vessel's GPS system or an internal sensor within the AIS unit. These data, along with other programmable information from the AIS unit (such as Maritime Mobile Service Identity (MMSI) number, vessel name, destination, and cargo type), are periodically transmitted. The system not only sends out information but also receives data from other vessels' AIS systems.

Each AIS message contains both static and dynamic information. Static information includes vessel attributes, while dynamic information covers the spatial-temporal data of the vessel [3]. The MarineCadastre website provides access to AIS data [49]. Table 1 displays the static and dynamic information contained in AIS data.

Table 1. Static and dynamic information in AIS data.

Static Information	Dynamic Information
MMSI Number	Ship's Position with Accuracy indication
IMO Number	Position timestamp (in UTC)
Name and Call Sign	Course Over Ground (COG)
Length and Beam	
Type of Ship	
Location of Position	

The first challenge in ETA analysis is to transform raw AIS messages into useful data required in ETA prediction. As shown in Table 1, AIS data do not include trip information such as trip number, trip origin, destination, start time, and end time. To get such information, we apply a trip separation algorithm to raw AIS data. The algorithm works based on comparing each vessel's AIS message and the previous one. It uses a hash table to capture the last AIS record of each vessel. Then, it calculates the time and length difference between the current and the last records. The calculated time difference and spatial distance help us to filter redundancies and noises. As the next step, it assumes a vessel's direction is "Stopped" if its speed is less than two knots (~1.151 miles per hour). Defining one stop to the next one as a vessel trip, the algorithm assigns trip numbers to the processed AIS data. Finally, the processed data are stored in a local database. This paper does not discuss the algorithm but rather uses its output to predict ETA for vessels. For an in-depth exploration of the methodology, we recommend reviewing our previous works [50,51].

In this study, we applied our trip separation method to extract trip data from raw AIS data. To estimate the time of arrival for vessels, this information has to be transformed into the form of a complete trip for each vessel. A complete trip dataset includes longitude and latitude coordinates from the origin to the destination, the time taken for each trip, and a sequence of segments representing the vessel's path. The methodology section provides a comprehensive explanation of the algorithm.

4. Methodology

The model proposed in this study consists of three modules. In the first module, AIS data undergo preprocessing, and new features are incorporated into the dataset. The second module utilizes the preprocessed data to train an XGBoost (v 2.0.1) model, incorporating hyperparameter optimization and defining a validation strategy. The final module is

dedicated to applying the trained model to a test dataset and comparing the results with the actual travel time. The three modules are illustrated in Figure 1.

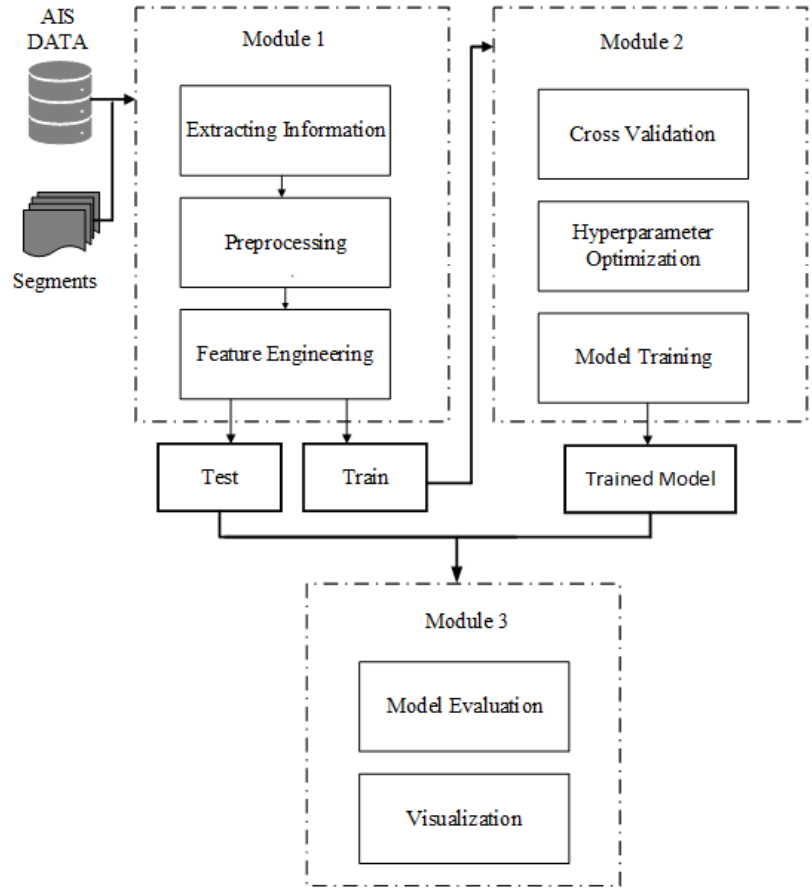

Figure 1. Overview of the proposed framework.

Subsequent sections provide a more detailed explanation of each module.

4.1. Module 1: Preprocessing

The model module cannot directly use raw AIS data; hence, preprocessing is essential before feeding the data into the model. This preprocessing module encompasses various steps, such as addressing missing values, feature engineering, and implementing Principal Component Analysis for dimensionality reduction. The subsequent sections elaborate on each of these steps in greater detail.

4.1.1. Segmentation of Area of Interest

The vessel's journey from its origin to its destination involves traversing a predefined route. To facilitate the tracking of the vessel's path, a network of segments is established using a Geographical Information System (GIS) layer, which maps all waterways within the designated Area of Interest (AoI). For this study, the AoI is the Gulf Intracoastal Waterway (GIWW). To segment the waterway, a function in QGIS (v 3.32.3) named "split line to maximum length" is utilized, dividing the channel's centerline into segments of 2 miles each. As the vessel moves through these segments, the IDs of the segments it passes are captured and stored as features in the database. A schematic representation of this network of segments is depicted in Figure 2.

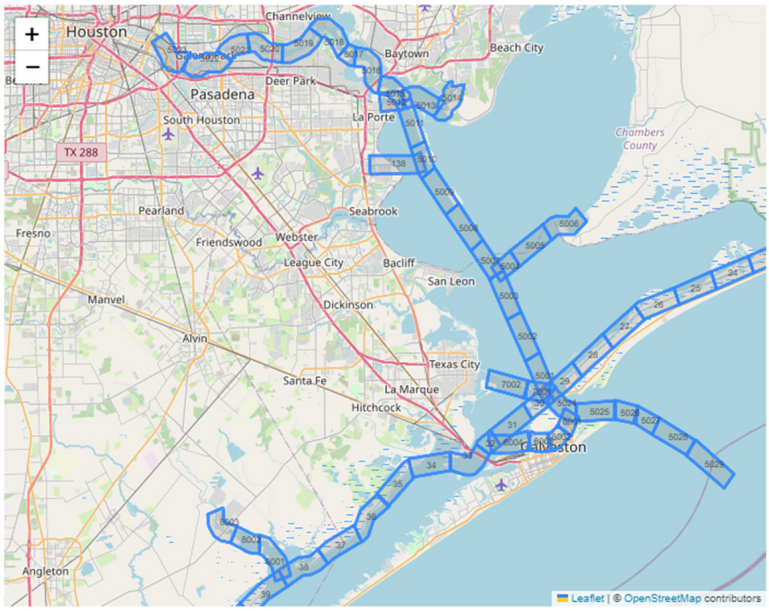

Figure 2. Schematic representation of segment network.

4.1.2. Extracting Features from Raw AIS Data

AIS data encompass details such as a vessel's location, speed, date, and time. In this section, the primary features are initially extracted from the raw data utilizing the Sedaghat [50,51] algorithm. Their approach enables the retrieval of information related to a vessel's route, velocity, direction, and trip number.

In the next step, the extracted information is processed to compute a complete trip for each vessel. A complete trip normally starts from the ocean and ends at a terminal and vice versa. A trip is deemed complete when there is a change in the vessel's trip number. Consequently, the data processed in the previous step are organized by trip number for each vessel, considering the latitude and longitude of the origin and destination. Subsequently, the travel time for the entire trip is computed. A visual representation of a complete trip is depicted in Figure 3.

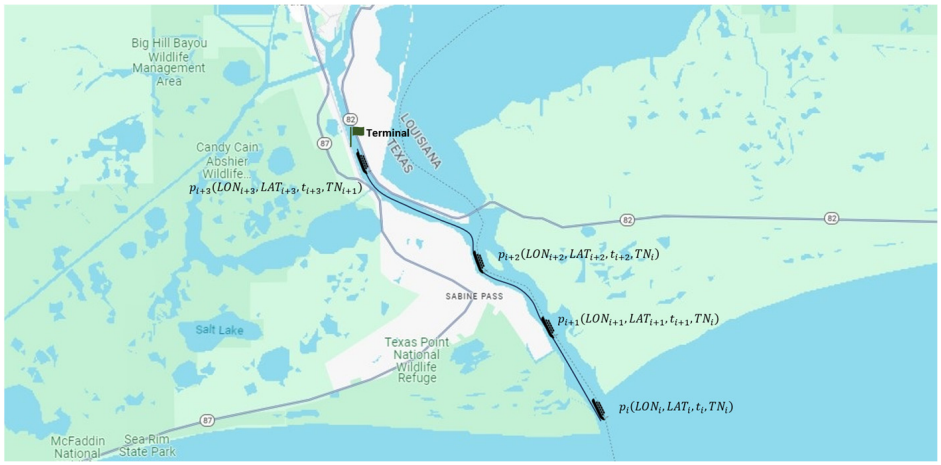

Figure 3. Visual representation of a complete trip.

As illustrated in Figure 3, when there is a change in the vessel's trip number, it signifies the completion of the previous trip. Consequently, the AIS information of the origin and

destination, along with the corresponding travel time, can be computed. For example, in Figure 3, a complete trip can be calculated using Equation (1)

$$OD_{trip} = P_{i+2}(LON_{i+2}, LAT_{i+2}, t_{i+2}, TN_{i+1}) - P_i(LON_i, LAT_i, t_i, TN_i) \quad (1)$$

where $LON_i, LAT_i, t_i,$ and TN_i represent the longitude, latitude, time, and trip number at location i, respectively.

4.1.3. Principal Component Analysis

Principal Component Analysis (PCA) is a powerful statistical technique used to reduce the dimensionality of data while preserving as much information as possible. This is achieved by identifying a set of orthogonal (uncorrelated) directions called principal components (PCs) that capture the greatest variance in the data. By projecting the data onto these PCs, we can obtain a lower-dimensional representation that is often sufficient for many tasks, including visualization, data analysis, and machine learning [52]. Principal components can be calculated by using singular value decomposition (SVD) of the dataset matrix. The SVD of the matrix X can be expressed as Equation (2)

$$X = U\Sigma V^T \quad (2)$$

where:

- U is an $N \times N$ orthogonal matrix containing the left singular vectors.
- Σ is an $N \times D$ diagonal matrix containing the singular values on the diagonal.
- V is a $D \times D$ orthogonal matrix containing the right singular vectors.

To compute the principal components (PCs) of the X dataset, the singular value decomposition of the dataset must be calculated initially. Subsequently, by selecting the top K right singular vectors (V_K) and their corresponding singular values (Σ_K), where K represents the desired dimensionality of the reduced data, the projected X in the new dimension can be calculated using Equation (3):

$$X_{new} = XV_K\Sigma_K^{(-1/2)} \quad (3)$$

where $\Sigma_K^{(-1/2)}$ is a diagonal matrix containing the reciprocal square root of the top K singular values.

PCA method also can be used as a method to reduce noise in the dataset by mapping data to a new space without reducing the dimensionality of the dataset. In this study PCA algorithm is applied to the latitude and longitude of the origin and destination feature to map the data to a new space with lower noise.

4.1.4. Feature Engineering

In machine learning, feature engineering involves extracting and manipulating data to transform them into a format suitable for training and improving the performance of machine learning models. This process allows the model to better understand the underlying patterns and relationships within the data, leading to more accurate and generalizable predictions. In this study, season, hour, minutes, and day of the week of trip beginning are extracted from raw AIS data and added to the dataset to improve model accuracy. Also, two new features are introduced including vessel segments path and path weight. These new features are explained in more detail in the following sections.

Vessel Segments Path

In the AIS data provided for ETA prediction, the trajectory of vessels that pass through different segments is introduced. In this study, all segments for every vessel are incorporated into the dataset as a one-hot-encoded vector appended to the original dataset. While this approach introduces additional sparsity to the dataset, it enriches the dataset with

more information about vessel paths from origin to destination, thereby aiding the model in achieving more accurate travel time estimations. The encoded representation of each segment is defined by Equation (4):

$$e_{segment} = \begin{cases} 1 \text{ if vessel path through segment} \\ 0 \quad \text{otherwise} \end{cases} \tag{4}$$

Path Weight

The trajectory each vessel follows plays a crucial role in determining its arrival time, with vessels navigating through busier routes expected to experience longer travel durations. Consequently, it is reasonable to infer that each path taken during shipment holds distinct weights that significantly influence arrival times. In this research, we propose a new feature termed "path weight", which serves as an indicator of the congestion level along a vessel's route. Given that our dataset represents vessel paths through discrete segments, we define the segment weight based on the frequency of vessel passages through these segments. As the number of vessels traversing a segment increases, the corresponding segment weight is proportionally amplified. The mathematical representation of segment weight is denoted as Equation (5):

$$w_{segment} = \frac{1}{N}\sum_{i=1}^{N} e_{segment} \tag{5}$$

where N is the total number of samples, and $e_{segment}$ represents one hot encode of the segment, as defined by Equation (4).

Path weight is the total sum of segments if the vessel path is through the segment and is defined by Equation (6):

$$W_{path,i} = \sum_{j=0}^{j=L} W_{segments} * I_{i,j}{}' \tag{6}$$

where L is the total number of segments, and $W_{segments}, I_{I,j}$ is defined by Equation (7).

$$\begin{gathered} W_{segments} = [w_0, w_1, w_2, \ldots..w_L] \\ I_{i,j} = [e_0, e_1, e_2, \ldots..e_L] \end{gathered} \tag{7}$$

The concept of path weight essentially represents the significance of a vessel's trajectory during each trip. As the path weight increases, it signifies a longer travel time for the vessel, indicating that the trajectory traverses through busier segments.

4.2. Module 2: Modeling

Once the raw data have been thoroughly preprocessed, they are fed into the modeling module for further analysis. This module encompasses the training of the model and the identification of optimal hyperparameters, accomplished through a predefined validation strategy. The subsequent paragraphs delve into a more comprehensive explanation of each of these steps.

In this research, the XGBoost (v 2.0.1) algorithm serves as the modeling module for Estimated Time of Arrival (ETA) prediction. XGBoost, a prominent member of the ensemble learning family, is chosen for its potency and widespread application in machine learning. XGBoost combines the strengths of both bagging and boosting techniques, creating a robust and highly accurate model. The algorithm works by iteratively training weak learners, typically decision trees, and boosting their performance by focusing on the mistakes made in previous iterations. It employs a unique regularization term in its objective function, which helps prevent overfitting and enhances generalization [53].

Based on gradient boosting method, XGBoost uses the k additive function to predict the output as expressed by Equation (8).

$$\hat{y}_i = \sum_{k=1}^{K} f_k(X_i), \quad f_k \in F \tag{8}$$

where f_k is an independent Classification and Regression Tree (CART) at each k step that maps input variable X_i to output variable y_i. And F is space of all possible CARTs. The XGBoost algorithm tries to minimize the regularized objective function which is defined in Equation (9):

$$\begin{gathered} Obj = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k) \\ \Omega = \gamma T + \tfrac{1}{2}\lambda \|w\|^2 \end{gathered} \tag{9}$$

where T is number of leaves in the tree, w is the score in corresponding leaves, and γ, λ are regularization coefficients. The regularized objective function comprises two parts. Training loss function l and regularization term Ω. The training loss l indicates the difference between the predicted $(\hat{y}_i)$ and actual (y_i) value. The regularization term shows the complexity of models, which helps the model to avoid overfitting to the dataset.

XGBoost incorporates two key techniques: shrinkage and column subsampling. Shrinkage reduces the impact of each tree by scaling down the weights added in each boosting step, which helps in mitigating overfitting. On the other hand, column subsampling enhances the training speed by selecting a random subset of input features for the construction of each tree.

XGBoost exhibits high sensitivity to its hyperparameters, with an increase in the tree size potentially leading to overfitting issues. Consequently, identifying appropriate hyperparameters is crucial for achieving a well-generalized model. This study attained optimal hyperparameters for the model by assessing its performance on a validation dataset. Employing the stratified cross-validation method, the model's hyperparameters were determined based on its performance in the validation dataset. The dataset is divided into 5 folds, ensuring consistency in the frequency of the day-of-the-week feature across all partitions. Consequently, the model's hyperparameters are estimated to minimize a predefined metric (Root Mean Square Logarithmic Error or RMSLE) on the validation dataset.

Metrics

In this study, five distinct metrics were employed to assess the model's performance. These metrics encompass R^2, Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Mean Absolute Percentage Error (MAPE), and Root Mean Square Logarithmic Error (RMSLE). The mathematical definitions of each of these metrics are provided in Equation (10).

$$\begin{gathered} R^2 = 1 - \frac{\sum_{i=1}^{N}(y(i)-\hat{y}(i))^2}{\sum_{i=1}^{N}(y(i)-\overline{y}(i))^2} \\ RMSE = \sqrt{\frac{\sum_{i=1}^{N}\|y(i)-\hat{y}(i)\|^2}{N}} \\ MAE = \frac{1}{N}\sum_{i=1}^{N}|y(i)-\hat{y}(i)| \\ MAPE = \frac{100}{N}\sum_{i=1}^{N}\left|\frac{y(i)-\hat{y}(i)}{y(i)}\right| \\ RMSLE = \sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(\log\frac{1+\hat{y}(i)}{1+y(i)}\right)^2} \end{gathered} \tag{10}$$

where $y(i)$ is the actual target value, and $\hat{y}(i)$ is the predicted target value for all evaluation metrics. Note that RMSLE is used as a metric to estimate the best hyperparameters of the model.

4.3. Module 3: Apply Model

In this module, the trained model with optimized hyperparameters is applied to the test dataset, and its performance is evaluated using the metrics defined in the Metric section. Additionally, this module includes visualizations to illustrate the efficiency of the model.

5. Experimental Study

The proposed methodology to estimate the time of arrival of vessel was evaluated by considering port of Houston in the United States. Port of Houston was ranked the second busiest seaport in the United States by total tonnage in 2013. The Houston Ship Channel (HSC), spanning 52 miles, is home to approximately 200 private and public industrial terminals. Annually, the HSC facilitates the transportation of over 247 million tons of cargo through the passage of more than 8200 vessels and 215,000 barges [2]. This vital waterway is integral to the support of Texas' energy and petrochemical sectors. The detailed use of the AIS data is summarized in Table 2.

Table 2. Summarized AIS data.

Variable	Description
Data	AIS data of 4330 Cargo and Tankers
Historical Period	January 2018 to April 2020
Features	MMSI, Vessel type, Date, Latitude, Longitude

In this study, data from 2018 were used for model training, and data from 2019 and 2020 were used in testing the model performance. These raw data were first fed to the preprocessing module to extract proper features for model training.

5.1. Preprocessing Module

Within this module, feature extraction, identification of missing values, feature engineering, and the division of data into training and testing sets were carried out. Following the approach suggested by Sedaghat et al. [50,51], the initial features were extracted from the raw data in the first step. Table 3 displays the list of features extracted from the raw data.

Table 3. Extracted initial features from raw AIS data.

Features		
MMSI	Vessel Type	Trip Number
Direction	Date	Location
Segment id		

To apply AIS data to the ETA problem, the initial features need to be transformed into the structure of a complete trip, encompassing a specific origin, destination, and corresponding travel time.

The resulting dataset, formatted with origin and destination information, is subsequently input into the dimension reduction and feature engineering section. This process is undertaken to compute the path weight and the path of vessel segments.

The path weight distribution of the training and test data is shown in Figure 4. As Figure 4 shows, the majority of shipment trips passed through lower-traffic routes. However, there are cases where vessels follow busier routes, leading to increased travel time and, consequently, elevating the significance of the corresponding segments in the estimation of travel time.

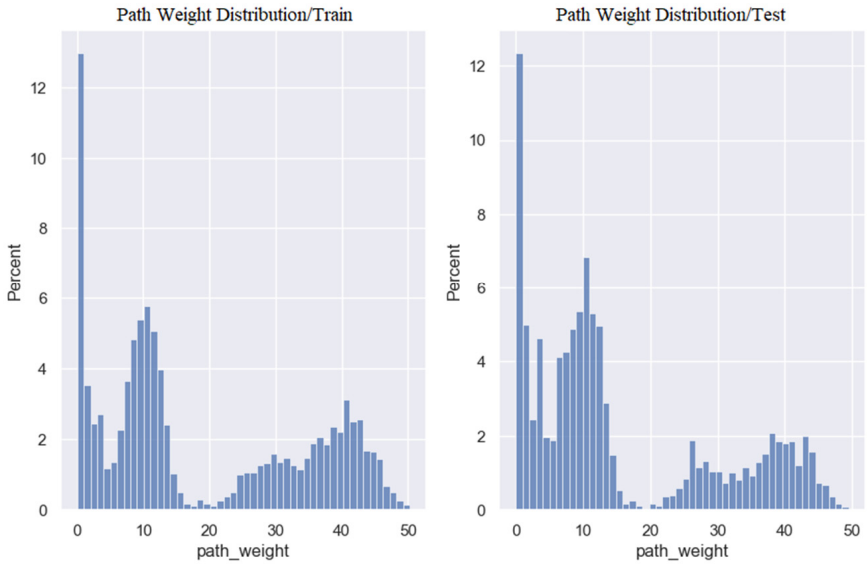

Figure 4. Path weight distribution.

5.2. Modeling Module

5.2.1. Hyperparameter Optimization

As previously mentioned, XGBoost's performance heavily relies on its hyperparameters. Therefore, selecting the right hyperparameters is crucial for enhancing the model's generalization abilities. In this research, the model's optimal hyperparameters were determined based on its validation dataset performance. To identify a suitable validation dataset, a stratified cross-validation approach was employed, dividing the training dataset into five distinct segments. Additionally, to ensure a balanced distribution of trips across all segments, the day of the week was used as the stratification feature. Furthermore, the Root Mean Square Logarithmic Error (RMSLE) was adopted as the metric for optimal parameter identification. Table 4 displays the optimal hyperparameters and their ranges. Our experiments achieved an overall validation score of RMSLE = 0.07.

Table 4. Optimal hyperparameters and their corresponding range.

Hyperparameter	Range	Increment Method	Optimal Value
λ	$(1 \times 10^{-3}, 10)$	Loguniform	0.009
α	$(1 \times 10^{-3}$,10)	Loguniform	1.03
Colsample by tree	(0.3, 1.0)	0.1	0.7
Sub sample	(0.4, 1.0)	0.1	0.4
Learning rate	(0.008, 0.02)	0.001	0.014
Max depth	(10, 80)	10	40
Min child weight	(1, 300)	uniform	5

5.2.2. Model Training

This section is divided into two distinct parts. The first part involves applying a tuned model to the training dataset with all predefined features, followed by an evaluation of its performance using the defined metrics in the Metrics section. In the second part, the effect of reducing the size of the feature is investigated, and the model is reassessed on metrics.

The performance of the tuned model is illustrated in Figure 5. Figure 5 demonstrates that the difference between the training and validation data is sufficiently minimal, indicating that the model is not overfitted to the training data across all folds.

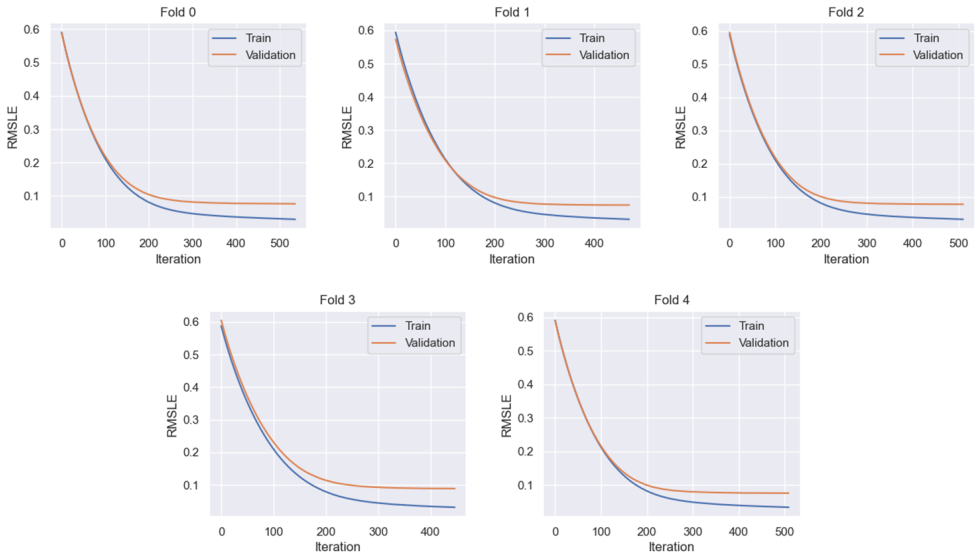

Figure 5. Model performance across folds.

Figure 6 illustrates the Cumulative Density Function of feature importance. The visualization in Figure 6 reveals that the number of features essential for model accuracy is considerably fewer than the total number of features. Our analysis indicates that merely 28% of features, encompassing all the segments traversed by ships, account for 99% of the model's feature importance. Critical among these are path weight, the latitude and longitude of both origin and destination and the busiest segments, which significantly influence the model's performance. Notably, our findings suggest that the date and time of the trip do not significantly impact the accuracy of the model.

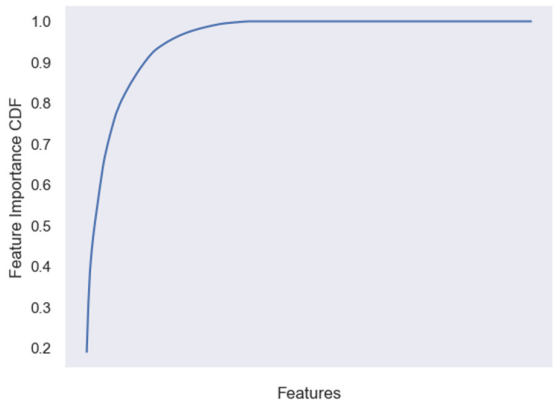

Figure 6. Feature importance cumulative density function.

Table 5 presents a comparison of the performance metrics for both the tuned model and the simplified model. This comparison highlights that removing less significant features has a negligible impact on the overall performance of the model. However, it significantly simplifies the model by reducing the sparsity of the dataset.

Table 5. Tuned full and simplified model accuracies.

Metric	Tuned Full Model	Reduced Model
R^2	0.99	0.98
MAE [min]	6.35	6.81
MSE [min]	95.01	105.21
MAPE	0.05	0.05

5.3. Apply Model

This module assesses the performance of the trained model using a test dataset, which comprises historical AIS data from trips taken during 2019 and 2020. The corresponding performance evaluation is depicted in Figure 7. To construct Figure 7, the travel times of vessels are segmented into 30 min intervals. Within each interval, the average travel time is compared with the distribution of the model's estimated travel times. As indicated by Figure 7, the model's estimated time distribution aligns closely with the average travel time for each interval. Notably, the model exhibits greater accuracy for shorter travel durations. However, for longer trips (longer than 180 min), the figure shows that the travel time estimates are more dispersed around the average, indicating reduced accuracy for these longer journeys.

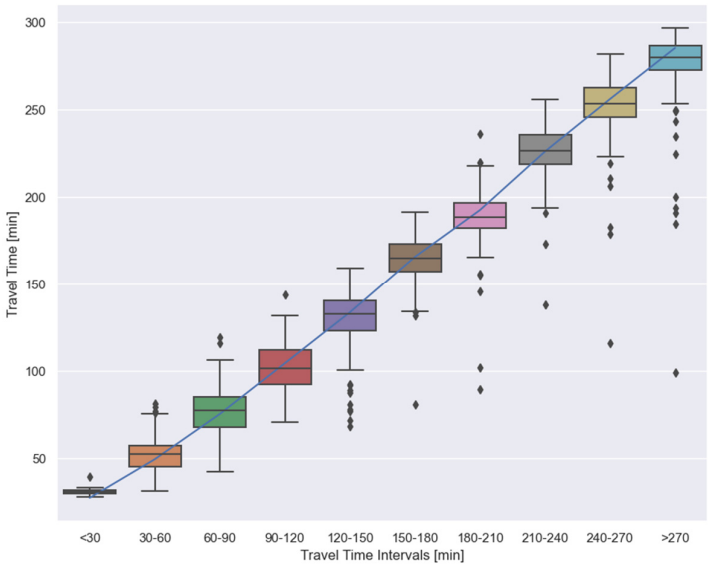

Figure 7. Model accuracy on the test dataset.

Table 6 displays the performance of the model on the test dataset. It demonstrates that both the trained model and the trained simplified model maintain consistent performance across all metrics on the test dataset, comparable to their performance on the training data.

Table 6. Model accuracy on test data.

Metric	Tuned Full Model	Reduced Model
R^2	0.98	0.98
MAE [min]	6.49	6.41
MSE [min]	136.01	127.61
MAPE	0.05	0.05

6. Conclusions

This research presents a comprehensive study on estimated time of arrival (ETA) for vessels in channels and narrow waterways. Using historical Automatic Identification System (AIS) data from 2018 to 2020, this study introduces a machine learning framework to transform raw data, perform feature engineering and preprocessing, and, finally, predict vessel arrival time in channels. The proposed XGBoost model shows high performance across all metrics. The model can predict travel time with only 5% mean absolute error and with 98% R^2. Moreover, the experimental results show that the model can maintain consistent accuracy even using a simplified structure. The less complex model not only preserves accuracy but also offers computational efficiency by addressing the sparsity in

the dataset. The model also shows varying accuracy across different trip durations. While it shows higher precision for shorter trips, its predictions for longer trips (over 180 min) display a wider dispersion around the average values. It should be noted that our study focuses on ETA prediction for vessels from anchorage areas in the ocean to their destination terminals and vice versa. Therefore, a small portion of trips are longer than 180 min. The other area for future research would be model production in real-time and integrating our model results with port and channel operations.

Author Contributions: Conceptualization, M.J.K.; Methodology, H.A.; Software, A.S. and M.J.K.; Validation, A.S.; Formal analysis, A.S. and M.J.K.; Data curation, H.A.; Writing—original draft, H.A.; Writing—review & editing, A.S., M.J.K. and M.H.; Visualization, H.A.; Supervision, M.H.; Project administration, M.H.; Funding acquisition, M.H. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The raw AIS data are extracted from https://marinecadastre.gov/ais/ (accessed on 1 January 2024).

Conflicts of Interest: The authors declare no conflict of interest.

References

1. *Review of Maritime Transport*; United Nations Publications: New York, NY, USA, 2022; ISBN 978-92-1-113073-7.
2. Park, K.; Sim, S.; Bae, H. Vessel Estimated Time of Arrival Prediction System Based on a Path-Finding Algorithm. *Marit. Transp. Res.* **2021**, *2*, 100012. [CrossRef]
3. Kang, M.J.; Hamidi, M. *Quantifying and Predicting Waterway Traffic Conditions: A Case Study of Houston Ship Channel*; Lamar University: Beaumont, TX, USA, 2021.
4. Kabir, M.; Kang, M.J.; Wu, X.; Hamidi, M. Study on U-Turn Behavior of Vessels in Narrow Waterways Based on AIS Data. *Ocean Eng.* **2022**, *246*, 110608. [CrossRef]
5. Cho, Y.; Park, J.; Kim, J.; Kim, J. Autonomous Ship Collision Avoidance in Restricted Waterways Considering Maritime Navigation Rules. *IEEE J. Ocean. Eng.* **2023**, *48*, 1009–1018. [CrossRef]
6. Agafonov, A.; Yumaganov, A. Bus Arrival Time Prediction with LSTM Neural Network. In *Proceedings of the Advances in Neural Networks—ISNN 2019*; Lu, H., Tang, H., Wang, Z., Eds.; Springer International Publishing: Cham, Switzerland, 2019; pp. 11–18.
7. Derrow-Pinion, A.; She, J.; Wong, D.; Lange, O.; Hester, T.; Perez, L.; Nunkesser, M.; Lee, S.; Guo, X.; Wiltshire, B.; et al. ETA Prediction with Graph Neural Networks in Google Maps. In Proceedings of the 30th ACM International Conference on Information & Knowledge Management, Gold Coast, QLD, Australia, 1–5 November 2021; Association for Computing Machinery: New York, NY, USA, 2021; pp. 3767–3776.
8. Carr, G.C.; Erzberger, H.; Neuman, F. Fast-Time Study of Airline-Influenced Arrival Sequencing and Scheduling. *J. Guid. Control Dyn.* **2000**, *23*, 526–531. [CrossRef]
9. Roy, K.; Levy, B.; Tomlin, C. Target Tracking and Estimated Time of Arrival (ETA) Prediction for Arrival Aircraft. In Proceedings of the AIAA Guidance, Navigation, and Control Conference and Exhibit; Guidance, Navigation, and Control and Co-located Conferences, Keystone, CO, USA, 21–24 August 2006; American Institute of Aeronautics and Astronautics: Reston, VA, USA, 2012.
10. Lim, A.; Rodrigues, B.; Zhu, Y. Airport Gate Scheduling with Time Windows. *Artif. Intell. Rev.* **2005**, *24*, 5–31. [CrossRef]
11. Narciso, M.E.; Piera, M.A. Robust Gate Assignment Procedures from an Airport Management Perspective. *Omega* **2015**, *50*, 82–95. [CrossRef]
12. Yang, Z.; Wang, Y.; Li, J.; Liu, L.; Ma, J.; Zhong, Y. Airport Arrival Flow Prediction considering Meteorological Factors Based on Deep-Learning Methods. *Complexity* **2020**, *2020*, 6309272. [CrossRef]
13. Ayhan, S.; Costas, P.; Samet, H. Predicting Estimated Time of Arrival for Commercial Flights. In Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining, London, UK, 19–23 August 2018; Association for Computing Machinery: New York, NY, USA; pp. 33–42.
14. Karbassi, A.; Barth, M. Vehicle Route Prediction and Time of Arrival Estimation Techniques for Improved Transportation System Management. In Proceedings of the IEEE IV2003 Intelligent Vehicles Symposium. Proceedings (Cat. No.03TH8683), Columbus, OH, USA, 9–11 June 2003; pp. 511–516.
15. Mazloumi, E.; Rose, G.; Currie, G.; Sarvi, M. An Integrated Framework to Predict Bus Travel Time and Its Variability Using Traffic Flow Data. *J. Intell. Transp. Syst.* **2011**, *15*, 75–90. [CrossRef]
16. Achar, A.; Bharathi, D.; Kumar, B.A.; Vanajakshi, L. Bus Arrival Time Prediction: A Spatial Kalman Filter Approach. *IEEE Trans. Intell. Transp. Syst.* **2020**, *21*, 1298–1307. [CrossRef]

17. Propp, D.A.; Rosenberg, C.A. A Comparison of Prehospital Estimated Time of Arrival and Actual Time of Arrival to an Emergency Department. *Am. J. Emerg. Med.* **1991**, *9*, 301–303. [CrossRef]
18. Fleischman, R.J.; Lundquist, M.; Jui, J.; Newgard, C.D.; Warden, C. Predicting Ambulance Time of Arrival to the Emergency Department Using Global Positioning System and Google Maps. *Prehospital Emerg. Care* **2013**, *17*, 458–465. [CrossRef]
19. Shmueli, G.; Bruce, P.C.; Patel, N.R. *Data Mining for Business Analytics: Concepts, Techniques, and Applications with XLMiner*; Wiley: Hoboken, NJ, USA, 2016; ISBN 9781118729137.
20. Pani, C.; Vanelslander, T.; Fancello, G.; Cannas, M. Prediction of Late/Early Arrivals in Container Terminals—A Qualitative Approach. *Eur. J. Transp. Infrastruct. Res.* **2015**, *15*, 536–550. [CrossRef]
21. Hart, P.E.; Nilsson, N.J.; Raphael, B. A Formal Basis for the Heuristic Determination of Minimum Cost Paths. *IEEE Trans. Syst. Sci. Cybern.* **1968**, *4*, 100–107. [CrossRef]
22. Shin, Y.W.; Abebe, M.; Noh, Y.; Lee, S.; Lee, I.; Kim, D.; Bae, J.; Kim, K.C. Near-Optimal Weather Routing by Using Improved A* Algorithm. *Appl. Sci.* **2020**, *10*, 6010. [CrossRef]
23. Alessandrini, A.; Mazzarella, F.; Vespe, M. Estimated Time of Arrival Using Historical Vessel Tracking Data. *IEEE Trans. Intell. Transp. Syst.* **2018**, *20*, 7–15. [CrossRef]
24. Chen, X.; Wang, M.; Ling, J.; Wu, H.; Wu, B.; Li, C. Ship Imaging Trajectory Extraction Via an Aggregated you only Look once (YOLO) Model. *Eng. Appl. Artif. Intell.* **2024**, *130*, 107742. [CrossRef]
25. Wu, X.; Roy, U.; Hamidi, M.; Craig, B.N. Estimate Travel Time of Ships in Narrow Channel Based on AIS Data. *Ocean Eng.* **2020**, *202*, 106790. [CrossRef]
26. Wu, X.; Mehta, A.L.; Zaloom, V.A.; Craig, B.N. Analysis of Waterway Transportation in Southeast Texas Waterway Based on AIS Data. *Ocean Eng.* **2016**, *121*, 196–209. [CrossRef]
27. Wu, X.; Rahman, A.; Zaloom, V.A. Study of Travel Behavior of Vessels in Narrow Waterways Using AIS Data—A CASE study in Sabine-Neches Waterways. *Ocean Eng.* **2018**, *147*, 399–413. [CrossRef]
28. Roy, U.; Wu, X. Ais-data based vessel traffic's characteristics and travel behaviour analysis: A case study at houston ship channel. *J. Ocean Technol.* **2019**, *14*, 58–74.
29. Kang, L.; Meng, Q.; Liu, Q. Fundamental Diagram of Ship Traffic in the Singapore Strait. *Ocean Eng.* **2018**, *147*, 340–354. [CrossRef]
30. Ghanim, M.S.; Shaaban, K.; Miqdad, M. An Artificial Intelligence Approach to Estimate Travel Time along Public Transportation Bus Lines. In Proceedings of the International Conference on Civil Infrastructure and Construction, Doha, Qatar, 2–5 February 2023.
31. Liu, H.; Xu, H.; Yan, Y.; Cai, Z.; Sun, T.; Li, W. Bus Arrival Time Prediction Based on LSTM and Spatial-Temporal Feature Vector. *IEEE Access* **2020**, *8*, 11917–11929. [CrossRef]
32. Petersen, N.C.; Rodrigues, F.; Pereira, F.C. Multi-Output Bus Travel Time Prediction with Convolutional LSTM Neural Network. *Expert Syst. Appl.* **2019**, *120*, 426–435. [CrossRef]
33. Ranjitkar, P.; Tey, L.-S.; Chakravorty, E.; Hurley, K.L. Bus Arrival Time Modeling Based on Auckland Data. *Transp. Res. Rec. J. Transp. Res. Board* **2019**, *2673*, 1–9. [CrossRef]
34. Larsen, G.H.; Yoshioka, L.R.; Marte, C.L. Bus Travel Times Prediction Based on Real-Time Traffic Data Forecast Using Artificial Neural Networks. In Proceedings of the 2020 International Conference on Electrical, Communication, and Computer Engineering (ICECCE), Istanbul, Turkey, 12–13 June 2020; pp. 1–6.
35. Alam, O.; Kush, A.; Emami, A.; Pouladzadeh, P. Predicting Irregularities in Arrival Times for Transit Buses with Recurrent Neural Networks Using GPS Coordinates and Weather Data. *J. Ambient. Intell. Humaniz. Comput.* **2020**, *12*, 7813–7826. [CrossRef]
36. Chondrodima, E.; Georgiou, H.; Pelekis, N.; Theodoridis, Y. Particle Swarm Optimization and RBF Neural Networks for Public Transport Arrival Time Prediction Using GTFS Data. *Int. J. Inf. Manag. Data Insights* **2022**, *2*, 100086. [CrossRef]
37. Milenković, M.; Milosavljevic, N.; Bojović, N.; Val, S. Container Flow Forecasting through Neural Networks Based on Metaheuristics. *Oper. Res.* **2021**, *21*, 965–997. [CrossRef]
38. Kourounioti, I.; Polydoropoulou, A.; Tsiklidis, C. Development of Models Predicting Dwell Time of Import Containers in Port Container Terminals—An Artificial Neural Networks Application. *Transp. Res. Procedia* **2016**, *14*, 243–252. [CrossRef]
39. Gao, M.; Shi, G.; Li, S. Online Prediction of Ship Behavior with Automatic Identification System Sensor Data Using Bidirectional Long Short-Term Memory Recurrent Neural Network. *Sensors* **2018**, *18*, 4211. [CrossRef]
40. Kim, J.; Lee, C.; Chung, D.; Kim, J. Navigable Area Detection and Perception-Guided Model Predictive Control for Autonomous Navigation in Narrow Waterways. *IEEE Robot. Autom. Lett.* **2023**, *8*, 5456–5463. [CrossRef]
41. Fancello, G.; Pani, C.; Pisano, M.; Serra, P.; Zuddas, P.; Fadda, P. Prediction of Arrival Times and Human Resources Allocation for Container Terminal. *Marit. Econ. Logist.* **2011**, *13*, 142–173. [CrossRef]
42. Pani, C.; Cannas, M.; Fadda, P.; Fancello, G.; Frigau, L.; Mola, F. Delay Prediction in Container Terminals: A Comparison of Machine Learning Methods. In Proceedings of the WCTR (World Conference on Transport Research), Rio de Janeiro, Brazil, 15–18 July 2013.
43. Pani, C.; Fadda, P.; Fancello, G.; Frigau, L.; Mola, F. A data mining approach to forecast late arrivals in a transhipment container terminal. *Transport* **2014**, *29*, 175–184. [CrossRef]
44. Pallotta, G.; Vespe, M.; Bryan, K. Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction. *Entropy* **2013**, *15*, 2218–2245. [CrossRef]

45. Parolas, I.; Tavasszy, L.; Kourounioti, I.; van Duin, R.; Cities, K. Prediction of Vessels' Estimated Time of Arrival (ETA) Using Machine Learning–a Port of Rotterdam Case Study. In Proceedings of the 96th Annual Meeting of the Transportation Research, Washington, DC, USA, 8–12 January 2017; pp. 8–12.
46. Noman, A.A.; Heuermann, A.; Wiesner, S.A.; Thoben, K.-D. Towards Data-Driven GRU based ETA Prediction Approach for Vessels on both Inland Natural and Artificial Waterways. In Proceedings of the 2021 IEEE International Intelligent Transportation Systems Conference (ITSC), Indianapolis, IN, USA, 19–22 September 2021; pp. 2286–2291.
47. Yu, J.; Tang, G.; Song, X.; Yu, X.; Qi, Y.; Li, D.; Zhang, Y. Ship Arrival Prediction and Its Value on Daily Container Terminal Operation. *Ocean Eng.* **2018**, *157*, 73–86. [CrossRef]
48. Notteboom, T.E. The Time Factor in Liner Shipping Services. *Marit. Econ. Logist.* **2006**, *8*, 19–39. [CrossRef]
49. Available online: https://marinecadastre.gov/ (accessed on 1 January 2024).
50. Sedaghat, A.; Kang, M.J.; Hamidi, M. A Heuristic ETL Process to Dynamically Separate and Compress AIS Data. In Proceedings of the 2023 Systems and Information Engineering Design Symposium (SIEDS), Charlottesville, VA, USA, 27–28 April 2023; pp. 159–164.
51. Sedaghat, A.; Arbabkhah, H.; Kang, M.J.; Hamidi, M. Deep Learning Applications in Vessel Dead Reckoning to Deal with Missing Automatic Identification System Data. *J. Mar. Sci. Eng.* **2024**, *12*, 152. [CrossRef]
52. Wold, S.; Esbensen, K.; Geladi, P. Principal component analysis. *Chemom. Intell. Lab. Syst.* **1987**, *2*, 37–52. [CrossRef]
53. Chen, T.; Guestrin, C. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, 13–17 August 2016; Association for Computing Machinery: New York, NY, USA; pp. 785–794.