Cyprus University of Technology
Department of Electrical Engineering and Computer Engineering and Informatics

Doctoral Dissertation

Exploiting AIS Data for Predicting Vessel Arrival Times and Trajectories: An Artificial Intelligence Approach

Nicos Evmides

Limassol, 2025

CYPRUS UNIVERSITY OF TECHNOLOGY

FACULTY OF ENGINEERING AND TECHNOLOGY

DEPARTMENT OF ELECTRICAL ENGINEERING,

COMPUTER ENGINEERING AND INFORMATICS

Doctoral Dissertation

Exploiting AIS Data for Predicting Vessel Arrival Times and Trajectories: An Artificial Intelligence Approach

Nicos Evmides

Supervisor:

Assoc. Prof. Herodotos Herodotou

Limassol, 2025

Advisory Committee

Doctoral Dissertation

Exploiting AIS Data for Predicting Vessel Arrival Times and Trajectories: An Artificial Intelligence Approach

Presented by

Nicos Evmides

Supervisor: Herodotos Herodotou, Associate Professor, Cyprus University of Technology

Member of the committee: Alexander Artikis, Associate Professor, University of Piraeus

Member of the committee: Andreas Andreou, Professor, Cyprus University of Technology

Cyprus University of Technology
Limassol, July 2025

Copyrights

The approval of the dissertation by the Department of Electrical Engineering, Computer Engineering and Informatics does not necessarily imply the approval by the Department of the views of the writer.

Acknowledgements

I would like to thank my professors at Cyprus University of Technology, Assoc. Prof. Herodotos Herodotou and Assoc. Prof. Michalis Michaelides, for guiding me through the process of my PhD research, providing valuable insights and resources to complete my studies. Furthermore, I would like to thank all my co-authors and team members for their hard work to complete the research work referenced in this doctoral dissertation.

ABSTRACT

Automatic Identification System (AIS) data has transformed maritime operations over the past two decades by enabling real-time situational awareness, vessel tracking, and collision avoidance. Mandated by the International Maritime Organization (IMO) since 2000, AIS systems automatically broadcast vital navigational and vessel-related information, such as position, speed, course, heading, vessel type, and destination, to nearby ships and coastal authorities. This dissertation presents a unified framework for the intelligent collection, processing, storage, and analysis of AIS data, addressing both the scalability and reliability challenges associated with real-time maritime data streams.

Building upon real-world deployments in the Eastern Mediterranean, the research contributes a suite of integrated algorithms and services designed to enhance maritime informatics. The proposed framework consists of three core components: AIS data cleaning, vessel Estimated Time of Arrival (ETA) prediction, and vessel trajectory forecasting. First, the system applies advanced data preprocessing techniques to remove redundancies, resolve inconsistencies, and impute missing values in AIS messages. Second, machine learning models are trained to provide accurate, dynamic ETA predictions, allowing ports and shipping stakeholders to optimize resource allocation and berth scheduling and reduce vessel idle times. Third, deep learning-based sequence models are used to forecast vessel trajectories over short and long horizons, supporting proactive traffic management, navigational safety, and operational planning.

By integrating these capabilities, this work advances the development of intelligent, scalable, and environmentally aligned maritime decision-support systems. The findings have practical implications for port authorities, shipping companies, and regulatory bodies, promoting safer, more efficient, and sustainable maritime logistics.

Keywords: Automatic Identification System (AIS), Maritime Informatics, Machine Learning, ETA Prediction, Trajectory Forecasting, Port Optimization, Data Cleaning, Deep Learning, Vessel Traffic Management

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ABBREVIATIONS

Important abbreviations that have been used in the text and need explanation are briefly presented.

ADE	Average Displacement Error
AIS	Automatic Identification System
ANN	Artificial Neural Network
ATA	Actual Time of Arrival
ATB	Actual Time of Berthing
ATD	Actual Time of Departure
ATUB	Actual Time of Unberthing
Bi-GRU	Bidirectional Gated Recurrent Unit
Bi-LSTM	Bidirectional Long Short-Term Memory
CNN	Convolutional Neural Network
CoG	Course over Ground
CUT	Cyprus University of Technology
DL	Deep Learning
DFD	Discrete Fréchet Distance
DNN	Deep Neural Network
DT	Decision Tree
EM	Expectation Maximization
ETA	Estimated Time of Arrival
ETD	Estimated Time of Departure
EV	Explained Variance
FDE	Final Displacement Error
GNB	Gaussian Naive Bayes
GP	Gaussian Process
GRU	Gated Recurrent Unit
IMO	International Maritime Organization
KNN	K-Nearest Neighbours
LSSVM	Least Squares Support Vector Machine
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percent Error
MDA	Maritime Domain Awareness
MDTC	Minimum Distance To Collision
ML	Machine Learning
MMSI	Maritime Mobile Service Identity
MC	Markov Chains
MSE	Mean Squared Error
R^2	Coefficient of Determination
RF	Random Forest

RMSE	Root Mean Squared Error
RNN	Recurrent Neural Network
RoT	Rate of Turn
SMAPE	Symmetric Mean Absolute Percentage Error
SoG	Speed over Ground
TSS	Traffic Separation Scheme
XGBoost	Extreme Gradient Boosting

LIST OF PUBLICATIONS

1. N. Evmides, S. Aslam, A. Televantos, A. Karagiannis, A. Paraskeva, M. Michaelides, and H. Herodotou.
   Employing Fuzzy Matching for Cleaning Manual AIS Entries.
   In Proceedings of the 2021 World of Shipping Portugal: An International Research Conference on Maritime Affairs
   Carcavelos, Portugal, January 28–29 2021.

2. A. Worth, A. Televantos, N. Evmides, M. Michaelides, and H. Herodotou.
   Online Analytical Processing of Port Calls for Decision Support.
   In Proceedings of the 23rd IEEE International Conference on Mobile Data Management (MDM), Paphos, Cyprus, June 6–9, 2022.
   DOI: 10.1109/MDM55031.2022.00095

3. N. Evmides, L. Odysseos, M. Michaelides, and H. Herodotou
   An Intelligent Framework for Vessel Traffic Monitoring Using AIS Data.
   In Proceedings of the 23rd IEEE International Conference on Mobile Data Management (MDM), Paphos, Cyprus, June 6–9, 2022, pp. 413–418.
   DOI: 10.1109/MDM55031.2022.00091

4. N. Evmides, S. Aslam, T. Ramez, M. Michaelides, and H. Herodotou.
   Enhancing Prediction Accuracy of Vessel Arrival Times Using Machine Learning.
   Journal of Marine Science and Engineering, 12(8), 1362, 2024.
   DOI: 10.3390/jmse12081362

5. N. Evmides, M. Michaelides, and H. Herodotou
   Vessel Trajectory Prediction with Deep Learning: Temporal Modeling and Operational Implications.
   Journal of Marine Science and Engineering, 13(8), 1439, 2025.
   DOI: 10.3390/jmse13081439

1 Introduction

Maritime transport accounts for more than 90% of the world's trade, while 74% of all goods imported or exported from Europe are carried by ships [3–6]. In 2022, container ports handled a total weight of more than 12,027 million metric tonnes around the globe, and it is expected to grow by 2.4% until 2028 [2], highlighting the importance of timely and efficient maritime transport.

The maritime domain is undergoing a digital transformation, driven by the proliferation of vessel tracking data and rapid advances in machine learning and data engineering. At the heart of this transformation lies the Automatic Identification System (AIS), a global standard mandated by the International Maritime Organization (IMO) for vessels exceeding specific gross tonnage thresholds [1]. AIS broadcasts contain information such as vessel identity, position, course, speed, and destination, transmitted at regular intervals over unencrypted very high frequency (VHF) signals. These transmissions not only enhance navigational safety through real-time traffic awareness but also provide a rich, longitudinal dataset that has become central to maritime analytics and operational decision-making [7].

This research work leverages AIS data to tackle several core challenges in maritime traffic monitoring, vessel arrival time prediction, and trajectory forecasting. Through the integration of scalable data engineering pipelines, intelligent learning models, and robust experimental validation, this work develops a unified framework to support real-time maritime situational awareness and decision-making.

An essential component of this framework is the real-time acquisition and processing of high-volume AIS data. A distributed architecture was developed to handle the high frequency and irregularity of AIS messages, addressing challenges such as duplication, timestamp drift, and missing data. Deployed operationally in the Eastern Mediterranean, this system has successfully ingested and processed over one billion AIS messages, providing the foundational infrastructure for higher-level analytics.

One of the most problematic fields in AIS messages is the manually entered destination field, which is often noisy, inconsistent, or entirely missing. To address this, a domain-specific fuzzy matching algorithm was designed and implemented using a combination of approximate string matching techniques and structured matching rules. This method achieved a high match rate on a large number of AIS messages, with unmatched cases mostly involving abbreviations, unofficial names, or corrupted entries. By integrating validated human-corrected matches into a growing mapping table, the system evolves over time and reduces computational overhead in future processing. This adaptive approach to semantic cleaning enhances the integrity of the destination field, which is critical for downstream analytics such as ETA prediction and traffic pattern analysis.

Building on this cleaned and structured data, the PhD Dissertation proposes a machine learning framework for estimating vessels' estimated time of arrival (ETA) at their respective ports. The modeling approach employs route-based data segmentation to preserve the spatial and temporal characteristics of trajectories, and it evaluates multiple algorithms—including deep neural networks, decision trees, and ensemble models—on a real-world dataset collected via coastal AIS base stations. Ensemble techniques such as Random Forest and XGBoost were shown to outperform others in both accuracy and robustness, providing meaningful improvements in port scheduling, vessel turnaround, and emission reduction.

To extend the operational utility of AIS analytics further, the PhD Dissertation also investigates the problem of vessel trajectory forecasting. Deep recurrent neural networks, specifically LSTM-based architectures, are employed to model sequential patterns in vessel movement. The experiments demonstrate that bidirectional LSTM networks perform particularly well in capturing complex spatiotemporal dependencies, especially for longer input sequences. However, the study also notes challenges associated with long-horizon forecasts, such as accumulated prediction errors and directional drift, suggesting a need for hybrid models or attention-based mechanisms in future implementations.

Together, these contributions form a vertically integrated framework that spans the entire AIS data lifecycle, ranging from real-time ingestion and semantic cleaning to predictive analytics for port operations and future-state navigation. Methodologically, the work combines data engineering with classical and deep learning paradigms, while practically, it addresses the pressing needs of port authorities, logistics providers, and regulatory bodies navigating an increasingly complex maritime environment.

The challenges of modern maritime logistics, characterized by congested sea lanes, growing vessel sizes, and tightening environmental regulations, require intelligent systems that are accurate, scalable, and adaptive. This PhD research meets that need by advancing the state of the art in AIS-based analytics, offering both foundational insights and deployable tools. The resulting systems support safer and more efficient maritime traffic management while aligning with broader goals of sustainability and digital transformation in the maritime sector.

1.1 Contributions

This doctoral research makes a series of novel and interrelated contributions to the field of maritime informatics, with a particular focus on the collection, semantic cleaning, and intelligent exploitation of AIS data. The contributions span systems design, data engineering, algorithmic development, and predictive modeling for operational applications in port and vessel management. Specifically, the thesis offers the following key contributions:

- **Design of a Scalable AIS Data Framework**: Proposes an end-to-end architecture for real-time AIS data ingestion, enrichment, and storage. The framework supports the scalable processing of high-volume AIS streams and provides a robust foundation for advanced maritime analytics and decision-support systems.
- **Development of Intelligent Data Cleaning Algorithms**: Introduces a comprehensive suite of algorithms for AIS message de-duplication, handling of missing or noisy values, and semantic correction of manually entered fields—most notably the vessel destination field. A novel fuzzy matching algorithm is developed and evaluated, employing structured domain-specific rules. This adaptive, rule-based fuzzy logic mechanism not only corrects inconsistent or abbreviated entries but also evolves over time through human-in-the-loop reinforcement, improving data quality for downstream tasks.
- **Implementation of Real-Time Maritime Services**: Designs and integrates intelligent services that utilize cleaned and enriched AIS data for real-time vessel tracking, maritime traffic visualization, anomaly detection, and collision risk estimation. These services demonstrate the operational utility of the framework in live maritime environments.

- **Enhanced ETA Prediction Using Machine Learning**: Conducts a comprehensive evaluation of six machine learning models–Deep Neural Networks (DNN), K-Nearest Neighbors (KNN), Decision Trees (DT), Random Forest (RF), XGBoost, and Gaussian Naive Bayes (GNB)–for predicting vessel ETA using cleaned and semantically validated AIS data. A route-based data segmentation methodology is proposed to improve model generalizability and reduce overfitting.
- **Feature Engineering for ETA Prediction**: Generates and systematically evaluates a diverse range of feature combinations derived from AIS messages, identifying optimal sets of input variables that enhance predictive accuracy across vessel types, voyage durations, and regional settings.
- **Deep Learning for Vessel Trajectory Forecasting**: Evaluates and benchmarks the performance of LSTM, Bi-LSTM, and Bi-GRU architectures for both short-term and long-term vessel trajectory forecasting. The findings highlight Bi-LSTM's superior ability to capture bidirectional temporal dependencies and maintain spatial coherence over extended forecast windows.
- **Empirical Insights into Model Robustness and Operational Limitations**: Provides a systematic assessment of degradation patterns in long-horizon predictions, model explainability, and implementation challenges in operational environments. This includes considerations for handling ambiguous or semantically overlapping destination entries and strategies for integrating adaptive learning loops into maritime decision systems.

Beyond these technical contributions, the research demonstrates how predictive intelligence can be operationalized in maritime contexts. Accurate ETA forecasting supports port authorities and logistics operators in optimizing berth allocation, crane scheduling, and tug dispatching, resulting in improved turnaround times and resource efficiency. Similarly, robust trajectory forecasting enhances safety by supporting early collision avoidance and route coordination in high-traffic areas.

The semantic cleaning of AIS data, especially the vessel destination field through fuzzy matching, also proves critical to the accuracy and reliability of all downstream analytics. By transforming messy, manually entered AIS fields into structured, machine-readable targets, the fuzzy matching component enables the effective use of machine learning and forecasting models at scale.

Importantly, the predictive services developed in this PhD research contribute to sustainability goals by reducing unnecessary emissions from vessels idling near ports and improving overall fleet efficiency. As such, the outcomes of this research not only advance academic understanding in maritime data systems and machine learning but also offer actionable solutions for building smarter, safer, and greener maritime infrastructures.

1.2 PhD Dissertation Structure

The remainder of this thesis is structured as follows:

- **Chapter 2** provides a comprehensive literature review of research in AIS-based maritime analytics, focusing on four core areas: frameworks, data quality management, vessel ETA prediction, and trajectory forecasting. It synthesizes findings from traditional approach models to advanced machine learning and deep learning techniques, highlighting key limitations in data integrity, model scalability, and operational deployment. The review identifies gaps in existing approaches to AIS data

cleaning and emphasizes the need for robust, adaptive methods to improve the semantic quality of critical AIS attributes. These insights serve as the foundation for the methodological innovations introduced in subsequent chapters.

- **Chapter 3** introduces an intelligent and scalable framework for real-time vessel traffic monitoring using AIS data. It outlines the system architecture, data ingestion mechanisms, preprocessing algorithms for cleaning and de-duplicating AIS messages, and the deployment of foundational services such as collision detection, traffic visualization, and area-of-interest monitoring.
- **Chapter 4** focuses on improving the semantic quality of AIS data through the development and evaluation of a fuzzy matching algorithm for cleaning manually entered vessel destination fields. It presents a rule-based fuzzy matching approach capable of handling abbreviations, misspellings, and non-standard formats. The chapter also examines error patterns and proposes strategies for incorporating human feedback and adaptive learning to improve long-term performance.
- **Chapter 5** presents a machine learning-based methodology for accurate vessel Estimated Time of Arrival (ETA) prediction. It evaluates a variety of algorithms, including Random Forests, XGBoost, and Artificial Neural Networks, while proposing novel feature engineering and route-based data partitioning strategies to mitigate overfitting and improve model generalizability.
- **Chapter 6** investigates deep learning techniques for vessel trajectory forecasting. It offers a comparative study of recurrent neural architectures (LSTM, Bi-LSTM, Bi-GRU) across short- and long-term prediction horizons, analyzing their performance degradation dynamics, spatiotemporal coherence, and potential for operational deployment.
- **Chapter 7** provides an integrated discussion of the results across all chapters. It synthesizes key insights, discusses methodological limitations, and reflects on the broader implications of the findings for maritime operations, sustainability, and decision-support systems.
- **Chapter 8** concludes the thesis by summarizing the main contributions, reinforcing the value of combining real-time AIS data with machine learning, deep learning, and data cleaning techniques. It also outlines future research directions, including transformer-based sequence models, automated retraining pipelines, and integration with external contextual data (e.g., weather, sea state, port congestion).

This structure reflects the interdisciplinary nature of the research, bridging maritime informatics, real-time systems, and predictive analytics to support intelligent, adaptive, and sustainable decision-making in the maritime sector.

2 Literature Review

The increasing availability of Automatic Identification System (AIS) data has enabled a wide array of innovations in maritime analytics, including situational awareness, vessel behavior modeling, port operation optimization, and maritime safety. This chapter reviews the current state of the literature in four core areas: maritime traffic monitoring frameworks, AIS data cleaning methodologies, machine learning approaches to vessel ETA prediction, and vessel trajectory forecasting using deep learning. In each domain, key contributions are examined, and limitations and gaps in the existing body of work are identified to provide direction for further study.

2.1 AIS-Driven Maritime Monitoring Frameworks

AIS data plays a central role for maritime surveillance, enabling applications spanning traffic monitoring, route analysis, destination prediction, and anomaly detection [4]. Early contributions focused on scalable data handling and infrastructure. Fu et al. [8] proposed a computational framework for compressing and efficiently transferring large-scale AIS data streams, facilitating subsequent analytics. Building on this, Vodas et al. [9] developed Maritime Situational Awareness (MSA) systems that incorporate operational event detection—such as illegal fishing and route deviations—while using data synopsis techniques to enhance scalability in processing high-volume AIS data.

The datAcron platform [10] further advanced AIS analytics by supporting streaming data processing, trajectory reconstruction, and data enrichment with external sources like weather and environmental context. Santipantakis et al. [11] introduced link discovery methods for contextual inference, enabling detection of suspicious vessel behaviors in real time. Bernabe et al. [12] applied self-supervised Transformer models to detect intentional AIS shutdowns with real-time capability. Maganaris et al. [13] proposed RNN encoder–decoder architectures for detecting anomalous motion patterns in AIS data, while Liang et al. [14] combined trajectory-to-image transformation with deep latent models for unsupervised anomaly detection.

Unsupervised learning approaches have also been used to model typical vessel behavior. Pallotta et al. [15] applied clustering to identify waypoints and construct route models from historical AIS data, and Sheng et al. [16] proposed trajectory clustering methods based on structural similarity. More recently, Jain et al.[17] introduced an unsupervised clustering methodology for marine vessel trajectories using a large historical AIS database, demonstrating improvements in pattern discovery and maritime traffic understanding. Additionally, AIS data has been utilized for collision risk assessment, by analyzing relative vessel positions, headings, and speeds [18, 19], as well as for broader situational awareness and environmental hazard detection [20, 21].

Despite these advances, many existing AIS frameworks still face challenges in handling the high volume, velocity, and noise inherent in real-time AIS streams. Several systems emphasize retrospective or batch analysis rather than fully integrated real-time monitoring with predictive capabilities. Moreover, few frameworks incorporate adaptive learning or robust data cleaning methods essential for downstream applications like ETA prediction and trajectory forecasting.

Our proposed framework addresses these gaps by providing a scalable, distributed architecture for ingestion, cleaning, deduplication, and semantic enrichment of AIS data. It supports near real-time analytics and foundational services such as collision warnings and area monitoring, creating a reliable data backbone for advanced machine learning models presented in later chapters. This work thus contributes to bridging the divide between raw AIS data acquisition and operationally relevant predictive analytics.

2.2 AIS Data Cleaning and Semantic Field Standardization

AIS data quality presents a critical challenge for downstream applications. Although much effort has focused on trajectory reconstruction and outlier removal, relatively few studies address semantic data cleaning, particularly in manually entered fields such as port destination. This field is prone to inconsistencies due to free-text inputs, abbreviations, misspellings, and language variations.

Abdallah et al. [22] addressed this issue using an edit distance-based matching algorithm to align raw AIS destination entries with a standardized port list. However, the method achieved only moderate success (58% accuracy) and was tested on a relatively small dataset. Other relevant studies, such as those by Tichavska et al. [23] and Wijaya et al. [24], focus on behavioral analysis using AIS, without addressing semantic standardization.

Few existing methodologies incorporate contextual or frequency-based features, and even fewer allow for adaptive learning from corrected records. This gap presents a challenge for research that depends on high-quality semantic AIS attributes, such as port-level arrival time prediction or berth scheduling.

2.3 Machine Learning Approaches to Vessel ETA Prediction

Accurate estimation of a vessel's estimated time of arrival (ETA) is fundamental for efficient port operation and logistics. Several studies have proposed data-driven methods that utilize historical AIS records, environmental factors, and port traffic conditions.

Park et al. [25] proposed a model that integrates Markov Chains and Bayesian sampling for trajectory and ETA prediction. Parolas et al. [26] compared artificial neural networks (ANNs), support vector machines (SVMs), and multi-linear regression. Their findings indicate that while regression models require minimal training, they struggle with complex mappings; in contrast, ANNs can model such complexities at the cost of increased training time and data requirements.

Incorporating environmental features, Ogura et al. [27] employed the Dijkstra algorithm to determine voyage routes, combining this with Bayesian sampling for ETA estimation under dynamic weather conditions. Other approaches include Alessandrini et al. [28], who proposed a hybrid method utilizing historical and real-time AIS data with a custom path-finding algorithm, and El et al. [29], who developed an ANN-based ETA model for long-range vessel trajectories.

Yoon et al. [30] introduced a historical voyage-based segmentation approach using spline interpolation, achieving fine-grained ETA estimation with real port data. Ensemble learning and gradient boosting have also been explored: Arbabkhah et al. [31] utilized XGBoost for ETA prediction in narrow waterways, demonstrating robust performance with minimal feature sets.

While these models vary in scope and methodology, many rely on static training data and lack provisions for real-time updates or adaptation to evolving vessel behaviors and traffic patterns. Moreover, most studies evaluate performance using aggregate metrics such as MAE or RMSE, with limited insight into temporal degradation or operational robustness.

2.4 Vessel Trajectory Forecasting with Deep Learning

Trajectory forecasting is essential for maritime safety and efficiency, especially in confined environments such as port areas. Early work employed rule-based or density mapping techniques. For instance, Tun et al. [32] applied density maps to identify motion patterns within port limits, while Rhodes et al. [33] used adaptive neural networks to cluster vessel speeds and headings around fixed reference points.

Recent developments in deep learning have led to the adoption of temporal models such as recurrent neural networks (RNNs), including LSTM, GRU, and Bi-GRU variants [34–36]. Chondrodima et al.[36] proposed an efficient LSTM-based framework for vessel location forecasting that combines grid-based spatial preprocessing with LSTM networks to handle irregular and sparse AIS data, improving both prediction accuracy and computational efficiency. Their approach also incorporates feature engineering techniques to better represent vessel movement patterns in maritime environments. These models are capable of capturing temporal dependencies and nonlinear motion patterns using AIS time series. Capobianco et al. [37] extended their LSTM framework with Bayesian modeling to incorporate prediction uncertainty. Grid-based preprocessing and spatial embedding schemes have also been proposed to address data sparsity in irregular AIS samples [36].

Hybrid architectures have also emerged. Wang et al. [38] combined LSTM with Kalman filters to account for observational noise and sequential updates. Wu et al. [39] proposed a multi-branch hybrid model incorporating convolutional, temporal convolutional, and ConvLSTM networks to simultaneously capture local and long-range dependencies.

Other studies have focused on denoising and preprocessing. Zhang et al. [40] applied Bi-LSTM models to cleaned and interpolated AIS sequences, while Shin et al. [41] assessed various RNN architectures for port-bound trajectory prediction, noting performance differences based on spatial constraints.

Most of these studies evaluate trajectory forecasting models using pointwise error metrics, such as mean squared error or displacement error. However, only a few incorporate trajectory-level spatial similarity measures, such as the Fréchet Distance, which are more suitable for capturing the overall shape and accuracy of predicted vessel paths.

2.5 Summary and Research Gaps

The literature demonstrates the extensive utility of AIS data in maritime monitoring, prediction, and optimization tasks. Several research themes have matured considerably, including route clustering, ETA prediction, and short-term trajectory forecasting. However, important gaps remain:

- Integrated frameworks that unify data cleaning, prediction, and trajectory modeling in a coherent pipeline are still lacking in the literature.

- AIS data cleaning, especially in manually entered fields such as port destinations, remains under-explored.
- ETA prediction models often rely on static training data and aggregate error metrics, limiting their responsiveness to real-time operational dynamics.
- Trajectory forecasting research has focused predominantly on short-term predictions, with limited attention to how predictive accuracy degrades over longer horizons.
- Most trajectory forecasting evaluations use basic point-wise metrics, while fewer adopt trajectory-aware measures that better reflect operational impact.

These limitations indicate several avenues for future investigation in maritime data analytics, particularly in real-time, scalable, and context-aware systems grounded in high-quality AIS input.

3 An Intelligent Framework for Vessel Traffic Monitoring using AIS Data

One of the main research goals of this research work is the end-to-end development of an intelligent framework that will utilize AIS data in higher-level applications for improving vessel traffic monitoring.

As previously mentioned, the Automatic Identification System (AIS) is a real-time tracking system for ships, transmitting data via VHF signals about a vessel's position, speed, course, and ID. Widely used in the maritime sector, AIS supports navigation, collision avoidance, security, search and rescue, etc. Despite its utility, AIS data management faces challenges like high volume, duplicate signals, delays, and inaccuracies in manually entered fields (e.g., destination ports). Nearly half of AIS destination entries are error-prone, complicating real-time analysis and decision-making.

3.1 AIS Vessel Monitoring Framework

The end-to-end architecture of our AIS framework is shown in Figure 3.1. In order to have the data readily available to be used by higher-level applications, the data needs to be collected, pre-processed, organized, and stored in a flexible way.

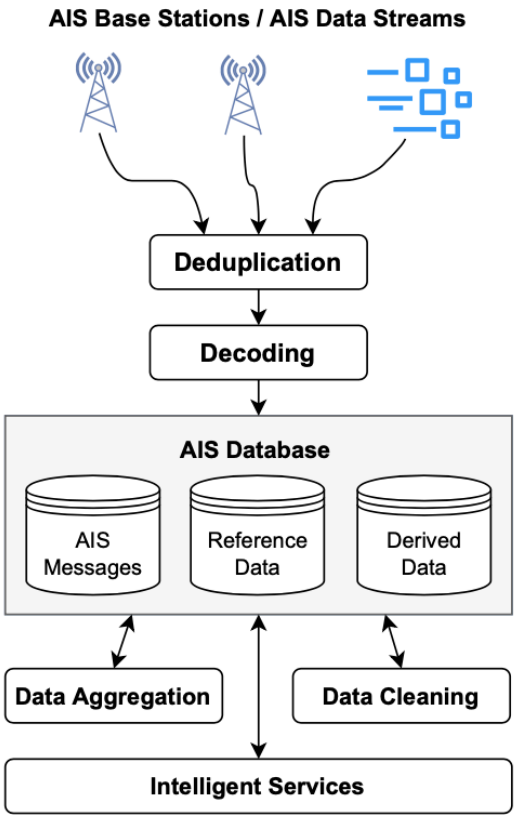


Figure 3.1: AIS framework architecture.

Data collection The process starts with the collection of AIS data originating from multiple sources, including live AIS base stations or other AIS data streams. Specifically, our framework currently collects AIS data from three distinct sources:

1. A custom AIS base station located at the Cyprus University of Technology in Limassol;
2. A stream of AIS data from five receivers deployed in coastal areas of Cyprus as part of the STEAM (2019) research project;
3. The official AIS stream from the Cyprus Shipping Deputy Ministry.

De-duplication. Since data is received from multiple sources, some of which may be located in close geographical proximity, the same AIS message may be received more than once. Only one version of the data must be kept as duplicate messages increase storage requirements, reduce performance, and may skew statistics. Hence, the first process run is de-duplication, which keeps only a single version of a message. De-duplication can take place without decoding the message by comparing encoded parts of the message, thereby minimizing decoding overheads.

Decoding. Further processing requires the message to be decoded to get the individual message fields, such as the unique vessel ID, position coordinates, speed, and course over ground, destination port, etc. The messages are decoded using libais 0.17 [42].

Storage. The processed data is stored in a relational database. According to the AIS standard, there are 27 different kinds of messages, each with its own set of fields and purpose. Thus, different messages are stored in different tables in the database. In addition, the database stores some reference data collected from other sources. For example, it stores a port data table, which lists over 17 thousand worldwide ports from the World Port Index, as well as a vessel registry containing over 23 thousand vessels.

Data cleaning. Some of the AIS fields are entered manually by the vessel crew and may be polluted with misspellings, truncations, unexpected abbreviations, or other irregularities. This data needs to be converted into a standardized form to be usable by high-level applications. To achieve this goal, we employ the use of fuzzy matching, a general technique for finding strings that match a pattern approximately rather than exactly.

Intelligent Services. The efficient processing of AIS data described above has enabled us to develop a set of intelligent services for providing online data analytics, real-time monitoring of areas of interest, and collision detection and avoidance, described next.

- **Online Data Analytics.** Our AIS framework provides a wide range of insightful statistics and intelligent data analytics for a particular period of time (e.g., last day, last week), which we also visualize through a web dashboard.
- **Real-time Monitoring of Areas of Interest.** Using AIS, it is possible to track vessels in real time that approach, enter, or exit various areas of interest such as port areas, marinas, fishing areas, or marine protected areas. The areas of interest are defined using coordinate points, which connect and form a location polygon. Then, we can check if a vessel is inside an area of interest by checking whether the coordinates of the vessel received in an AIS message are located within the corresponding enclosed area of the polygon.

- **Collision Detection and Avoidance.** Another beneficial use of AIS data involves making a prediction whether a collision between two vessels is likely to occur. Given the speed and course of two vessels (from AIS messages), we can calculate the positions of the two vessels in the near future (e.g., in 20 minutes), assuming the vessels maintain their course and speed.
- **Estimated time of arrival (ETA) Prediction** is the time when a vessel is expected to arrive at the particular port terminal [43]. Accurate ETA prediction helps terminal managers and stakeholders make quick and efficient collaborative decisions to enhance the terminal's performance. A study by Michaelides et al. analyzes the arrival time punctuality of container vessels at the Port of Limassol in Cyprus [44]. The study reveals that 45% of container vessels arrive 30 minutes late, 13% arrive 2 to 6 hours late, and another 13% arrive 1 to 3 days later than their ETA. When there is a significant discrepancy between the ETA and the actual arrival time, shipping companies face hefty penalties, and the entire berth allocation plan needs to be revised [7]. A method that accurately predicts vessel ETA arrivals, such as the one discussed in Chapter 3 of this dissertation, will be very beneficial to the community.
- **Route Prediction** is the ability to predict the path followed by a vessel to reach from one destination to another. The need for accurate prediction has never been more important. As the world economy has grown, sea traffic, threats to vessels, and our dependence on shipping have increased exponentially. Along with that dependence, the complexity of monitoring maritime assets in order to safeguard them has soared. Using AIS data can be a clever way of making predictions as to the position of a vessel in the future and act preemptively to address hazards such as potential collisions with other vessels, hazard areas (such as areas where planed works or military exercises are taking place) and other potential risks such as bad weather, etc. Chapter 4 of this dissertation discusses methods for route prediction.

3.2 AIS Data Intelligent Processing

This section presents the key processing tasks, namely deduplication, data cleaning, and online data aggregation, that perform pre- and post-processing of the data to shape it in a more organized and efficient manner.

3.2.1 Deduplication

AIS messages are collected by the framework from multiple receiving stations or data streams. Hence, it is frequent to receive duplicate messages within a very short period of time, typically within a few seconds. In addition, it is also possible to receive messages out of order, that is, receive a message that refers to an event that took place before an already received message. The deduplication module is able to handle these cases efficiently by utilizing a custom data structure that is used to buffer the messages received in a certain time window (e.g., the last 60 seconds). This data structure maintains a doubly linked list of time-ordered messages along with a hash table that maps a message's payload to the list's node that stores the message, as shown in Figure 3.2. The presence of the hash table makes checking for an existing duplicate message very efficient, while the linked list enables traversing the messages from both ends.

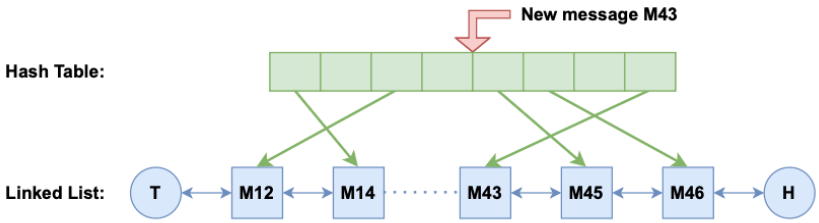


Figure 3.2: Custom data structure used during deduplication.

Algorithm 1 shows the overall deduplication process. When a new AIS message is received, the custom buffer is checked to determine whether the same message has arrived during the configured time window (line #2). This check happens only based on the encoded payload to avoid unnecessary decoding of duplicate messages. If the message is distinct from the existing messages, then the list is traversed from its head backwards until a node is found with a message containing an earlier (or equal) timestamp compared to the new message (lines #3-6). Note that this check is typically very fast because most messages arrive in the correct order. Next, the message is inserted at the correct position in the list and sent to the decoder (lines #7-8). Finally, the list is traversed from its tail forward in order to delete from the buffer all messages that are older than the current time minus the configured time window (lines #10-14).

Algorithm 1 Deduplication Process

```
1:  procedure ProcessNewMessage( message )
2:      If not buffer.contains(message .payload) then
3:          node = buffer.head
4:          while node.message.time > message.time do
5:              node = node.previous
6:          end while
7:          buffer.insertAfter(node , message )
8:      end if
9:      decoder.process(message)
10:     node = buffer .tail
11:     while node.message.time < currTime − interval do
12:         node = node.next
13:         buffer.delete(node )
14:     end while
15: end procedure
```

3.2.2 Data Cleaning

As mentioned earlier, some of the AIS fields are entered manually by the vessel crew and may be polluted with misspellings, truncations, unexpected abbreviations, or other irregularities. This data needs to be converted into a standardized form to be usable by high level applications [45]. To achieve this goal, we employ the use of fuzzy matching, a general technique for finding strings that match a pattern approximately rather than exactly. In simple terms, given an input string, fuzzy matching will find the most similar string from a list of strings based on a given similarity function. We have developed a fuzzy matching algorithm that receives dirty port destinations entries from AIS messages and finds the corresponding clean port names based on matching similarities with clean, standardized port records from a

reference ports table. Fuzzy Lookup builds and utilizes an Error-Tolerant Index (ETI) for finding matching rows in the reference table. Each record in the reference table is broken up into words, known as tokens, and the ETI keeps track of all the places in the reference table where a particular token occurs. Moreover, the algorithm uses a set of domain-specific rules as well as a custom distance function that takes into account the edit distance, the number of tokens (e.g., port codes, country codes, port names), token order, and relative frequencies. Hence, the matching process is resilient to a variety of errors that are present in the input records, while the cleaned records are associated with a score that indicates the quality of the match. More details can be found in Evmides et al. [45]. The output of the Fuzzy Matching algorithm is a mapping table that maps the dirty destination data to the clean data from the reference table. The mapping table can then be used by applications for converting dirty entries to clean ones very efficiently. In addition, the mapping table gradually grows over time as more data is cleaned. Hence, as time goes by, new dirty entries are more likely to already exist in the mapping table and do not need to go through the Fuzzy Matching algorithm.

3.2.3 Online Data Aggregation

Data aggregation concerns the procedure of calculating accumulating metrics like counts, max, min and average on AIS messages in order to provide insightful statistics. Since AIS messages are received at a rapid data income rate, we needed a way of optimizing the aggregation process in order to report various statistics with a decreased latency time. To achieve this, we have configured events that are scheduled to execute within our database at frequent time intervals. These events calculate and store various statistics that concern the previous day across various dimensions, such as the number of messages and unique ship visits per day, per port area, per ship status (e.g., moored, anchored), and more. Whenever data statistics are requested for reporting, we dynamically calculate the metrics of the current day and aggregate them with the previously pre-calculated data. This enables our framework to be fast and report live aggregate results back to the user with very low latency time (i.e., within milliseconds). The reason is because we avoid re- calculating past metrics that will not change in the future. More specifically, since AIS messages are received at real time, it is not expected to receive past messages and thus, previously calculated aggregations will not change. This is a key feature that significantly boosts performance.

3.3 Intelligent Services

The efficient processing of AIS data described above have enabled us to develop a set of intelligent services for providing online data analytics, real-time monitoring of areas of interest, and collision detection and avoidance, described next.

3.3.1 Online Data Analytics

Our AIS framework provides a wide range of insightful statistics and intelligent data analytics for a particular period of time (e.g., last day, last week), that we also visualize through a web dashboard. First, we numerically report the total messages received, discriminated by message type (e.g., Class A, Class B, Base Station, etc.). Then, we provide the number of unique ships seen grouped by country of registration using a pie chart. Furthermore, we have also utilized line graphs to visually report metrics like number

of signals received per day, number of ships seen per day, number of ships moored at specific ports per day, and number of ships at anchoring area per day. We have also integrated a few multiline graphs. The first one, again visualizes the number of ships moored at a particular port per day grouped by the different berth areas at the port. The other multiline graph reports the minimum, maximum, and average waiting time of the ships that are located at the anchoring area per day. All of these data analytics and visualizations provide the user and higher-level applications with important information and meaningful representations that can be further analyzed as needed.

3.3.2 Real-time Monitoring of Areas of Interest

Using AIS, it is possible to track vessels in real time that approach, enter, or exit various areas of interest such as port areas, marinas, fishing areas, or marine protected areas. The areas of interest are defined using coordinate points, which connect and form a location polygon. Then, we can check if a vessel is inside an area of interest by checking whether the coordinates of the vessel received in an AIS message are located within the corresponding enclosed area of the polygon. Given that messages are received every few seconds, the key challenge lies in ensuring that only distinct notifications are sent when a vessel is approaching, entering, or exiting an area the first time. Algorithm 2 shows (a simplified version of) the logic used to generate these notifications to interested applications. The input corresponds to one area and information about one vessel received in an AIS message, including its current geolocation, speed, and course. First, the past state of the vessel is retrieved from the database, representing the state reported by the last seen message and corresponds to whether the vessel was outside, approaching, or inside the area (line #1). If the vessel was outside the area and it is now inside the area, a notification is sent for the vessel entering this area (lines #3-5). If the vessel is not inside the area, its future location is projected to see if the vessel will enter the area in the near future; if so, a notification is sent for the vessel approaching this area (lines #6-8). Currently, the projected location is computed based on the current speed and course of the vessel but more advanced (e.g., machine learning-based) techniques can be used. Similarly, if the vessel was approaching the area and it is now inside the area, a notification is sent for the vessel entering the area (lines #9- 12). Otherwise, if the vessel was inside the area and now is outside, then a notification is sent for the vessel exiting the area (lines #13-16). Finally, the new state is saved in the database.

3.3.3 Collision Detection and Avoidance

Another beneficial use of AIS data involves making a prediction whether a collision between two vessels is likely to occur. Given the speed and course of two vessels (from AIS messages), we can calculate the positions of the two vessels in the near future (e.g., in 20 minutes), assuming the vessels maintain their course and speed. These calculations create two line segments, one for each vessel, from their current to their future position, as seen in Figure 3.3. If the two segments intersect, then the two vessels will pass from the same location in the near future. However, for a collision to occur, the two vessel paths must intersect at approximately the same time. Hence, we next compute the time it will take for each vessel to reach the point of intersection, given their current speed. If the two times are within a small delta from each other (e.g., 4 minutes), then a warning is issued that a collision is likely to occur. More advanced closest point of approach (CPA) algorithms are also investigated. Finally, the list of warnings is maintained in the database to avoid issuing repeated collision warnings, unless the warning parameters

Algorithm 2 Monitoring areas of interest in real time

```
1:  procedure NotifyAreasOfInterest( message )
2:      pastState = getPastState(area , vessel )
3:      if pastState == outside then
4:          if isInArea(area,vessel) then
5:              notifyEntering(area , vessel )
6:          else if isInArea(area, project(vessel)) then
7:              notifyApproaching(area , vessel )
8:          end if
9:      else if pastState == approacing then
10:         if isInArea(area, vessel) then
11:             notifyEntering(area, vessel)
12:         end if
13:     else if pastState == inside then
14:         if not isInArea(area, vessel) then
15:             notifyExiting(area, vessel)
16:         end if
17:     end if
18:     setState(area,vessel)
19: end procedure
```

change. The collision detection procedure must be checked frequently between pairs of vessels navigating at a close distance. To optimize this process, we collect the most recent AIS messages in frequent intervals and sort them based on latitude and longitude. Next, we check for collision only between vessels whose distance is smaller than a predefined threshold (e.g., 50 Km) instead of generating all possible pairs of vessels. We also have a slightly modified version of the collision detection procedure for the scenario where one vessel is moving and the other vessel is stationary. In particular, we create a line segment for the stationary vessel based on its length (plus some buffer length) and set it perpendicular to the course of the moving vessel.

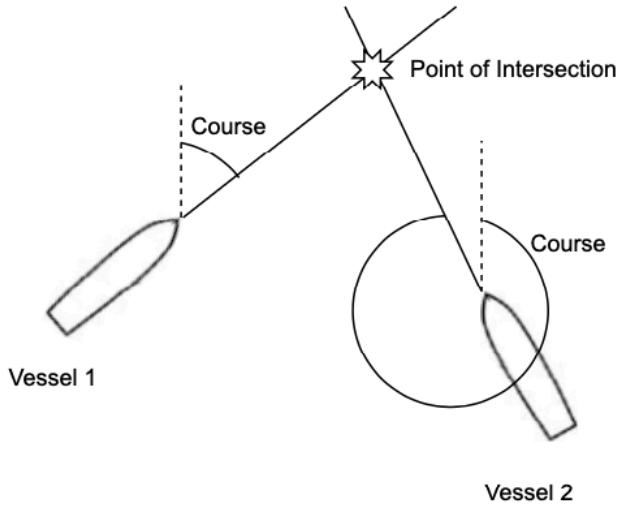


Figure 3.3: Vessel course intersection.

3.4 Conclusions

Our AIS framework is currently collecting AIS data from one in-house AIS base station and two AIS streams gathering data from 15 other base stations located along the coast of Cyprus and covering most of Eastern Mediterranean sea. Over the past few years, we have collected, processed, and stored close to 1 billion AIS messages, the majority of which (78%) belong to the category Class A Position Report, reporting detailed location information for commercial vessels. Table 3.1 shows the number of messages collected per AIS message category to date. Moreover, Figure 3.4 shows the AIS messages across all categories collected over the past two years, grouped by year and month. On average, 13 messages per second and over 1.1 million messages per day are received and stored in the database. Notice the seasonal pattern that reveals that more messages are collected over the summer months instead of the winter months as VHF signals are impacted by weather.

Table 3.1: AIS Message Types and Counts.

Category	Message Type	Count
Class A Position Report	1, 2, 3	1,458,342,558
Base Station Report	4	162,164,685
Class A Static and Voyage Data	5	74,019,300
Class B Position Report	18	102,028,357
Other Messages	6–17, 19–27	107,201,559

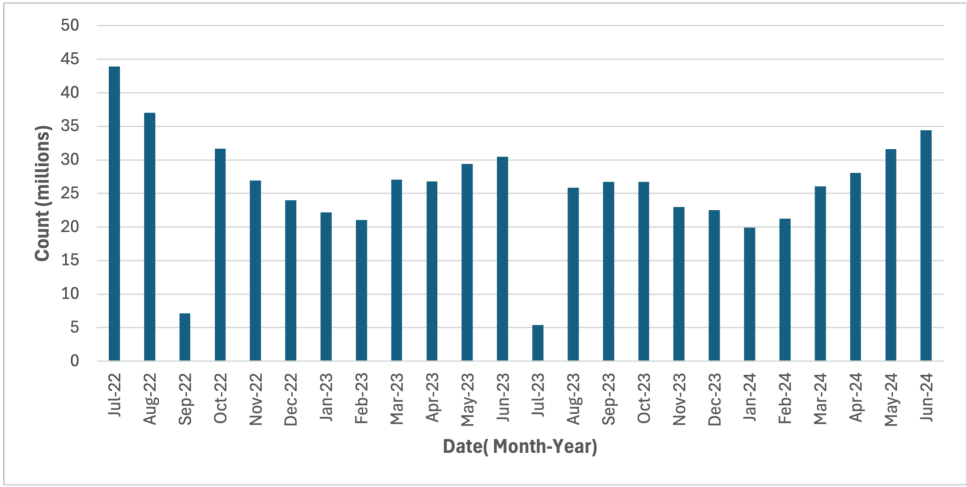


Figure 3.4: Collected AIS messages per month.

Another interesting observation from Figure 4 is that prior to October 2020, more messages were collected and stored in the database as that was the time before deploying our deduplication process. Regarding the data cleaning process, 91.6% of the Class A Static and Voyage Data messages containing destination ports were successfully cleaned, while 6.6% contained empty or invalid fields, and only 1.9% of the fields remained unmatched; highlighting the high accuracy of our fuzzy matching approach. Finally, Figure 3.5 shows a screenshot from a web interface visualizing collision warnings and notifications regarding vessels entering or exiting various areas of interest near Cyprus. In total, 291K notifications were sent for areas

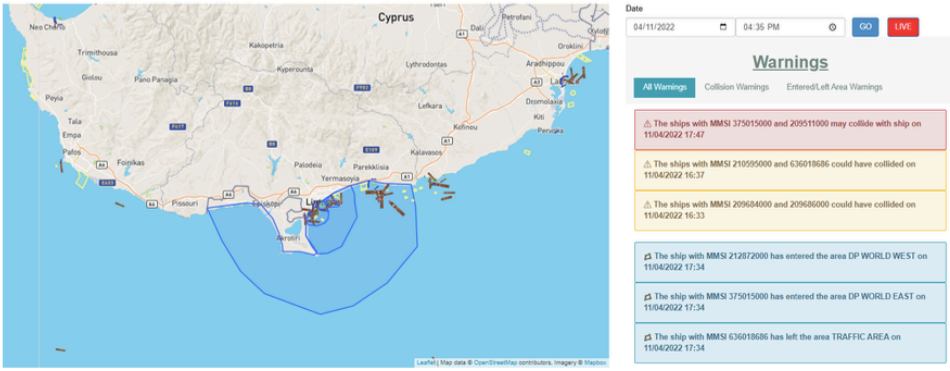


Figure 3.5: Collision warnings and notifications sent for vessels entering or exiting areas near Cyprus, as visualized in our framework.

within and around the Port of Limassol, Cyprus, showcasing the fine-grained vessel traffic monitoring that our framework can achieve.

AIS data have a wide spectrum of applications such as route prediction, ETA estimation, collision detection, risk assessment, etc. However, their use is surrounded with various challenges including their huge volume and velocity, duplicate messages, dirty fields, etc. This research study presents a new framework boasting intelligent processing approaches and services that cover the full cycle of AIS data collection and management in an efficient and effective way, that enables its easier use by current and future higher level applications. In the future, more data sources can be incorporated into the framework such as satellite AIS data or streams from other regions to complement the collected AIS information.

Compared to other AIS-based maritime analytics frameworks discussed in the literature, the framework introduced in this chapter represents a significant step toward integrating scalable real-time processing with robust data quality management and operational services. While prior systems such as Fu et al. [8] and Vodas et al. [9] have addressed large-scale AIS data handling and event detection, they tend to focus either on efficient data compression, scalability, or specific operational tasks in isolation. Similarly, platforms like datAcron [10] have pioneered streaming and contextual enrichment capabilities, and works such as Santipantakis et al. [11] and Bernabe et al. [12] have demonstrated advanced inference or anomaly detection methods, but these are often implemented as standalone analytics components rather than as part of an end-to-end operational infrastructure. In contrast, the proposed framework consolidates multiple essential capabilities—high-frequency data ingestion from multiple sources, intelligent deduplication, semantic cleaning (e.g., destination field correction), and contextual enrichment—within a single distributed architecture that is already deployed and operational at scale in the Eastern Mediterranean. This allows it to not only match the streaming and enrichment capabilities of existing platforms but also to provide higher semantic consistency for downstream tasks. Overall, the proposed framework bridges the gap between raw AIS data acquisition and advanced operational analytics more comprehensively than most existing solutions. By embedding both robust preprocessing and predictive service capabilities into a unified, scalable infrastructure, it not only aligns with but extends beyond prior state-of-the-art systems, providing a replicable blueprint for regional and potentially global maritime surveillance architectures.

The framework is deeply embedded in the research presented throughout the subsequent chapters, serving as the backbone for the development and evaluation of higher-level applications. In Chapter 4, the desti-

nation field cleaning methods represent a key component of the broader data preprocessing pipeline, significantly enhancing the semantic quality and consistency of AIS data. This refined data is then utilized in Chapter 5 to train machine learning models for Estimated Time of Arrival (ETA) prediction, where accurate and reliable input is essential for performance. In Chapter 6, the same high-quality AIS data underpins the development of deep learning models for vessel trajectory forecasting, further demonstrating the framework's ability to support complex, real-time maritime analytics. Together, these chapters showcase the framework's central role in enabling scalable, data-driven solutions for maritime situational awareness and decision-making.

While prior research (e.g., Fu et al. [8], Vodas et al. [9], Santipantakis et al. [11]) has explored ETA estimation, these efforts often lack systematic handling of noisy AIS inputs and bias in model training. Our route-based data splitting methodology mitigates overfitting and supports robust generalisation across voyages. The resulting ETA predictions not only enhance operational decision-making but also feed directly into long-term trajectory forecasting models discussed in Chapter 6, demonstrating the value of integrated, high-quality AIS preprocessing for downstream maritime analytics.

4 Employing Fuzzy Matching for Cleaning Manual AIS Entries

As established in the introduction, AIS data has evolved into a vital resource for maritime analytics, underpinning a wide array of applications that extend far beyond its original safety-centric purpose. The quality and consistency of this data—particularly the accuracy of destination information—are critical for downstream applications such as Estimated Time of Arrival (ETA) prediction and vessel trajectory forecasting, both of which depend on reliable input to generate trustworthy outputs.

While most AIS information is generated automatically via onboard instruments, the destination field is manually entered by the ship's personnel as part of AIS Message Type #5 and may contain up to 20 ASCII characters. Even if some systems were to provide a menu list of port destinations to choose from, there is no agreed standard for how the destination field should be reported. As a result, the destination entries are often polluted with misspellings, truncations, inserted punctuation marks, unexpected abbreviations, and other irregularities.

For example:

- `Limassol CY`
- `--Limmasol--`
- `LMSCY`
- `GRPIR=>CYLIM`

All of these refer to the same location: Port of Limassol, Cyprus.

Past studies estimate that 49% of reported AIS destination entries are polluted with similar errors. Such dirty data not only hinders maritime authorities' ability to comprehend and act on information but also makes this information unusable by higher-level digital applications, which lack human-like cognitive capabilities to interpret noise.

This dissertation proposes an automated method to clean and normalize dirty port destination data with minimal human intervention. Using fuzzy matching against a reference table of over 17,000 standardized port names and their UN/LOCODE identifiers, the method applies a domain-specific similarity function and produces a mapping table between dirty and clean entries. This mapping grows over time, reducing computation and increasing reliability for future inputs.

Experimental validation was conducted using real-world AIS data from the Eastern Mediterranean collected over three years. The system processed 2.7 million dirty destination entries and achieved a 91.6% success rate in mapping them to standardized port names.

4.1 AIS Data Collection and Cleaning Overview

Figure 3.1 presented in Chapter 3 shows the overall system architecture of our AIS framework for collecting and analysing AIS data. The framework consists of the following major steps: (1) AIS message

collection, (2) deduplication, (3) decoding, (4) database storage, (5) port destination cleaning, and (6) utilization of AIS data in higher-level applications.

The process begins by collecting AIS messages transmitted by ships equipped with AIS transponders. These messages, which can be received by any AIS receiver within range, include encoded data such as vessel identity, geographic position, speed over ground, course over ground, and intended port of destination.

Due to overlapping coverage, the same AIS messages may be received from multiple sources. To mitigate storage redundancy, a deduplication module ensures only a single copy of each message is retained. Additionally, since AIS messages are transmitted in encoded formats, a decoding module processes and extracts the individual fields upon receipt of each deduplicated message.

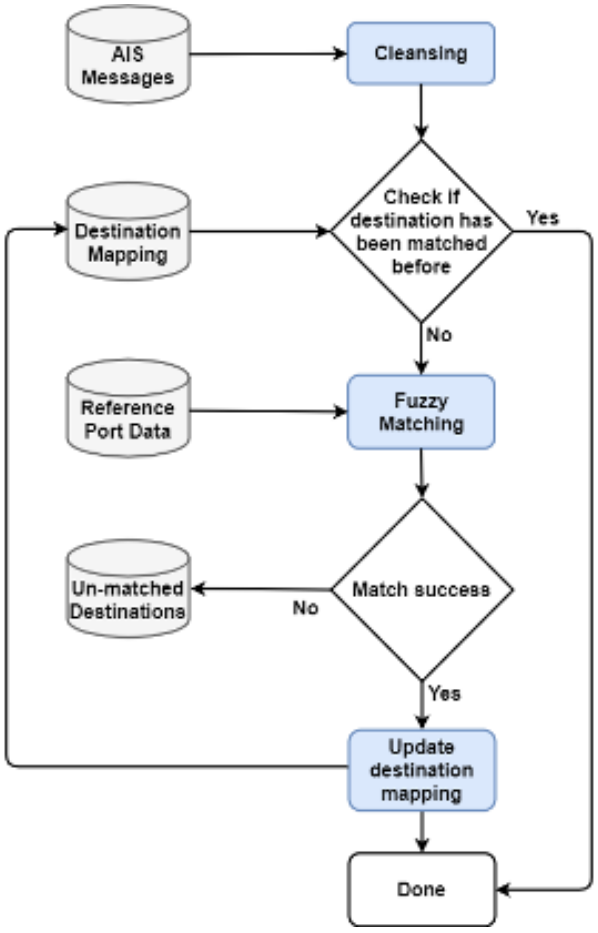


Figure 4.1: Destination cleaning process diagram

The decoded messages are stored in a relational database. AIS defines 27 different message types, each with a unique structure and purpose. These are stored in separate tables. The port destination field, of particular interest in this study, is part of the 'Static and Voyage Data' message (type #5), broadcast every 6 minutes by Class A vessels.

In addition to AIS messages, the database includes two auxiliary tables:

- **Reference Port Data Table:** Contains over 17,000 global ports from the World Port Index (2019), including port names and UN/LOCODEs.

- **Destination Mapping Table:** Maps raw destination field entries to cleaned and standardised port names.

The Destination Mapping Table is updated with new entries upon the completion of each Destination Cleaning process. Applications can then efficiently use this mapping to convert dirty AIS destination strings to their clean equivalents.

The Destination Cleaning process is periodically executed in batch mode. As illustrated in Figure 4.1, each port destination string is first preprocessed by removing punctuation, special characters, and numbers, and by converting all letters to uppercase. The cleaned entry is checked against the Destination Mapping Table. If a match is found, it is reused; otherwise, the entry is passed to the fuzzy matching algorithm.

The algorithm applies nine domain-specific matching rules and calculates similarity using a custom distance metric based on edit distance, token count, token order, and frequency. If a match is found, it is logged in the Destination Mapping Table with a quality score. This improves the efficiency of future lookups. Entries that cannot be matched are stored separately for future human review or improved algorithms.

4.2 Fuzzy Matching Algorithm

Fuzzy matching is a general technique for finding strings that match a pattern approximately rather than exactly. In layman's terms, given an input string, fuzzy matching will find the most similar string from a list of strings based on a given similarity function.

We implemented our fuzzy matching algorithm using Fuzzy Lookup transformations that are a part of the SQL Server Integration Services (SSIS). Fuzzy Lookup matches 'dirty' input records—polluted with misspellings, truncations, unexpected abbreviations, or other irregularities—with clean records from a reference table [46].

Fuzzy Lookup builds and utilises an Error-Tolerant Index (ETI) for finding matching rows in the reference table. Each record in the reference table is broken up into words, known as tokens, and the ETI keeps track of all the places in the reference table where a particular token occurs. For example, if the reference table contains the entry `Limassol Cyprus CY, CYLMS`, the ETI will contain entries for `Limassol`, `Cyprus`, `CY`, and `CYLMS`. Additionally, Fuzzy Lookup indexes substrings (q-grams), enabling it to match records even with extra or missing characters.

Since the ETI can grow substantially with a large number of unique tokens and reference entries, we construct the index once and reuse it during fuzzy matching.

Given an input string, the algorithm uses the ETI to extract potential matches from the reference table that share common tokens or q-grams. These candidate entries are then compared in a second round of matching, and the best match is selected and returned, with a similarity score between 0.0 and 1.0 (where 1.0 denotes an exact match).

Our empirical analysis led us to set a minimum similarity threshold of 0.7, as matches with lower scores tended to be incorrect. Therefore, any match with a similarity score below 0.7 is considered unmatched. Automatically adjusting this threshold remains an open topic for future work.

Table 4.1: Dirty destination examples following nine different reporting patterns and their cleaned counterparts

Dirty Destination	Port Code	Port Name	Country Code	Country Name
BEIRUTE>< ,	BEY	Beirut	LB	Lebanon
SHKELON IL	AKL	Ashkelon	IL	Israel
ABU QIR / EGYPT	AKI	Abu Kir	EG	Egypt
CHKLB	CHK	Chekka	LB	Lebanon
OPL CYPRUS	–	–	CY	Cyprus
EGYPT, DAMEITTA	DAM	Dumyat (Damietta)	EG	Egypt
>EG PSD	PSD	Port Said	EG	Egypt
CY LARNACA	LCA	Larnaca	CY	Cyprus
GRVOL=>ILHFA	HFA	Haifa	IL	Israel

Apart from misspellings, truncations, or punctuation, destination entries may appear in various formats. Since there is no agreed reporting convention, fields might contain port names, UN/LOCODEs, country codes, or combinations thereof. To understand the common formats used, we collected 140,000 random destination entries and conducted exploratory data analysis. This led to the identification of nine prevalent patterns, examples of which are shown in Table 4.1.

Based on these patterns, we formulated nine domain-specific matching rules that reflect how ship captains or crew members report destinations. Each rule corresponds to a specific combination of fields. Table 4.2 outlines the rules used in our system.

Table 4.2: Rules for matching the port destination field

ID	Rule Name	Description	Clean Example
1	PortName	Port name (usually a city or locality)	Limassol
2	PortName CountryCode	Port name + 2-letter ISO country code	Limassol CY
3	PortName CountryName	Port name + full country name	Limassol Cyprus
4	PortCode CountryCode	UN/LOCODE + 2-letter country code	LMS CY
5	CountryName	Country name only	Cyprus
6	CountryName PortName	Country name + port name	Cyprus Limassol
7	CountryCode PortCode	Country code + UN/LOCODE	CY LMS
8	CountryCode PortName	Country code + port name	CY Limassol
9	CountryCode PortCode (origin) CountryCode PortCode (destination)	Origin and destination using country codes and UN/LOCODEs	GR PIR CY LMS

For each destination field, Fuzzy Lookup is invoked nine times in parallel—once per rule—with each invocation returning a similarity score. The highest score among the nine is selected as the best interpretation, assuming it exceeds the 0.7 threshold. If no rule produces a high enough score, the entry is marked as unmatched. Additionally, destination fields that are empty or contain no alphabetic characters are also excluded from the matching process.

4.3 Experimental Evaluation

A total of 2,778,520 unique AIS ‘Static and Voyage Data’ messages were collected and stored (after deduplication) from the received AIS data that cover the Eastern Mediterranean region near Cyprus. The collection period starts on April 11, 2017 and ends on May 01, 2020. In total, the Fuzzy Matching algorithm was able to match the destination field for 2,543,802 messages, while 52,375 fields remained

unmatched. The remaining 182,343 fields were either empty or invalid (i.e., lacked any alphabetic characters). Table 4.3 highlights our key results, showing that our algorithm is able to achieve a 91.6% matching rate, while it fails to match the destination field for only 1.9% of the cases. The remaining 6.6% corresponds to empty or invalid destination fields.

Table 4.3: Aggregated matching results

	Count	Percentage
Total dirty destination records	2,778,520	
Total matched records	2,543,802	91.6%
– Records matched from Destination Mapping table	2,349,847	
– Records matched using Fuzzy Matching algorithm	193,955	
Unmatched Records	52,375	1.9%
Records with empty or invalid destination	182,343	6.6%

As discussed in Section 3, the destination cleaning process was run in periodic batches. In total, we run 29 batches of approximately 100,000 AIS messages each, corresponding to the average number of AIS messages we receive each month. Figure 3 shows how many records were matched using the Fuzzy Matching algorithm and how many records were matched from the Destination Mapping table (recall Section 3) for each batch. During the first batch, the Destination Mapping table was empty, and thus all records were directed to the Fuzzy Matching step. Afterwards, we observe a downward trend in the data being matched using the fuzzy matching algorithm. In particular, only 5–15% of the data go through fuzzing matching. Collectively, out of the 2,543,802 matched records, only 193,955 records were fuzzy matched. This demonstrates our algorithm's adaptive behaviour in 'learning' from previous matches as time goes by and utilising that knowledge to interpret future data.

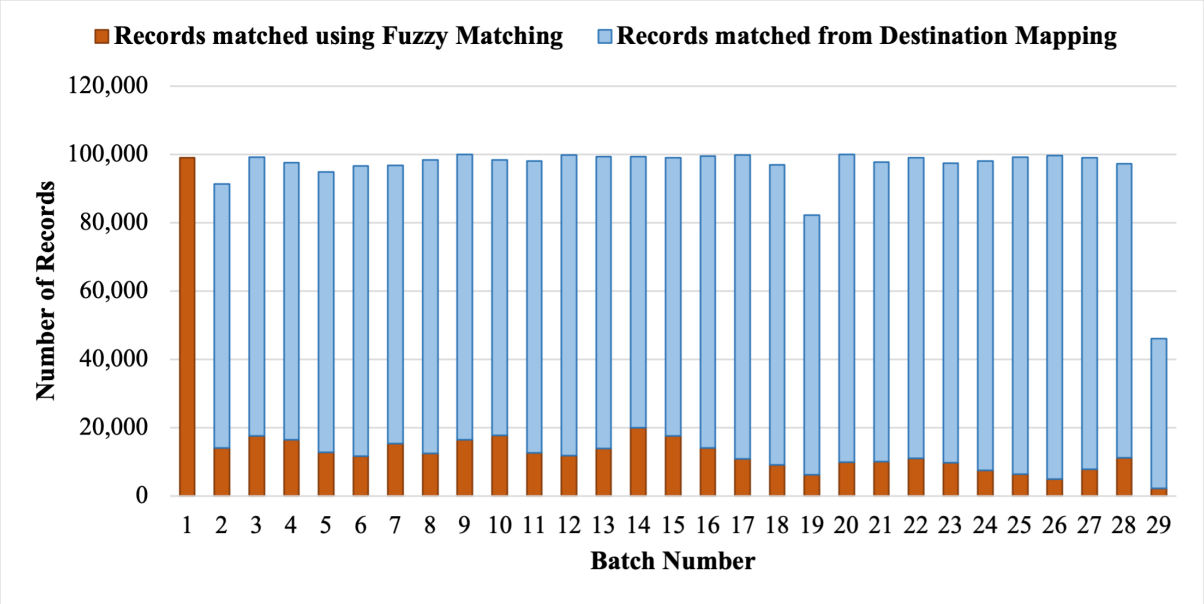


Figure 4.2: Number of records matched using Fuzzy Matching or Destination Mapping for each batch

Next, we analyse the contribution to the matching results of each of the nine matching rules we defined in Section 4. The results are shown in Table 4.4. The most popular form for the destination field is simply

Table 4.4: Number and percent of destination entries matched for each matching rule

Rule Name	Matched Destination Entries		Matched to Unique Ports	
	Count	Percentage	Count	Percentage
PortName	1,821,728	71.61%	1,103	27.54%
PortName CountryCode	13,664	0.54%	166	4.14%
PortName CountryName	1,474	0.06%	89	2.22%
PortCode CountryCode	6,887	0.27%	84	2.10%
CountryName	64,024	2.52%	89	2.22%
CountryName PortName	4,326	0.17%	24	0.60%
CountryCode PortCode	490,796	19.29%	1,968	49.14%
CountryCode PortName	21,074	0.83%	92	2.30%
CountryCode PortCode (origin) & CountryCode PortCode (destination)	119,829	4.71%	390	9.74%
Total	2,543,802	100%	4,005	100%

the port name, as 71.6% of the messages are matched using the 'PortName' rule, while the second most popular (at 19.3%) is specifying the country code followed by the port's UN/LOCODE. The remaining patterns have a varying contribution between 0.1% and 4.7%. All these destination fields are matched to unique ports from the Reference Port Data table. Interestingly, the destinations declaring only the port name lead to 1102 unique ports (27.5%), while the destinations with the 'CountryCode PortCode' form lead to 1968 unique ports (49.1%). Overall, the 2.5 million matched dirty destinations are mapped to 4005 unique ports. Such disparity is expected since each vessel broadcasts an AIS 'Static and Voyage Data' message every 6 minutes with the same destination field for the duration of its trip from one port to another.

In an attempt to understand the limitations of the fuzzy matching algorithm, we manually investigated the 52,375 destination fields that remained unmatched. We found 107 distinct entries that a human expert could match to a port. Those distinct entries account for 21,242 destination fields, and thus our algorithm failed to find a match for only 0.83% of the total entries that could be matched. The failed matches are attributed to the following scenarios:

- **Use of abbreviations:** In several cases, the ship's crew may abbreviate the name of the destination port, making it hard for the fuzzy matching algorithm to work. For example, the entry 'ALEX-EYG' refers to the El Iskandariya (Alexandria) Port in Egypt. If the entry contained a longer substring of 'Alexandria', the 'PortName CountryCode' matching rule would have been able to match it correctly (even with the misspelt country code).
- **Use of incorrect UN/LOCODE:** For some ports, a wrong UN/LOCODE is used almost as frequently as the correct entry. For instance, SZC is often used to refer to the Suez Canal even though the official code is SCN. The same is true for the Port of Limassol (LIM instead of LMS) and the Port of Larnaca (LAR instead of LCA). By having two incorrect letters out of three leads to a very low similarity score in the case of matching port codes, even when the country code is present.
- **Use of unofficial names:** Instead of reporting the official name of the destination port, sometimes the crew uses an unofficial name or synonym. For example, the Suez Canal is sometimes referred to as 'SUEZPASSAGE' or 'SUEZSTRAIT'.

- **Use of nearby location names:** In some cases, the destination field will contain the name of a nearby location or the name of the general area instead of the name of the actual port. According to domain experts, for example, it is common for a ship that is headed to the Port of Canakkale to report that it is headed to the Dardanelles Strait (also known as the Canakkale Strait).

Except for using abbreviations, the other scenarios could be addressed by extending the Reference Port Data table with other commonly used entries. For example, we could extend the clean record of Suez Canal to include the code SZC and the unofficial names 'Suez Passage' and 'Suez Strait' so that the fuzzy matching algorithm can match those entries as well.

It is important to note that the 107 manually matched entries have been added to the Destination Mapping table. Hence, any future destination entries that suffer from those specific issues will now be correctly matched and not go through the fuzzy matching algorithm. This aspect reveals how our framework can easily incorporate human expertise in order to improve further its operations and accuracy in matching and resolving dirty destination entries.

The remaining 31,133 unmatched destination entries contain values that do not correspond to any valid port, country, or location, and were correctly marked as unmatched by the fuzzy matching algorithm. In many cases, those values correspond to the status of the ship (e.g., 'cruising', 'awaiting orders', 'U/W'), or the function of the ship (e.g., 'cable work', 'drilling'), or simply nonsense (e.g., 'BMXCPRE@8P?? B0', 'Z3-P,F').

Finally, we investigated the correctly matched entries to ensure that the fuzzy matching algorithm is generating meaningful results. We have found two potential limitations:

- **Ports with the same name:** There exist ports in the world that have the same name such as 'Alexandria' (located in both Egypt and Washington DC, USA) or 'Naples' (located in both Italy and Florida, USA). Given that the AIS messages were collected in the Eastern Mediterranean sea, it is far more likely that the previous entries refer to Alexandria, Egypt and Naples, Italy, respectively, rather than locations in the USA.
- **Ambiguous references:** Some destination fields contain names or locations that are either not specific enough or sound similar to other ports. For example, the destination field 'Athens', most likely refers to the Port of Piraeus (which is located in Athens, Greece) and not Athena in Florida, USA. Another example is the destination 'Delta Nile', which refers to some port in the Delta Nile, Egypt, but without specifying which one.

No string-based approach, including fuzzy matching, can correctly identify the destination port in such cases. One option would be to add multiple entries in the Destination Mapping table so that higher-level applications can know that a dirty destination may refer to multiple locations (such as the example with the Port of Alexandria). Another option would be to extend our approach to consider more information when resolving the destination field, such as the current path trajectory or historical destinations of the vessel. We leave such an investigation for future work.

4.4 Conclusions

In this dissertation, a framework is proposed for collecting, storing, and automatically cleaning AIS data with respect to the destination port entry, thereby putting this data in a standardised format so that it can be further utilised by other digital applications. Towards this end, a novel fuzzy matching algorithm was developed and implemented using nine domain-specific matching rules. Overall, the Fuzzy Matching algorithm makes the AIS data collection process more resilient to several common errors observed in real data, while automatically determining the correct port destination of vessels can improve the efficiency of the maritime transportation industry as well as traffic safety. Besides, the complete methodology is adaptive in the sense that it can reuse previous results in order to reduce its computational resources over time. The performance of the algorithm was evaluated against real data, collected from the Eastern Mediterranean sea, and achieved an overall matching success rate of 91.6% in interpreting the dirty data.

The fuzzy matching–based AIS cleaning approach in this chapter focuses specifically on normalising the destination field, which is often noisy due to free-text input by vessel operators. Unlike conventional methods that rely on exact string matching or manually curated dictionaries, the fuzzy matching approach can recognise partial matches and correct typographical errors, abbreviations, and inconsistent formatting. While other works on AIS preprocessing (e.g., [15]) have applied rule-based standardisation or port-code mapping, these approaches are less flexible in handling ambiguous or non-standard entries. Our method strikes a balance between automation and accuracy, enabling more consistent identification of vessel destinations and improving the quality of features used in downstream ETA prediction and trajectory modelling.

In the future, we plan to perform our destination cleaning process in real-time, that is, clean the dirty destination fields as soon as AIS messages are decoded and before they are stored (instead of running the cleaning process periodically in batches). This will enable the use of our framework by applications that have strict real-time requirements. Another issue that needs further investigation is simplifying (or even eliminating) the reviewing process of the unmatched results. For example, the unmatched results can be presented to a human with possible suggestions for a match, based on the similarity score outcome of the fuzzy matching. We also plan to explore automated approaches for verifying the matched entries and flagging any potential inconsistencies or ambiguous results. For example, a ship sailing eastbound in the Eastern Mediterranean is unlikely to have as the next destination any port in the USA. For this purpose, information from other types of AIS messages will need to be utilised, pertaining mainly to speed and direction of vessel movement.

The framework proposed in this chapter serves as a foundational building block for the research presented in subsequent chapters. When combined with the AIS data cleaning approach introduced in Chapter 3, it enables the construction of a comprehensive and reliable dataset, which is subsequently leveraged in the advanced modeling and predictive tasks explored in Chapters 5 and 6.

5 Enhancing Prediction Accuracy of Vessel Arrival Times Using Machine Learning

The containerized volume handled by ports is increasing every day, leading to several seaside problems, i.e., port congestion, long waiting times, and accidents. According to a report presented by the United Nations Conference on Trade and Development (UNCTAD) [2], an increase in the waiting time of container ships at the port was noted following the COVID-19 pandemic, especially for developed countries, as depicted in Figure 5.1. To address this issue, a logistics plan can be developed to enhance the efficiency of sea transportation. The logistics plan may include berth scheduling, quay crane scheduling, vessel path planning, and accurate prediction of vessel arrival times. However, a significant obstacle lies in the fact that vessels must arrive punctually for this plan to succeed. Unfortunately, only 55-89% of vessels follow their schedules and arrive on time at the destination port [25, 47].

Even in some cases, timely arrived ships need to wait before mooring due to inefficient berth allocation plans [48]. Uncertainty in vessel arrival times also lowers the schedule's reliability, causing delays and congestion problems and decreasing productivity levels for inland transport operators. Late arrivals of ships at the port cause high costs of vessel operation and the whole supply chain [49]. For example, when ships arrive late, port management authorities must adjust the entire berth allocation schedule. This can result in delays for other ships as new plans are implemented or routes are rerouted. Additionally, frequent changes to berth plans can significantly decrease the performance of the port terminal, as berth planning is fundamental to its operations [50–52]. Given the numerous decisions tied to vessel arrival times, precise prediction of vessel arrivals at the terminal is essential for the optimal functioning of any container terminal.

The estimated time of arrival (ETA) is the time when a vessel is expected to arrive at the particular port terminal [43]. Accurate ETA prediction helps terminal managers and stakeholders make quick and efficient collaborative decisions to enhance the terminal's performance. A study by Michaelides et al. analyzes the arrival time punctuality of container vessels at the Port of Limassol in Cyprus [53]. The study reveals that 45% of container vessels arrive 30 minutes late, 13% arrive 2 to 6 hours late, and another 13% arrive 1 to 3 days later than their ETA. When there is a significant discrepancy between the ETA and the actual arrival time, shipping companies face hefty penalties, and the entire berth allocation plan needs to be revised [7]. At the Port of Limassol, passenger ships demonstrate high arrival punctuality, with 61% arriving within 30 minutes of their ETA and an additional 31% arriving within 60 minutes [53]. This high level of punctuality is expected since passenger ships adhere to very strict schedules. However, for other vessel types, the 30-minute arrival punctuality ranges only between 41% and 49% [53]. Some other studies from existing literature also deal with ETA prediction and punctuality analysis, such as a study presented in [54] that predicts ETA for container ships in short sea shipping while exploiting meteorological and automatic identification system (AIS) data. Another study [55] performs ETA prediction using AIS data and proposes a deep learning-based approach. A neural network-based solution is proposed in [56] for ETA predictions using historical ship position data of two dedicated areas in the Netherlands and Germany.

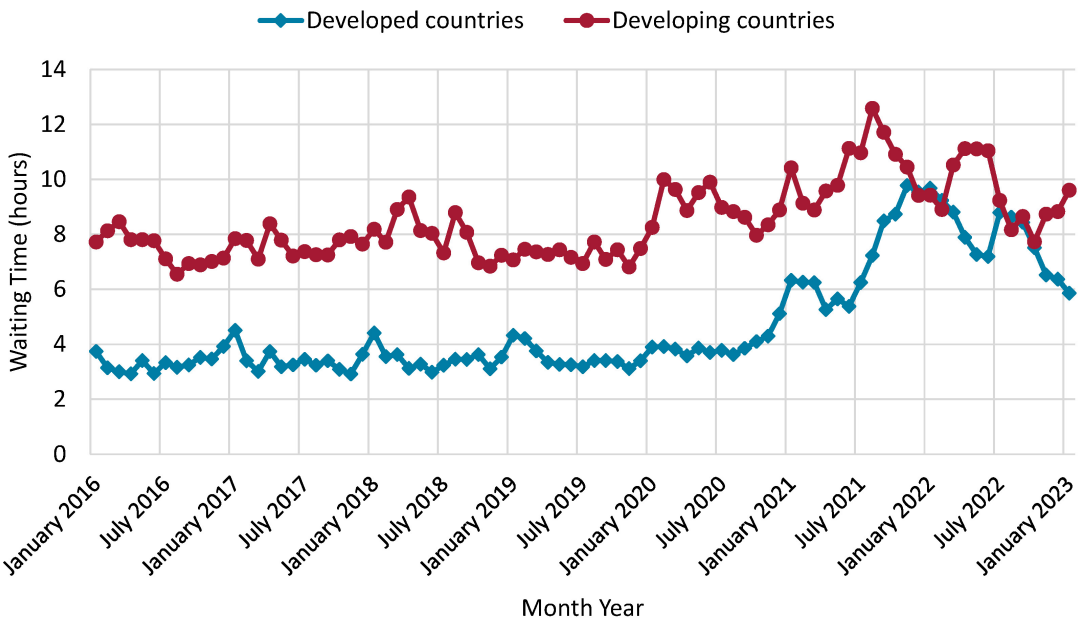


Figure 5.1: Average waiting times in hours of container vessels. Data provided by [2]

Most of the recent studies on ETA predictions of vessels use machine learning-based methods trained on historical data related to ships' paths, positions, and other parameters. Although these approaches yield satisfactory results, they are often inadequate for continuous real-time predictions. Therefore, this study aims to propose an accurate data-driven methodology that can predict ship arrival times in real time using machine learning. Furthermore, our proposed methodology evaluates and selects the most suitable algorithm to perform ETA based on the current situation and creates the best feature/input combination from the available AIS data. This study develops six models, namely Deep Neural Networks (DNN), K-Nearest Neighbours (KNN), Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Gaussian Naive Bayes (GNB). This research study makes the following key contributions.

- The study generates and evaluates a comprehensive list of feature combinations based on AIS data that can enhance the prediction accuracy of vessel ETA;
- The study proposes a new methodology of splitting data based on vessel routes to avoid bias and overfitting during the training and testing of machine learning models;
- The study evaluates six machine learning algorithms for predicting vessel ETA and compares them against multiple recently proposed, machine learning-based approaches;

The rest of the chapter is structured as follows. The methodology employed in this study is outlined in Section 5.1. Section 5.2 details the simulation settings, including results and discussions. Section 5.3 discusses the benefits that can be achieved with the accurate prediction of vessel ETAs, along with future work. Finally, Section 5.4 concludes the study.

5.1 Methodology

Compared to prior work, this study systematically evaluates and selects a more extensive collection of features originating from the two position reports (A and B) and the static voyage data of AIS. Our final

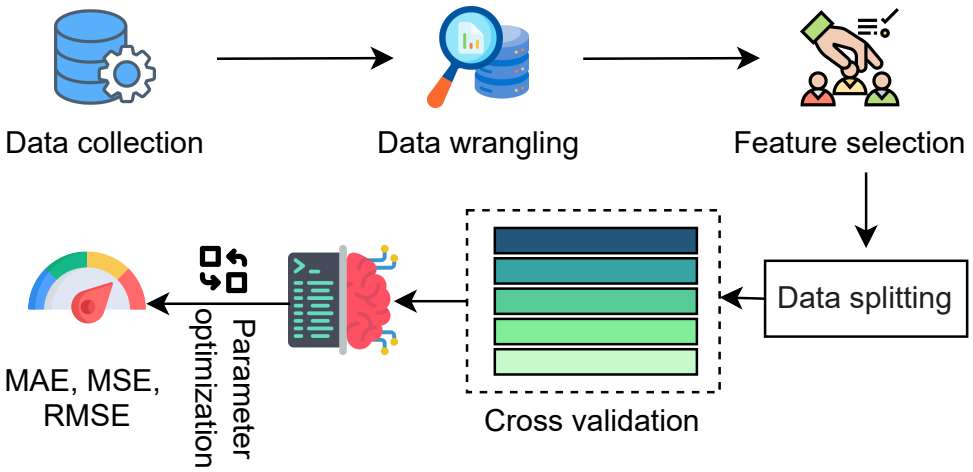


Figure 5.2: Overall methodology for vessel arrival times prediction

selection of only AIS-based features makes the training and inference much more efficient, enabling the real-time prediction (and correction) of vessel ETAs. In addition, this study performs a comprehensive comparison of six different machine learning algorithms against six other approaches, namely the estimated time of arrival provided by shipping agents, a simple calculation-based approach, and four other machine learning models proposed recently in the literature.

Figure 5.2 shows the overall methodology employed by this study for performing vessel arrival times prediction using machine learning models. The process involves collecting AIS data and transforming it (i.e., wrangling) into a format that is readily consumable by the downstream methodology steps (Section 5.1.1). Next, a long list of features were evaluated and selected for model construction (Section 5.1.2). To measure the fitness of predictions across multiple models, the popular approach of cross-validation was utilized, while the data was split based on vessel routes to avoid overfitting (Section 5.1.3). Finally, hyperparameter tuning is performed to identify the best hyperparameters to use for each machine learning model that optimizes its performance (Section 5.1.4). All steps are further elaborated below.

5.1.1 Data Collection and Wrangling

This study uses AIS data collected pertaining to the Eastern Mediterranean Sea from 17 AIS base stations operated by the Cyprus Shipping Deputy Ministry, Tototheo Ltd, and Cyprus University of Technology (CUT) between 2020 and 2022 and stored in the CUT-AIS platform [57]. AIS data contain several pieces of information about the vessels moving in the area, such as the position (longitude, latitude), size (length, breadth, depth), type, speed over ground (SoG), course over ground (CoG), rate of turn (RoT), and heading of the vessel as well as the next destination port and ETA provided by the agent [45]. The data related to the destination and actual time of vessel arrivals (ATA) was retrieved from the Port Community System of the Cyprus Ports Authority.

The raw AIS data needs to be transformed into a usable format containing only numeric values, which can then be used for training and testing the machine learning models. After decoding the AIS messages, most AIS data values are already in a numerical format (e.g., longitude, latitude, SoG, CoG, etc.) and can be used as is. String values such as the destination and vessel type are converted into numbers using ordinal encoding. In addition, the distance to the destination is computed given the current location and

the coordinates of the destination. Finally, the ETA provided by the shipping agent, as well as the ATA, are converted to minutes to the destination by computing the time difference between the time of the AIS message and the ETA and ATA, respectively.

After analyzing the dataset, it is observed that the ETA provided by the shipping agents is often incorrect, with a mean absolute error of 178 minutes (almost 3 hours). Indicatively, the shipping agent, on a certain occurrence, indicated that the ship would arrive in 1745 minutes (1 day and 5 hours) while the ship actually arrived in 2445 minutes (1 day and 17 hours). The worst error observed in our dataset was 4319 minutes (3 days), showing just how inaccurate the provided ETA can be. Furthermore, it was observed that the shipping agent, in many cases, does not update the ETA as the vessel progresses through the route. In other cases, the ETA indicates that the vessel has arrived while it is actually still on its way to its destination. The reverse is also observed sometimes, where the ETA provided by the shipping agent states that the vessel will arrive much later than the actual arrival time of the vessel. These observations further motivate the need for more accurate predictions of the vessel arrival times.

Finally, errors have been observed in relation to the validity of the vessel location, where some longitude and latitude values are much further away from the vessel route, as depicted by the majority of the position data. This error occurred in about 4.25% of the data and is considered rare. It was also observed that the longitude and latitude are updated every 2-3 seconds, and wrong values are corrected within 1-2 minutes. This error can be easily detected while calculating the remaining distance of the vessel route using outlier detection. Hence, data points containing invalid position values are eliminated from the dataset to avoid confusing the model and reducing its performance.

5.1.2 Feature Selection

The process of selecting features for training predictive models is of utmost importance [58]. However, there is a scarcity of research that thoroughly analyzes the most suitable AIS input data for accurate predictions. This study has drawn insights from three relevant papers to guide our feature selection process while also considering several additional features extracted from AIS data. The list of all features considered, as well as the ones employed in past papers and this research study, are shown in Table 5.1.

Using numerous features can lead to overfitting, where the model performs well on the training data but poorly on new data. Therefore, this study prioritizes feature combinations that have been identified as most important in past research. When considering new features, this study only includes those deemed highly relevant for predicting ETA. A recursive feature elimination with cross-validation process was employed for computing the importance score for each feature. Based on our extensive analysis, the following features used in prior research are excluded from this research study: change in speed compared to the last three hours, average speed of last 12 hours, timestamp when the AIS message was received, and the International Maritime Organization (IMO) number that uniquely identifies a vessel. The first two features strongly correlate with the speed over ground, which is a very important feature, and hence do not provide any additional insight to the model [26]. The timestamp of AIS message reception was omitted as an input feature based on the observation that a route can be traversed multiple times without significant variations in duration. Moreover, the feature did not demonstrate significance according to other researchers' findings [59]. Finally, the IMO was not included in this study as a feature because it is

Table 5.1: Features identified and employed in past papers and this research study.

Feature	Parolas et al [26]	Flapper et al [59]	Kolley et al [60]	This study
Longitude	✓	✓	✓	✓
Latitude	✓	✓	✓	✓
Speed over ground (SoG)	✓			✓
Course over ground (CoG)				✓
Rate of turn (RoT)				✓
True heading			✓	✓
Vessel type		✓	✓	✓
Vessel length	✓	✓		✓
Vessel breadth	✓	✓		✓
Vessel draught				✓
Destination				✓
Distance to destination	✓	✓		✓
ETA provided by the agent	✓			✓
Navigation Status		✓		✓
Speed change compared to 3 previous hours	✓			
Average speed of the last 12 hours	✓			
AIS time received		✓	✓	
IMO number			✓	

not related to the vessel's mobility nor to the route the vessel takes. The complete list of selected features along with their computed feature importance is presented in Table 5.2.

Distance to destination along with speed over ground (SoG) are intuitively important features, serving as the primary metrics of estimation for route completion timeline. Nevertheless, the accuracy of the estimation is affected by other factors represented by the other features. Latitude and longitude are defined as separate values in addition to distance to destination since the Coriolis deflection increases with increasing latitude. A vessel traveling the same distance in higher latitude seas will face much harsher conditions than a vessel traveling in lower latitude [61]. Furthermore, by utilizing the longitude and latitude, the route-specific characteristics can be factored in; e.g., a vessel currently traveling from north to south may be on a route that allows faster transition rather than traveling from east to west. The navigation status indicates idle periods during the course of a vessel, extending the ETA. Metrics of the vessel (i.e., length, breadth, draught, and type) indicate the potential behavior of the vessel related to its physical characteristics, e.g., a large vessel may start decelerating several miles before the port to reach a full stop. Course over ground (CoG) indicates the direction where the ship moves, factoring in longer paths vessels must follow to avoid obstacles and hazards or comply with regulations. The heading indicates where the ship's bow is pointed and may differ from CoG. Usually, this occurs to counteract strong currents that may cause the vessel to drift sideways. The rate of turn (RoT) is a potential indication of the vessels' maneuverability. A less maneuverable vessel in need of performing certain maneuvers to reach its destination will have an increased ETA.

Table 5.2: Features importance based on recursive feature elimination with cross-validation.

Feature Name	Importance (%)
Distance to destination	36.96
ETA provided by the agent	15.50
Navigation Status	14.49
Longtitude	8.13
Speed over ground (SoG)	6.88
Vessel length	3.59
Vessel breadth	3.52
Destination Port	2.78
Heading	2.56
Course over ground (CoG)	1.70
Vessel Draught	1.47
Latitude	1.33
Vessel Type	0.95
Rate of turn (RoT)	0.14

5.1.3 Model Selection and Evaluation

The next important decision after selecting the input features is choosing the most appropriate machine learning algorithm for the ETA prediction model. Three studies were identified during the literature review process that presented good vessel ETA predictions. As previously mentioned, one of the purposes of this work is to test new machine learning algorithms that have not been studied or tested in-depth by other researchers in studies to predict vessel ETAs, along with creating a better feature/input combination. For this purpose, the following six machine/deep learning algorithms were considered:

- **Deep Neural Networks (DNN):** Computational models inspired by the human brain, consisting of layers of interconnected nodes (neurons) that learn from data. The input layer has as many nodes as features while the output layer has only one node for generating the predicted value. Each node has its own associated weight and threshold. When the processed data is above that threshold, the node is activated, sending data to the next layer of the network for further processing. The outcome is a prediction factoring in the assigned weights [56];
- **K-Nearest Neighbors (KNN):** An algorithm that stores all (or a sample of) available data points and classifies new data points based on a similarity measure (e.g., distance functions) and their distance from other data points. The predicted value for the new case is the arithmetic mean of the target values of the K nearest neighbors [60];
- **Decision Trees (DT):** A tree-like model of decisions and their possible consequences, with linear regressors at the leaves. The algorithm splits the data recursively into child nodes at each branch based on their value, attempting to group similar data until a stop criterion is met (e.g., tree depth). A prediction for a new data point is made by following the tree downwards to a leaf node starting from the root. Each leaf contains a linear regression model used to compute the final predicted value [62];

- **Random Forest (RF):** An ensemble method that builds multiple decision trees and merges them to get a more accurate and stable prediction. When a test data point arrives, each decision tree generates a prediction, which is average to generate the final predicted value [63];
- **Extreme Gradient Boosting (XGBoost):** An optimized distributed gradient boosting library designed to be highly efficient, flexible, and portable. XGBoost builds a predictive model by combining the predictions of multiple decision trees, similar to Random Forest, and employs regularization techniques to enhance model generalization [63];
- **Gaussian Naive Bayes (GNB):** A probabilistic classifier based on Bayes' Theorem, assuming strong (naive) independence between features. Each new data point is assigned to a class by calculating the maximum value of the posterior probability of the class. The mean and variance of features are calculated for each class during training, and these statistics are used to predict the class of new points based on the likelihood and prior probabilities.

A common practice for evaluating machine learning models is to randomly split the data into a training and test dataset (e.g., 80% training data, 20% testing data). However, the performance estimate of the model can vary significantly depending on how the data is split, while a single random split provides only one snapshot of the model's performance, which might not be reliable. In addition, if the model is fine-tuned based on the performance on the test set, it can lead to overfitting. To mitigate these downsides, 10-fold cross-validation is employed, which involves splitting the data into 10 folds. For each fold, the other 9 folds are used as the training set and that fold as a test set. A model is fit on the training set and evaluated on the (unseen) test set. This process repeats 10 times, one for each fold, and the evaluation scores are averaged to produce the final evaluation score. This procedure generally results in a less biased or less optimistic estimate of the model's performance than the simple train/test split.

In our specific scenario of arrival time predictions, there is an additional important consideration that is not discussed nor addressed in prior work. In particular, the dataset contains multiple training instances originating from multiple vessel routes. Randomly splitting the data, even for cross-validation, means that training instances from the same vessel route will be contained in both the training and testing splits/folds. As a result, the testing data is infected with data that is very similar to the training data, resulting in in-sample testing. Hence, the evaluation will be biased, and overfitting will occur. To avoid this issue, the study employs an out-of-sample testing methodology that randomly splits the vessel routes into different folds, thereby keeping all data from a single route within a single (training or testing) fold.

To evaluate the quality of the model predictions, key statistical measurements are computed in relation to the difference between model predictions and actual vessel arrival values. The following statistics were recorded in relation to errors: *Mean Absolute Error*, *Mean Squared Error*, and *Root Mean Squared Error*. In addition, to record the relevance of the predictions to the actual values, the *Coefficient of Determination* (R^2 score) and the *Explained Variance* score are also recorded.

5.1.4 Hyperparameter Tuning

In the context of machine learning, hyperparameters are variables specific to each model that control the learning process itself and can impact the model's results. Choosing the optimal hyperparameter values is called *hyperparameter tuning* and is an imperative part of the overall prediction process. Several meth-

ods are available for choosing the optimal hyperparameters, and the three most popular ones are Grid Search, Random Search, and Bayesian Optimization. *Grid Search* will run all possible combinations of the values provided for hyperparameters and choose the combination that produces the best results. The disadvantage of this approach is that it is the most resource-intensive and slow. *Random Search* will select random combinations for testing optimal hyperparameters and return the combination that provided the best outcome after several iterations. While faster and less resource-intensive than Grid Search, it may not always produce the most optimal set of hyperparameters. *Bayesian Optimization* addresses the inefficiency of evaluating several unsuitable combinations seen in Grid and Random Search. It considers the previous evaluation results and uses a probability function to select the following hyperparameter combination that will likely generate better results. This approach reduces the number of iterations needed to identify the (near) optimal hyperparameters. Considering the hyperparameter tuning is needed only once, and enough resources were available, the Grid Search was employed to find the optimal combination of hyperparameter values. For each hyperparameter, the study identified a list of parameter values based on the literature and the specifications manual of each machine-learning model. Next, the Grid Search method generated all possible combinations of the specified hyperparameter values. For each combination in the grid, the model was trained and evaluated through cross-validation using R^2 as the evaluation metric. With this process, the combination of hyperparameters is identified that results in the best model performance. Table 5.3 lists the hyperparameters, values tested, and the optimal values for each model.

Table 5.3: Hyperparameter domains and selected optimal values for all tested machine learning models

	Hyperparameter	**Domain**	**Optimal Value**
DNN	hidden layer sizes	(50,50,50), (50,100,50), (100,1)	(50,50,50)
	activation	identity, relu, tanh, logistic	identity
	alpha	0.0001, 0.05	0.0001
	learning rate	constant, invscaling, adaptive	invscaling
	solver	lbfgs, sgd, adam	lbfgs
KNN	n neighbors	4, 5, 6	6
	weights	uniform, distance	distance
	algorithm	auto, ball_tree, kd_tree, brute	brute
	leaf size	20, 30, 40	20
	p	1,2	1
DT	splitter	best, random	best
	max depth	1, 5, 15, 20, 25	15
	min samples split	10, 50, 100, 136, 150, 200	200
	min samples leaf	100, 1000, 2000, 2221, 3000	100
	min weight fraction leaf	0.0, 0.1, 0.2	0.0
	max features	1, 2, 3, 4	4
	ccp alpha	0.0, 0.1	0.0
RF	trees	60, 80, 100	100
	max features	1, 2	sqrt correct values
	max depth	3, 5, 6, 10	40
	bootstrap	0.001, 0.05, 0.1, 0.15, 0.2	false correct values
	minimum sample split	0.5, 0.8	5
	minimum sample leaf	0.5, 0.8	4
XGB	n estimators	60, 80, 100	100
	min child weight	1, 2	1
	max depth	3, 5, 6, 10	3
	learning rate	0.001, 0.05, 0.1, 0.15, 0.2	0.2
	colsample bytree	0.5, 0.8	0.5
	subsample	0.5, 0.8	0.5
GNB	variance smoothing	1e-11, 1e-10, 1e-9	1e-9

5.2 Evaluation Results

This study evaluates six machine learning algorithms for vessel ETA prediction, namely Multi-layer Perceptron Deep Neural Networks (DNNs), K-Nearest Neighbours (KNN), Decision Tree (DT), Random Forest (RF), Extreme Gradient Boosting (XGBoost), and Gaussian Naive Bayes (GNB). For comparison

purposes, this study implements four model approaches from prior work: ANN by Parolas et al. [26], Gradient Boosting (GBoost) by Flapper et al. [59], DNN by Kolley et al. [60] and KNN by Kolley et al. [60]. For completeness, a comparison is performed against the ETA provided by the shipping agents as well as a simple estimation of the ETA produced using a time calculation based on the current speed and distance remaining to complete the route. All algorithms are implemented in Python 3.6 using the 'scikit-learn' library and all tests were run on a server equipped with two Intel Xeon Silver 4214Y CPU (24 cores @ 2.20GHz) and 128 GB of RAM.

For the six compared algorithms, this study used the features resulting from our evaluation (recall Section 5.1.2) and listed in Table 5.1, as well as the optimal hyperparameter values listed in Table 5.3. For the four algorithms from prior work, the original features and hyperparameter values are employed as presented in the respective papers. The dataset consists of AIS data collected from the Eastern Mediterranean region covering a period of two months. The data contains 172 unique vessel routes towards ports in Cyprus and originating from Europe, Asia, and the Middle East. By using a consistent dataset across all tests, the study ensures the comparability of our results with the approaches proposed in the previous research works. All tests were run using the cross-validation method where the dataset is split in folds while taking into account the vessel routes, as discussed in Section 5.1.3.

Table 5.4: Cross-validation test results for six machine learning algorithms using the features and hyperparameter values identified in this research study.

Evaluation metric	**DNN**	**KNN**	**DT**	**XGBoost**	**RF**	**Gaussian NB**
R^2	0.829	0.817	0.719	0.862	0.880	0.859
Explained Variance	0.847	0.829	0.728	0.871	0.890	0.890
MAE	136.339	129.647	147.077	105.036	99.924	383.269
MSE	47123.605	51873.630	76904.310	39887.408	35862.213	471123.601
RMSE	203.052	210.929	252.011	177.324	163.278	646.277
Max Error	1415.880	1567.755	1726.287	1325.695	1287.346	2936.000

Table 5.4 shows the cross-validation test results for the six compared machine-learning algorithms. To get a holistic view of the results, five statistical evaluation measurements are used, namely Coefficient of Determination (R^2), Explained Variance (EV), mean absolute error (MAE), mean squared error (MSE), root mean squared error (RMSE), and maximum error. R^2 and EV measure how closely predictions align with actual results, with values closer to one indicating better performance. EV does not account for systematic offsets and can be biased, while R^2 remains unbiased by such deficiencies. Hence, a high EV score coupled with a low R^2 score may suggest systematic prediction bias. However, both metrics were consistently close across all models, which indicates the absence of systematic bias in our results, largely due to our proposed data-splitting technique. The highest values for both R^2 and EV were achieved by the RF algorithm with scores of 0.88 and 0.89, respectively. Gaussian NB achieved the same high EV score of 0.89 as RF but had a slightly lower R^2 score (0.86). XGBoost had slightly lower EV (0.87) and R^2 (0.86) scores than RF and Gaussian NB, while DT, as a much simpler algorithm, yielded the lowest performance with 0.72 R^2 and 0.73 EV.

The other four statistical measures allow us to obtain further insights into the results and to better compare our top-performing algorithms. In particular, MAE indicates how much off predictions are from the actual

values in the dataset. MSE serves a similar purpose with the differentiation that squaring amplifies larger differences; hence, it is useful in scenarios where large errors have a higher impact, but it is harder to interpret. RMSE is easier to interpret and combines the advantage of sensitivity to large errors found in MSE. As seen in Table 5.4, RF achieves the lowest scores across all evaluation metrics with an MAE of 99.9 minutes and RMSE of 163.3 minutes. These results are twice as good as the corresponding results of the ETA provided by the shipping agents (discussed further below), which have an MAE of 178.4 and RMSE of 305.2 minutes. Despite the strong performance of Gaussian NB in R^2 and EV, it exhibits the highest maximum error (2936 minutes) and RMSE (646 minutes) among all algorithms, which is 2-2.5$\times$ higher than the values of the next highest DT. Hence, Gaussian NB is able to follow the overall trends of arrival times but makes several large mis-predictions. XGBoost, on the other hand, is only slightly worse (<10%) than RF across all metrics. Finally, DNN and KNN achieve similar performance, which is 20-30% worse than RF across all metrics.

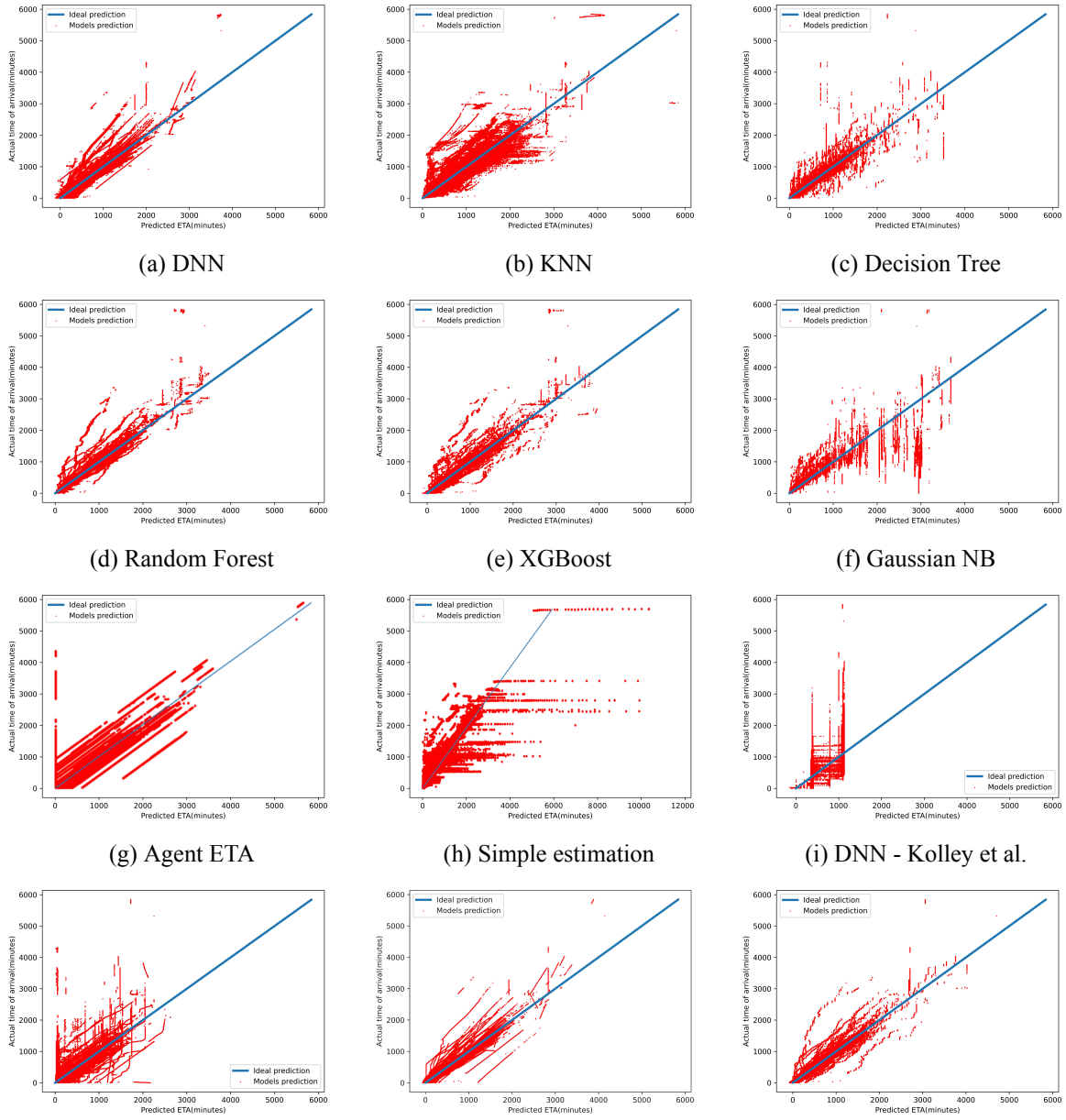
Table 5.5: Cross-validation test results for the ETA provided by the agents, a simple predictor, and four models proposed by other researchers.

Evaluation metric	**Agent ETA**	**Simple Predictor**	**ANN by Parolas et al. [26]**	**GBoost by Flapper et al. [59]**	**DNN by Kolley et al. [60]**	**KNN by Kolley et al. [60]**
R^2	0.655	0.090	0.800	0.827	0.408	0.401
ExplainedVariance	0.656	0.091	0.830	0.837	0.515	0.472
MAE	178.360	175.690	151.309	126.607	261.932	242.204
MSE	93139.420	3212373.610	53686.013	50223.234	147731.612	151881.195
RMSE	305.180	1792.310	215.853	198.778	353.183	379.071
MaxError	4319.000	306747.000	1182.795	1213.396	2289.798	2728.885

Table 5.5 shows the cross-validation test results for the ETA provided by the agents, a simple predictor, and four models proposed by other researchers. The ETA provided by the agents yields an R^2 score of 0.66, 178 minutes ($\sim$3 hours) MAE, and 305 ($\sim$5 hours) RMSE. These values reveal a big problem with ETA predictability that causes various scheduling problems in the destination ports, as discussed in the beginning of the manuscript. Interestingly, these evaluation metrics are all worse compared to the six machine learning models (with the exception of MAE, MSE, RMSE of Gaussian NB), showing that machine learning is a viable alternative. While the MAE achieved by the simple predictor is very close to the agent's ETA, all other metrics are much worse, with an extremely low R^2 score of 0.09. Hence, simple calculations for computing arrival times are unreliable because vessels typically exhibit complex navigational patterns, moving through specific waterways and with variable speeds.

The ANN model proposed by Parolas et al. [26] and the GBoost model proposed by Flapper et al. [59] produced similar results (albeit a bit worse) to the corresponding DNN and XGBoost models. In particular, ANN by Parolas et al. achieved 0.8 R^2 and 151 minutes MAE, while GBoost by Flapper et al. 0.83 R^2 and 126 minutes MAE. On the other hand, the DNN and KNN models by Kolley et al. [60] yielded much lower performance than our models, with R^2 scores of 0.41 and 0.40, respectively, that are even lower than the agents' ETA prediction method. The biggest differentiating factor between these models is the features that are used for training and testing the models. In particular, the models by Parolas et

al. and Flapper et al. use 7 features (plus a few others) out of the 14 features proposed in this research study. On the other hand, Kolley et al. use only 4 out of the 14 proposed features, leaving out some of the most important features, such as distance to destination and navigation status, according to our feature evaluation. Overall, the features identified in this study are important for predicting vessel ETA and can be used effectively for developing strong machine learning models.



(a) DNN (b) KNN (c) Decision Tree

(d) Random Forest (e) XGBoost (f) Gaussian NB

(g) Agent ETA (h) Simple estimation (i) DNN - Kolley et al.

(j) KNN - Kolley et al. (k) ANN - Parolas et al. (l) GBoost - Flapper et al.

Figure 5.3: ETA vs ATA for all proposed machine learning models, the agent's ETA, the simple estimation method, and the models from prior studies.

To further understand the predictive power of each compared approach, Figure 5.3 shows results comparing the ETA and ATA values. The blue diagonal line represents the ideal prediction and the red markings the predictions made by the various models. Our first observation is that the predictions generated by the algorithms tend to be worse for larger ETA and ATA values, which is expected since it is harder to

predict the arrival time when the vessel is far away (e.g., more than 48 hours in advance). The individual plots also give us various interesting insights for the different approaches. For instance, KNN (Fig. 5.3b) seems to underpredict the ETA, while Gaussian NB (Fig. 5.3f) typically over-predicts the ETA. In the plot of the agent ETA (Fig. 5.3g), it is observed that a vertical line of predictions for zero ETA, which indicates that the ETA has already passed but the vessel, in reality, is still a few hours away from its destination. The simple estimation plot (Fig. 5.3h) has many points spread far away to the right of the blue line, indicating a lot of inaccurate over-predictions. Finally, the RF plot (Fig. 5.3d) visually verifies the strong predictive power of RF shown in Table 5.4 since most predictions are mostly gathered around the blue line, with only a few outliers further away.

5.3 Discussion

Accurate ETA prediction has several practical applications, such as management decision support where resources such as pilot boats, tug boats, cranes, berths, personnel, etc., can be scheduled ahead of time using data-driven decisions. Such decisions help in optimizing resource utilization, reducing idle times, cutting delays, reducing costs for port users, and increasing profit for port operators. According to Michaelides et al., [53], an average idle time at the port of Limassol for bulk carriers and tankers is over 8 and 12 hours, respectively, with a significant possibility of reduction when more accurate planning is introduced. The effects of an efficient port stretch far beyond the port itself since ports serve as hubs through which goods are sourced in and out of local markets. Having an efficient data-driven port scheduling process through which resources are optimized means faster delivery times for goods at reduced costs that will benefit the local community being served by the port.

The Maritime industry is a major contributor to greenhouse emissions, which places it under scrutiny [64]. Furthermore, more than 940 million tons of carbon dioxide emissions are estimated to be emitted through marine transportation. Therefore, understanding ship emissions' magnitude and spatiotemporal distribution are crucial for developing effective strategies to reduce emissions [65]. The International Maritime Organization commits to reducing at least 50% of greenhouse emissions from the shipping industry by 2025 [66]. Therefore, accurate ETA predictions are crucial for ports worldwide to minimize vessel waiting times and ensure efficient utilization of port resources. Furthermore, precise ETA prediction supports green shipping practices and benefits the environment. Vessels emit significant amounts of greenhouse gases while idling outside ports, waiting for berths and other resources to become available. Proper planning with accurate ETA prediction, as highlighted by Michaelides et al. [53], can significantly reduce vessel idle time, thereby lowering their carbon footprint. Since ports are often located near densely populated areas, improved air quality resulting from reduced emissions can bring substantial health benefits. Additionally, ports and shipping companies that actively implement green policies are more attractive to investors and more likely to attract business, ultimately increasing their profits and overall value.

Despite the strong predictive performance of the proposed methodology, there are also some potential limitations. First, the accuracy of the predictions for vessels traveling in a particular region is likely to be affected if the model was trained with AIS data from a different region. For example, a model trained with data from the Eastern Mediterranean may not be appropriate for making ETA predictions for vessels traveling in the North Sea or narrow waterways due to differences in vessel routes, sea currents, and

weather conditions. In addition, to ensure accurate predictions over time, it might be needed to retrain the model periodically with fresh AIS data to account for changing vessel routes and the presence of newer ships with more advanced navigational technologies. Finally, any ML-based model cannot make accurate predictions when unexpected events occur. For example, when the Suez Canal was blocked for six days because a container ship had run aground, the ETA of vessels crossing the canal increased dramatically. The aforementioned limitations are all strong candidates for deeper investigation in future research.

AIS data often suffer from inaccurate and missing data that can potentially impact any attempt to accurately predict a vessel's ETA. Prior work has developed a method based on fuzzy matching to automatically clean and correct the manually entered destination field in AIS signals [45]. Others, like Riviero et al. [67], have proposed models to recognize potential anomalies and outlier points in ship routes. This study does not take these issues into account and raw AIS data is used for both training and testing the ML models. The results reveal the ML models are robust and still able to accurately predict ETA despite the potential presence of these issues. Nonetheless, it would be interesting for future work to investigate the impact of AIS data inaccuracies as well as integrate preprocessing techniques to clean, normalize, and remove abnormalities from AIS data into the proposed pipeline.

5.4 Conclusions

The ETA of vessels plays a vital role in terminal operations, as several operations and decisions at terminals depend on it, including berth allocation, berth scheduling, and quay crane assignment. Therefore, this study deals with enhancing ETA predictions of incoming vessels. For this reason, multiple machine/deep learning models, including Deep Neural Networks, K-Nearest Neighbours, Decision Trees, Random Forest, and Extreme Gradient Boosting, have been developed and trained on real-world data collected from multiple AIS base stations located in Cyprus. The study also proposes a new methodology of splitting data based on vessel routes to avoid bias and overfitting during the training and testing of machine learning models By identifying and employing important features and optimal hyperparameters specific to each algorithm, the study achieves higher performance across all relevant metrics (i.e., R^2, Explained Variance, MAE, MSE, RMSE, and Max Error) compared to baseline and other state-of-the-art approaches. Based on the comparative analysis, the study concludes that Random Forest is the best model in predicting ETA (with MAE of 99.92 and RMSE of 163.28), closely followed by XGBoost (with MAE of 105.04 and RMSE of 177.32). In the future, we plan to integrate preprocessing techniques to alleviate any irregularities or errors that are associated with AIS data, as well as implement automated periodic retraining of the model to ensure it can retain accurate ETA predictions over time.

The ETA prediction research in this chapter directly leverages the AIS data acquisition and preprocessing framework, including fuzzy matching techniques, presented in earlier chapters. The reliability of the AIS data processing pipeline ensures high-quality, well-aligned datasets, enabling the robust training and evaluation of machine and deep learning models for ETA estimation. While this chapter focuses on predicting arrival times to optimise port operations and resource allocation, it complements the trajectory prediction work presented in the next chapter by addressing the temporal dimension of vessel movements.

6 Vessel Trajectory Prediction with Deep Learning: Temporal Modeling and Operational Implications

Maritime transportation is a key enabler of global trade, facilitating the movement of goods and people across vast oceanic, sea, and inland waterways. A critical component in maritime navigation and safety is vessel trajectory prediction, which is the task of forecasting the future positions or movement paths of ships based on their past and current movement data, such as location, speed, and heading. It can be used for short-term predictions (e.g., the next few minutes) or long-term forecasts (e.g., hours ahead), and is essential for optimizing maritime logistics [68,69], enhancing navigational safety [70,71], and supporting autonomous and unmanned vessels [72,73]. Given the increasing complexity of maritime operations and the advent of digitalization in the shipping industry, leveraging Automatic Identification System (AIS) data has become an integral aspect of maritime research and operational planning [74].

Vessel trajectory prediction is a complex task influenced by a multitude of factors, including environmental conditions, regulatory constraints, vessel interactions, and human decision-making factors [75]. Traditional methods, such as statistical models, kinematic models, and physics-based simulations, often fail to capture the intricate and dynamic nature of ship movements [76]. Recent advancements in Deep learning (DL) techniques have demonstrated significant potential in modeling complex spatiotemporal dependencies, making them well-suited for vessel trajectory prediction [77,78]. Furthermore, deep learning (DL) models such as recurrent neural networks (RNNs), long short-term memory (LSTM) networks, and transformer-based architectures, have exhibited superior predictive accuracy by learning from vast historical AIS datasets [34,35,40,79].

Despite the progress in this domain, several challenges remain unresolved. Data sparsity, varying sampling rates, and the presence of anomalous AIS signals introduce complexities in predictive modeling [80]. Furthermore, the integration of external factors such as weather conditions, ocean currents, and geopolitical influences into predictive frameworks remains an area of active research [81]. While some studies have successfully implemented hybrid models combining ML with physics-based simulations, a consensus on the most effective approach is yet to be established [76]. Controversies persist regarding the trade-offs between model interpretability and predictive performance, particularly when deploying black-box ML models in safety-critical maritime applications [82].

The primary objective of this study is to develop a robust deep learning-based framework for vessel trajectory prediction using real-world AIS data, with an emphasis on understanding temporal prediction dynamics and model generalizability. The main contributions of this study are as follows:

- A systematic evaluation of three deep recurrent neural architectures, LSTM, Bi-LSTM, and Bi-GRU, for short-term vessel trajectory prediction, extending prior maritime sequence modeling work.
- A comparative analysis of model performance across short-term prediction horizons, providing insights into the temporal dynamics and degradation patterns of each architecture.

- An in-depth investigation of the most robust model's behavior under extended prediction windows, establishing practical thresholds for reliable long-term trajectory forecasting.

The study contributes to the ongoing discourse in maritime informatics and intelligent transportation systems by addressing key methodological gaps and offering insights into the practical deployment of DL-driven predictive analytics in real-world maritime operations [83]. In doing so, it provides a structured and empirical foundation for selecting and deploying trajectory forecasting models in operational contexts. The findings have practical implications for stakeholders in maritime logistics, port management, and regulatory oversight, where predictive accuracy and horizon-specific model robustness are critical to informed decision-making and navigational safety.

The subsequent sections elaborate on the proposed methodology (Section 6.1), detail the experimental evaluation and results (Section 6.2), discuss the findings and implications of this study (Section 6.3), and conclude with implications and directions for future research (Section 6.4).

6.1 Methodology

6.1.1 Data Collection and Processing

This study utilizes Automatic Identification System (AIS) data collected from 16 terrestrial AIS base stations installed along the coast of Cyprus, covering the eastern Mediterranean Sea [57]. AIS data spanning a two-month period (June 1, 2024 – July 31, 2024) from the Eastern Mediterranean region was selected for training and evaluating the deep learning models. To ensure high-quality and structured input for model training, the following pre-processing steps were applied:

1. **Filtering:** Only AIS records where the vessel was actively navigating (i.e., with a navigation status *NavStatus* = 0) were retained, eliminating transmissions from stationary or inactive vessels.
2. **Sorting:** AIS messages were ordered by vessel identifier and timestamp to preserve temporal continuity.
3. **Route Segmentation:** AIS messages were grouped by the vessel's unique IMO number to identify distinct voyages. If a transmission gap exceeding 6 hours was detected, the trajectory was split into separate segments to ensure temporal coherence.
4. **Outlier Detection and Validation:** Outliers, such as unrealistic jumps in location or implausible speeds, were removed to preserve smooth and realistic trajectories.
    - **Speed-Based Distance Filtering:** Each point was validated against the maximum possible travel distance, computed using:

$$\text{max\_distance} = \frac{\text{speed}_{\text{km/h}} \times \text{time\_diff}}{3600}$$

      Points exceeding this threshold were flagged as physically implausible and removed.
    - **Rate of Turn (RoT) Constraints:** The RoT, measured in degrees per minute, was checked against known operational limits for different vessel types (see Table 6.1). RoT values were

normalized using:

$$RoT = \left(\frac{\text{raw RoT}}{4.733} \right)^2 \times \text{sgn}(\text{raw RoT})$$

This normalization is commonly used in maritime anomaly detection to improve data consistency [84, 85].

5. **Temporal Resampling and Interpolation:** All trajectories were resampled at uniform 30-second intervals. Missing values were interpolated, and in cases with multiple messages per interval, mean aggregation was applied to ensure data smoothness.

Table 6.1: Typical RoT ranges for different vessel types. Data provided by [1].

Vessel Type	**Reasonable RoT (°/min)**	**Notes**
Small Boats/Yachts	20–60	Agile, capable of sharp turns
Tugboats	30–80	Designed for high maneuverability
Cargo Ships (Large)	3–6	Slow, wide turning radius
Container Ships	2–5	Inertia-limited maneuvering
Tankers (VLCC, ULCC)	1–4	Extremely slow due to mass
Passenger Ships	5–10	Moderate agility
Naval Warships	10–30	High-speed maneuvering capacity

The final dataset was randomly partitioned into a training set containing 4094 vessel routes with a total of 10,236,895 AIS points and a testing set containing 100 vessel routes with an additional 232,766 AIS points. To prevent data leakage and model overfitting, entire vessel trajectories were assigned exclusively to one of the two sets (further elaborated in Section 6.1.5). The two datasets include vessels from 37 different ship types, such as general cargo, container, Ro-Ro, and tanker vessels, capturing diverse navigational behaviors. Table 6.2 lists the frequencies of each vessel type in the training and testing datasets. Moreover, the testing data spans 43 unique destination ports across 16 countries, as shown in Figure 6.1, introducing variability in routing patterns and operational contexts (e.g., international shipping lanes, regional cargo routes, and port approaches).

In real-world deployments, the data pre-processing pipeline described in Steps 1–5 can be implemented in real time prior to inference. This ensures that incoming AIS data is cleaned and standardized on the fly, enabling the model to operate reliably under dynamic, real-time conditions.

Table 6.2: Frequency of vessel types in the training and testing datasets.

Vessel Type	Training Frequency	Testing Frequency
General Cargo	2200	51
Tanker	927	21
Other	218	5
Container Ship	101	7
Bulk Carrier	94	6
Passenger	92	
Oil/Chemical Tanker	90	
Unknown	49	
Ro-Ro Cargo	42	4
Conventional	27	
Livestock Carrier	26	
Crude Oil Tanker	24	1
Bulk/Conventional	24	
Passenger Ship	22	1
Container	18	1
LPG Tanker	18	1
Chemical Tanker	13	1
Oil Products Tanker	12	
Tug	9	
Multi Purpose Offshore Vessel	7	
Offshore Supply Ship	7	
Special Purpose Vessel	7	
Cargo/Container Ship	6	
Vehicles Carrier	5	
Offshore Support	4	1
Yacht	4	
Edible Oil Tanker	3	
Supply Vessel	3	
Platform	2	
Ro-Ro/Container Carrier	2	
Ro-Ro/Passenger Ship	2	
Ro-Ro	1	
Cement Carrier	1	
Cruise	1	
Hopper Dredger	1	
Military	1	
Offshore Support	1	

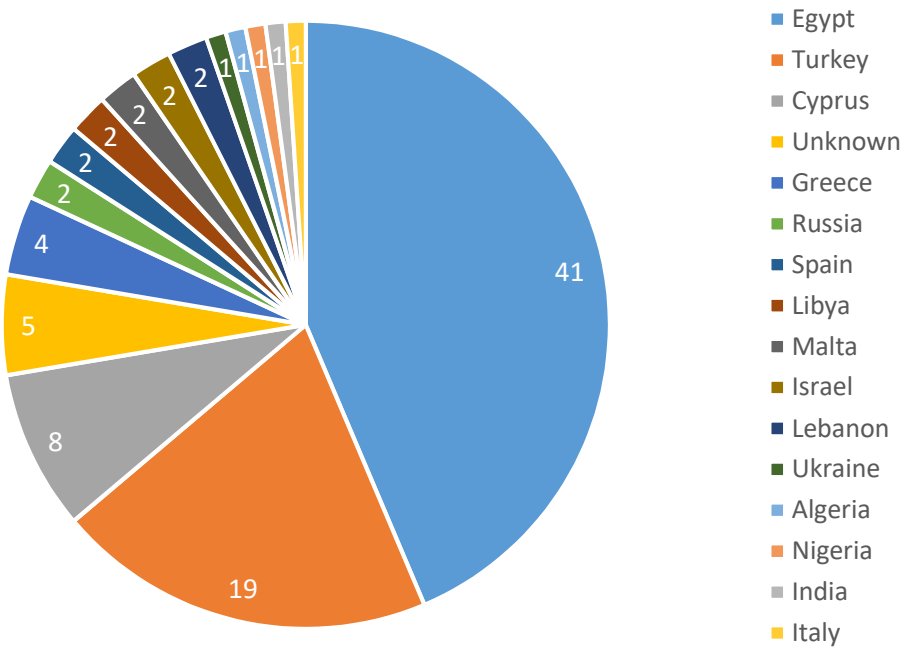


Figure 6.1: Frequency of destination countries in the testing dataset.

6.1.2 Feature Selection

The selection of input features is a critical component in developing accurate predictive models. Despite the wide availability of AIS data attributes, there is limited research focused on a systematic analysis of which features most effectively contribute to predictive performance. A review of relevant studies reveals a consistent use of four key dynamic features: longitude, latitude, course over ground (CoG), and speed over ground (SoG). These features have been empirically validated across multiple papers [34, 35, 40] and strike a balance between informativeness and simplicity.

In this study, we adopt the same four core features to ensure comparability with prior work and to mitigate the risk of overfitting associated with high-dimensional input spaces. We also evaluated the inclusion of additional static and dynamic attributes, such as vessel type, rate of turn, and true heading, and observed that their contribution to short-term trajectory prediction was consistent with findings reported in earlier studies. These results confirm that while such features may offer marginal improvements under specific conditions, they do not generalize well, and the core set remains robust and effective for the task at hand.

6.1.3 Model Selection

The next phase of this study involves the selection and systematic evaluation of deep learning architectures previously validated in the maritime domain. The selected models, Long Short-Term Memory (LSTM), Bidirectional Long Short-Term Memory (Bi-LSTM), and Bidirectional Gated Recurrent Units (Bi-GRU), have demonstrated superior performance in predictive modeling tasks involving spatiotemporal data, particularly in AIS-based vessel trajectory forecasting [34, 35, 40]. These architectures are adopted directly from influential peer-reviewed studies, thereby ensuring methodological consistency with established benchmarks in the literature.

The LSTM model, originally proposed by Hochreiter and Schmidhuber [86], has been widely employed in trajectory modeling due to its ability to capture long-term dependencies and mitigate the vanishing gradient problem [87]inherent in standard RNNs. In maritime informatics, LSTMs have shown strong predictive capability across tasks such as route prediction and vessel behavior modeling [35, 88]. Most existing LSTM-based studies focus on short input sequences (e.g., using four or five past positions, often spaced at 1–2 minute intervals) and typically generate a single next-step prediction, as seen in [34, 35].

To further improve the modeling of bidirectional temporal dependencies, the Bi-LSTM architecture extends LSTM by introducing a second recurrent layer that processes input sequences in reverse order. This allows the model to incorporate both past and future contextual information, which is particularly advantageous when predicting vessel trajectories in dynamic maritime environments. Prior studies, such as Wang et al. [89], have demonstrated the superior performance of Bi-LSTM in predicting complex vessel movement patterns. However, these works also tend to rely on short input horizons and typically predict vessel positions at fixed short-term intervals—such as 1, 2, 3, or 6 minutes ahead—rather than forecasting the vessel's full trajectory or assessing prediction stability over extended time horizons.

Finally, the Bi-GRU model, employed in this study as presented in Li et al. [34], offers a computationally efficient alternative to Bi-LSTM while maintaining the capability to capture bidirectional dependencies. GRUs simplify the memory gating mechanisms of LSTM, and the bidirectional configuration further enhances their capacity to model intricate temporal structures in AIS data. Li et al.'s study evaluated twelve architectures using four input timesteps (spaced at 2-minute intervals) to predict a single 5th position, and concluded that Bi-GRU performed best in that short-term context.

By adopting these architectures, the present study builds upon a well-established empirical foundation in the literature, while introducing a methodologically structured approach to temporal sequence modeling.

6.1.4 Hyperparameters

In the context of deep learning, hyperparameters are predefined variables that govern the model's learning behavior and can significantly influence predictive performance. To ensure methodological consistency and alignment with established best practices in AIS-based vessel trajectory prediction, this study adopts hyperparameter configurations directly from the literature. For the Bi-GRU model, hyperparameters such as the number of number of neurons, dropout rate, and learning strategy are applied as specified in Li et al. [34]. Similarly, the LSTM and Bi-LSTM models utilize configurations from Capobianco et al. [35] and Zhang et al. [40], respectively. These settings are implemented without modification to maintain fidelity to validated architectures and to minimize variability arising from manual tuning. This approach prioritizes reproducibility and enables a controlled investigation of input sequence design and comparative model performance. Table 6.3 lists the hyperparameters used for each model.

6.1.5 Experimental Design and Evaluation

In this study, model evaluation is conducted using an out-of-sample testing methodology tailored to the specific challenges of AIS-based vessel trajectory prediction [90]. Rather than relying on traditional random data splits, which risk contamination between training and testing sets due to the presence of overlapping route patterns, the evaluation strategy randomly isolates 100 entire vessel routes exclusively

Table 6.3: Hyperparameter values for the three considered deep learning models.

Hyperparameter	**LSTM**	**Bi-LSTM**	**Bi-GRU**
Learning Rate	0.0001	0.001	0.0001
Dropout	[0.2, 0.2]	[0.2, 0.2]	[0.5, 0.5]
Neurons	[64, 64]	[200, 100]	[128, 128]
Activation	tanh	tanh	–
Epochs	100	100	100

for testing purposes. This ensures that the model is assessed only on completely unseen navigational data, offering a more rigorous and realistic measure of generalization performance. The selection of a fixed number of routes balances the need for statistical robustness with computational feasibility, as the scale of the dataset and the intensive testing demands make exhaustive evaluation across all routes impractical. This approach mitigates the risk of overfitting and yields performance metrics that more accurately reflect the model's behavior in operational maritime environments.

The models will be trained using input sequences of varying temporal lengths: 30 seconds, 2 minutes, 10 minutes, and 20 minutes, enabling a nuanced investigation of model performance across varying temporal granularities. Testing will be conducted in two distinct modes. In the first mode, each temporal sequence will consist of real, observed AIS data, and the model will be tasked with predicting the subsequent data point. In the second mode, previously predicted points will be recursively used as inputs for the next prediction step, thereby simulating a sliding window mechanism. This dual-mode evaluation is designed to assess both the one-step-ahead predictive accuracy and the robustness of each model when relying on its own prior outputs, offering a more comprehensive understanding of model performance in practical deployment scenarios.

Furthermore, we assess the forecasting capabilities of the best-performing configuration identified in our experiments, namely, the Bi-LSTM model, at prediction horizons of 10, 20, and 60 minutes. This facilitates a systematic comparison between short-term and long-term predictive robustness under conditions that reflect realistic maritime operational settings. The proposed experimental framework not only examines model sensitivity to input resolution, but also evaluates the Bi-LSTM model's capacity to mitigate the compounding effects of recursive error over extended forecast durations. Ultimately, the study aims to provide a comprehensive assessment of the predictive accuracy and temporal adaptability of recurrent neural architectures across diverse forecasting regimes and input configurations, with a particular focus on Bi-LSTM performance in long-term trajectory prediction.

To evaluate the quality of the model predictions, key evaluation metrics that capture both spatial accuracy and trajectory shape over time are computed in relation to the difference between model predictions and actual vessel locations [91]. Specifically, we compute (1) point-wise error metrics to measure how close the predicted coordinates are to the actual ones, including Mean Absolute Error (MAE), Mean Squared Error(MSE), Root Mean Squared Error(RMSE), and Symmetric Mean Absolute Percentage Error (SMAPE); (2) trajectory-level metrics that evaluate the entire predicted path vs. the ground truth path, namely, the Average Displacement Error (ADE) and the Final Displacement Error (FDE), which are used to evaluate the positional accuracy of the predicted trajectories. The ADE computes the average

distance between each predicted position (x_i^{pred}) and corresponding actual vessel position (x_i^{actual}) in a vessel trajectory over a prediction horizon of n time steps, using the following equation:

$$\text{ADE} = \frac{1}{n} \sum_{i=1}^{n} H(x_i^{\text{pred}}, x_i^{\text{actual}}) \tag{6.1}$$

The position distances are computed using the Haversine formula H, which accounts for Earth's curvature when computing the distance between two latitude/longitude points. The FDE measures the distance between the predicted and actual position at the final forecasted time step, providing an assessment of the endpoint accuracy.

$$\text{FDE} = H(x_n^{\text{pred}}, x_n^{\text{actual}}) \tag{6.2}$$

To assess the similarity of the overall trajectory shape, we also use the Discrete Fréchet Distance (DFD) [92]. Unlike ADE and FDE, which focus on individual point errors, DFD captures the similarity between the entire predicted and actual trajectory paths. It takes into account the spatial arrangement and order of points along both curves, making it particularly suitable for evaluating the global structure of vessel movements.

All experiments were conducted on a server equipped with two Intel Xeon Silver 4214Y CPUs, with 24 logical cores, 2.20 GHz, and a 64-bit architecture each, 128 GB memory, and a 2 TB hard drive.

6.2 Results

This section presents a comprehensive analysis of the performance of the deep learning models applied to the vessel trajectory prediction task. The evaluation focuses on quantifying predictive accuracy across varying temporal input lengths and sequence conditions, using both real and recursively predicted inputs for short-term and long-term predictions, respectively. Multiple error metrics are employed to ensure a robust assessment, including traditional point-wise statistical measures as well as spatial distance-based indicators. These metrics collectively enable a nuanced comparison of model behavior in short-term versus long-term prediction scenarios and offer insights into each model's ability to capture spatiotemporal vessel movement patterns under different operational regimes.

The results not only facilitate a direct comparison between LSTM, Bi-LSTM, and Bi-GRU architectures, but also support the evaluation of the models' robustness when exposed to input sequences of increasing duration and complexity. Moreover, by systematically distinguishing between modes using actual versus recursively predicted inputs, the study aims to reveal the compounding effect of prediction errors in real-time applications. These outcomes directly address the core objectives of the study: to evaluate the impact of sequence input methods (Section 6.2.1), assess model performance under varying temporal horizons (Section 6.2.2), and determine the threshold beyond which long-term predictions begin to deteriorate (Section 6.2.3). In doing so, the findings offer practical implications for the deployment of trajectory prediction systems in operational maritime environments.

6.2.1 Short-Term Prediction Analysis

Short-term prediction serves as an idealized scenario in which future inputs are assumed to be available during inference. This setup is commonly used to assess the upper bound of a model's learning capacity, as it typically results in lower prediction errors and reduced sensitivity to cumulative forecasting inaccuracies. Under such conditions, recurrent models, particularly those based on LSTM variants, tend to perform well, benefiting from the availability of accurate context and temporal features, which mitigate the effects of error propagation and enhance both spatial and kinematic forecast precision.

This section provides a comprehensive evaluation of short-term prediction performance, specifically focusing on single-step ahead forecasts for key vessel trajectory parameters: latitude, longitude, speed over ground (SoG), and course over ground (CoG). Three recurrent neural network architectures, Bi-LSTM, LSTM, and Bi-GRU, were assessed across multiple sequence lengths (1, 4, 20, and 40 points), utilizing the previously defined feature set. The corresponding results for each model and sequence length combination are presented in Tables 6.5, 6.4, and 6.6, respectively. All tables show the average metric across the 100 test routes.

6.2.1.1 LSTM Model

As shown in Table 6.4, the LSTM model demonstrated reasonably good performance with increasing training sequence size. Among the tested configurations, the model trained with a sequence length of 20 offered the most balanced trade-off between accuracy and stability. It achieved an Average Displacement Error (ADE) of 5.34 km and a Final Displacement Error (FDE) of 5.11 km, alongside a $MAE_{lat} = 0.0225$ and $MAE_{lon} = 0.0472$. However, despite its relative improvement in spatial prediction over shorter sequence lengths, the model's kinematic accuracy remained limited, with $MAE_{sog} = 0.31$ and $MAE_{cog} = 2.35$.

Notably, increasing the sequence length beyond 20 (i.e., to 40 points) did not yield substantial performance gains; in some metrics, such as MAE_{lat} and ADE, slight degradations were observed. The model trained with only a single timestep exhibited particularly poor results, with a FDE of 16.48 km, $MAE_{lat} = 0.08$, and $MAE_{lon} = 0.14$, illustrating the importance of temporal context in accurate short-term forecasting.

6.2.1.2 Bi-LSTM Model

The Bi-LSTM architecture consistently demonstrated strong predictive capabilities across all evaluated metrics listed in Table 6.5. Notably, the model trained with a sequence length of 40 outperformed all others, achieving the lowest errors in most categories. It recorded the smallest Mean Absolute Error (MAE) for both spatial and kinematic features ($MAE_{lat} = 0.009$, $MAE_{lon} = 0.013$, $MAE_{sog} = 0.15$, and $MAE_{cog} = 1.58$. The corresponding Root Mean Squared Error (RMSE) values further confirmed this trend, with the model achieving $RMSE_{lat} = 0.016$, $RMSE_{lon} = 0.028$, and the lowest angular error $RMSE_{cog} = 8.10$.

In terms of spatial accuracy, the model also achieved the best results with an ADE of 1.68 Km and an FDE of 1.77 Km, indicating close alignment between the predicted and actual trajectories. At open sea, such slight deviations are expected and often acceptable. The Discrete Fréchet Distance (DFD) of 8.43 km reveals that the overall shape of the predicted trajectory deviates in at least one region of the route,

Table 6.4: Short-term prediction metrics when using LSTM with multiple training sequence sizes.

Training Sequence	1 point	4 points	20 points	40 points
ADE	16.7346	6.7630	5.3435	5.5197
FDE	16.4824	7.0457	5.1156	5.1840
DFD	26.7758	14.5577	13.2993	13.0049
MAE				
Latitude	0.0761	0.0401	0.0225	0.0289
Longitude	0.1402	0.0445	0.0472	0.0444
SoG	0.3516	0.2474	0.3144	0.2413
CoG	7.5847	1.8875	2.3544	2.3091
MSE				
Latitude	0.0132	0.0024	0.0010	0.0012
Longitude	0.0344	0.0037	0.0037	0.0039
SoG	3.2432	1.9329	2.4626	1.9372
CoG	167.0330	75.6857	75.7494	72.6262
RMSE				
Latitude	0.1149	0.0489	0.0308	0.0354
Longitude	0.1854	0.0609	0.0605	0.0626
SoG	1.8009	1.3903	1.5693	1.3919
CoG	12.9229	8.6997	8.7028	8.5209
SMAPE				
Latitude	0.2374	0.1237	0.0664	0.0872
Longitude	0.4480	0.1462	0.1536	0.1419
SoG	17.2383	14.7116	15.4403	13.8862
CoG	6.4324	2.6641	2.8363	2.4974

potentially during sharp turns, but eventually corrects itself, especially given that FDE is much lower. These results demonstrate that increasing the sequence length significantly enhances Bi-LSTM's ability to capture temporal dependencies and mitigate predictive error propagation, particularly for short-term trajectory forecasting.

6.2.1.3 Bi-GRU Model

As seen in Table 6.6, the Bi-GRU model demonstrated inconsistent performance across sequence lengths, with notable degradation at longer input horizons. At shorter sequence lengths (1 and 4), the model achieved reasonably competitive results. For instance, the configuration with a sequence length of 4 yielded ADE = 6.37 Km, FDE = 6.10 km, MAE_{lat} = 0.04, and MAE_{lon} = 0.04, indicating satisfactory short-term predictive performance. The associated DFD of 13.76 km further reflects adequate trajectory alignment.

However, as the sequence length increased to 20 and 40, the model's predictive accuracy deteriorated markedly. At 40 points, the errors escalated sharply (MAE_{lat} = 0.37, MAE_{lon} = 0.30, and MAE_{cog} = 6.42), while both ADE and FDE exceeded 26 km. The combination of spatial and directional errors suggests a breakdown in trajectory coherence. Similarly, large increases in MSE and RMSE, especially in CoG and SoG, indicate numerical instability and high predictive error. These results suggest that while

Table 6.5: Short-term prediction metrics when using Bi-LSTM with multiple training sequence sizes.

Training Sequence	1 point	4 points	20 points	40 points
ADE	3.1380	3.8658	2.4288	1.6815
FDE	3.2490	3.7644	2.5727	1.7685
DFD	9.8687	9.8279	9.9159	8.4290
MAE				
Latitude	0.0178	0.0242	0.0116	0.0086
Longitude	0.0220	0.0245	0.0195	0.0130
SoG	0.2818	0.1750	0.1978	0.1545
CoG	2.8532	2.0957	1.7525	1.5799
MSE				
Latitude	0.0008	0.0013	0.0003	0.0002
Longitude	0.0012	0.0016	0.0012	0.0008
SoG	3.2278	2.0311	1.7693	1.8695
CoG	111.0046	74.2322	71.2995	65.6492
RMSE				
Latitude	0.0277	0.0367	0.0182	0.0155
Longitude	0.0346	0.0399	0.0341	0.0281
SoG	1.7955	1.4252	1.3292	1.3664
CoG	10.5359	8.6158	8.4380	8.0999
SMAPE				
Latitude	0.0538	0.0741	0.0349	0.0262
Longitude	0.0719	0.0791	0.0630	0.0417
SoG	15.7211	13.3554	13.9570	13.9390
CoG	4.5636	2.8249	2.3403	2.1848

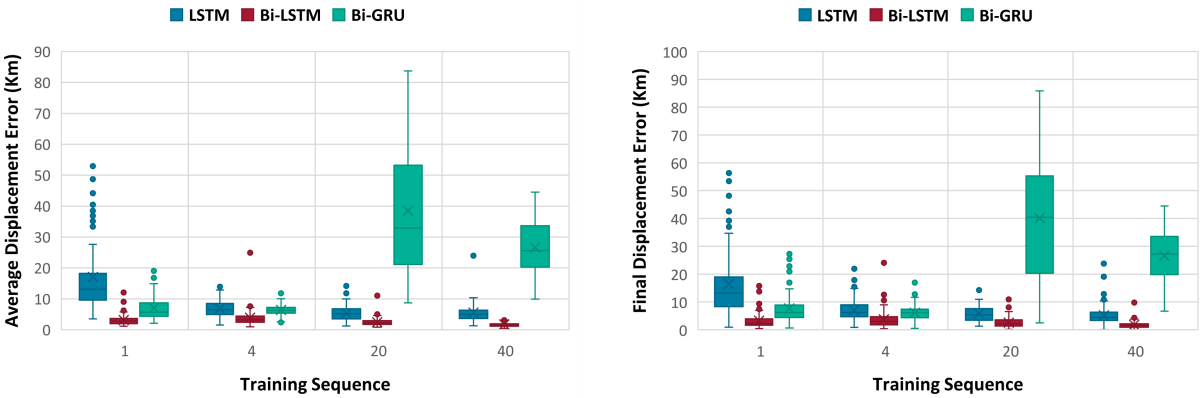
Bi-GRU is capable of handling short input sequences effectively, it struggles to scale with longer temporal dependencies, likely due to limitations in its gating mechanism or vanishing gradient effects.

6.2.1.4 Summary and Comparative Insights

Figure 6.2 visualizes the ADE and FDE distributions using box and whisker plots for the short-term prediction of the three models across the four training sequences. Across all models and configurations, Bi-LSTM with longer sequence lengths demonstrated the most robust and accurate short-term forecasting. LSTM models performed well overall but experienced the most outliers, i.e., the trajectory predictions for some routes are very inaccurate with very high ADE and FDE. While LSTM models benefit from increased sequence length (up to a point), their capacity to capture bidirectional dependencies and mitigate error propagation remains limited in comparison to Bi-LSTM. In contrast, Bi-GRU models struggled to maintain performance at higher sequence lengths, raising concerns about their scalability and stability for short-term vessel trajectory forecasting. This performance divergence underscores the relative fragility of the Bi-GRU architecture in deep sequence modeling compared to Bi-LSTM and LSTM counterparts. These findings underscore the critical role of architectural design and temporal depth in achieving reliable short-horizon predictions.

Table 6.6: Short-term prediction metrics when using Bi-GRU with multiple training sequence sizes.

Training Sequence	1 point	4 points	20 points	40 points
ADE	6.9709	6.3675	38.0837	26.5513
FDE	7.6473	6.1018	40.1789	26.6074
DFD	11.9392	13.7604	49.4036	32.4359
MAE				
Latitude	0.0337	0.0405	0.2291	0.3712
Longitude	0.0512	0.0411	0.7686	0.2978
SoG	0.4495	0.3475	0.5529	0.3488
CoG	2.7593	2.6614	14.6638	6.4242
MSE				
Latitude	0.0027	0.0021	0.0800	0.1615
Longitude	0.0046	0.0029	0.8034	0.1351
SoG	3.3883	2.1620	2.4171	2.1014
CoG	107.7442	79.5674	444.1331	130.2928
RMSE				
Latitude	0.0522	0.0459	1.8782	1.9616
Longitude	0.0698	0.0529	2.1494	2.8986
SoG	0.6889	0.5239	1.0310	1.5765
CoG	4.2650	3.6621	16.9843	13.0935
SMAPE				
Latitude	0.1045	0.1212	0.7100	1.1107
Longitude	0.1701	0.1352	2.4374	0.9743
SoG	18.7340	14.3645	19.5429	16.6608
CoG	4.0088	2.6420	8.6845	5.0689



(a) ADE (b) FDE

Figure 6.2: Distribution of average displacement error (ADE) and final displacement error (FDE) for the short-term prediction of the three models across the four training sequences.

In addition to predictive accuracy, we evaluated the computational performance of the models to better understand their suitability for real-world deployment. Specifically, we measured training time (in minutes per epoch) and inference time (in milliseconds per prediction) across varying input sequence lengths, shown in Table 6.7. As expected, increasing the input training sequence from 1 to 40 data points led to a

rise in training time for all models from approximately 30 minutes per epoch for LSTM to 400 minutes for Bi-LSTM. Inference time also increased but very modestly, from 52 ms to 62 ms per prediction for all models. While training time scales significantly with sequence length due to increased model complexity and temporal depth, the relatively low inference latency indicates that the models remain suitable for near-real-time trajectory prediction applications.

Table 6.7: Training and inference times for models with different training sequence sizes.

Model	Training Sequence	Training Time (min/epoch)	Inference Time (ms/prediction)
LSTM	1	28.75	52.88
LSTM	4	38.91	52.96
LSTM	20	103.58	56.01
LSTM	40	193.53	59.43
Bi-LSTM	1	51.04	53.70
Bi-LSTM	4	75.15	54.49
Bi-LSTM	20	236.73	58.67
Bi-LSTM	40	418.18	62.83
Bi-GRU	1	46.14	53.60
Bi-GRU	4	63.22	54.16
Bi-GRU	20	180.15	58.38
Bi-GRU	40	296.40	62.01

6.2.2 Long Term Prediction Analysis Across Horizons

In trajectory forecasting tasks, the prediction horizon plays a critical role in determining model performance. While short-term prediction assumes access to ground-truth inputs at each time step, long-term prediction presents a more realistic and challenging scenario. In this setting, the model recursively uses its own previous outputs as inputs for future steps, potentially introducing cumulative error propagation.

In this section, we evaluate long-term prediction performance using the Bi-LSTM architecture trained on 40-point input sequences, since it exhibited the best predictive performance for single-point predictions. The evaluation is conducted across three prediction horizons: 20 points (10 minutes), 60 points (30 minutes), and 120 points (60 minutes). These configurations allow us to systematically assess how prediction accuracy behaves as the forecasting window increases, thereby providing insight into the model's ability to maintain reliable performance over time.

Table 6.8 shows the long-term prediction results across the multiple prediction horizons. As expected, even high-performing models such as Bi-LSTM experience performance degradation under long-term prediction scenarios as the horizon length increases. Predictions for the next 20 points (i.e., next 10 minutes) yield satisfactory results with ADE = 5.3 Km and FDE = 8.5. The DFD of 8.7 km suggests the worst-case trajectory deviation occurs early or stabilizes, leading the predicted path to remain within a rough spatial corridor, even if points are farther off on average. As the forecast horizon extends beyond the initial 40-point training sequence, errors accumulate due to the model's recursive reliance on its own previous outputs, resulting in compounded inaccuracies. This deterioration is evident across displacement-based metrics as ADE and FDE increase to 18.6 Km and 32.8 Km, respectively, reflect-

ing diminished spatial precision. Similarly, trajectory shape divergence, as measured by DFD, intensifies with longer horizons up to 33.5 Km. Further, point-based error metrics for key dynamic features (i.e., latitude, longitude, SoG, and CoG) exhibit pronounced growth, with angular variables such as CoG showing particularly steep error escalations.

Table 6.8: Long-term prediction metrics when using Bi-LSTM with a 40-point training sequence across multiple prediction horizons.

Prediction Sequence	**20 points (10 mins)**	**60 points (30 mins)**	**120 points (60 mins)**
ADE	5.3052	11.1432	18.6331
FDE	8.5026	19.0673	32.7800
DFD	8.7431	19.6947	33.5118
MAE			
Latitude	0.0226	0.0462	0.0771
Longitude	0.0448	0.0953	0.1602
SoG	0.8982	1.7570	2.4021
CoG	12.6513	30.6923	46.0783
MSE			
Latitude	0.0027	0.0062	0.0143
Longitude	0.0048	0.0187	0.0507
SoG	8.6617	22.5692	39.0892
CoG	644.5075	2406.9327	4742.5689
RMSE			
Latitude	0.0517	0.0786	0.1197
Longitude	0.0695	0.1367	0.2252
SoG	2.9431	4.7507	6.2517
CoG	25.3866	49.0604	68.8845
SMAPE			
Latitude	0.0675	0.1379	0.2302
Longitude	0.1499	0.3207	0.5389
SoG	24.1447	35.9006	44.4687
CoG	8.7546	21.3748	32.1960

To qualitatively illustrate this behavior, Figure 6.3 presents visual comparisons between the ground-truth routes and Bi-LSTM predictions at 20, 60, and 120-point horizons for three representative vessel trajectories. The figures highlight the increasing divergence between predicted and actual paths as the forecast horizon lengthens, particularly in terms of directional drift and spatial displacement. The highest prediction errors in long-term vessel trajectory forecasting typically occur near regions where the vessel undergoes significant changes in direction. These directional transitions are often associated with navigational decisions, environmental responses (e.g., wind, current), or operational constraints, which introduce abrupt changes in movement patterns. Predictive models, especially those that rely heavily on short-term temporal dependencies, tend to struggle in capturing such non-linear dynamics. As a result, prediction accuracy degrades in these regions, as evidenced by localized spikes in displacement error.

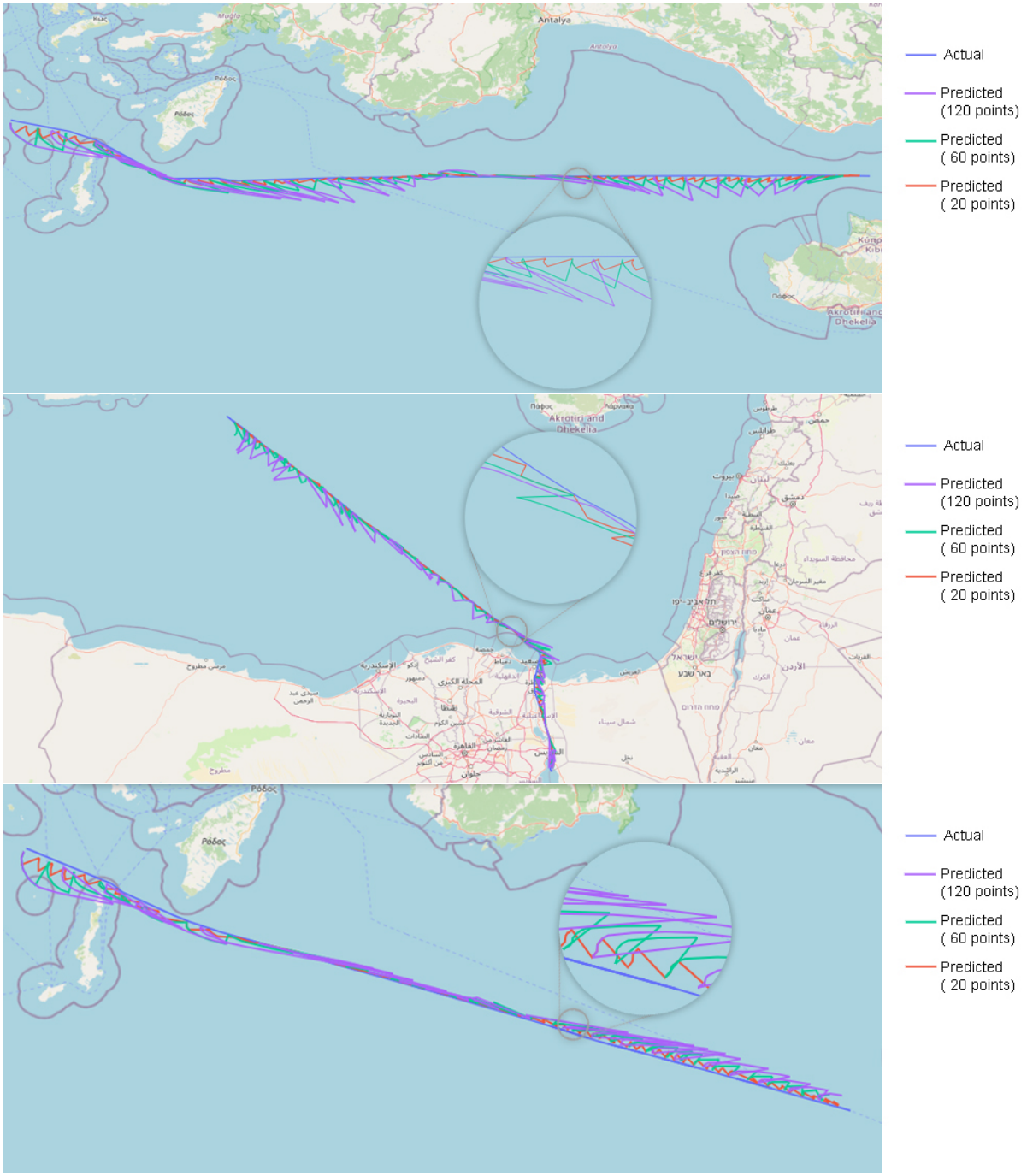


Figure 6.3: Comparison of three actual routes with predicted routes across multiple prediction horizons.

These findings underscore the inherent limitations of autoregressive sequence models in maintaining directional and spatial stability over extended predictions, highlighting the importance of evaluating models across both short- and long-term horizons to fully assess their practical viability in real-world applications. Moreover, they call attention to the necessity of developing strategies aimed at mitigating error propagation to enhance long-range forecasting reliability.

6.2.3 Long Term Horizon to Threshold Analysis

In our final experiment, we evaluate the Bi-LSTM model's long-term predictive performance in an open-loop setting, where the model generates the vessel's future trajectory continuously from a single initial input sequence. Unlike segmented approaches, this setup does not introduce new observational inputs during inference; instead, the model relies entirely on its own recursive predictions to progress along the route. The initial input consists of a 40-point sequence, and the model proceeds without interruption for as many steps as required to reach the end of the actual route, enabling a realistic and unbounded assessment of long-term forecasting behavior.

To further assess the robustness of the Bi-LSTM model over extended prediction horizons, we conducted a time-to-threshold analysis, where we analyzed how quickly prediction errors surpass predefined spatial thresholds during recursive inference. Specifically, we measured the average number of minutes required for the prediction error to surpass a set of predefined spatial thresholds (5, 10, 20, 50, and 100 Km). The results, summarized in Table 6.9, show that on average, it takes around 10 minutes for the error to exceed 5 km, 20 minutes to exceed 10 km, while exceeding 100 km takes nearly 297 minutes (around 5 hours). In addition, we tracked the proportion of routes that ever crossed each threshold. We observed that almost all routes (98) exceed the 5 km threshold eventually, whereas only 75 reach an error exceeding 100 km.

Table 6.9: Time-to-threshold analysis results for long-term vessel trajectory predictions. The table reports the average number of minutes required for the prediction error to exceed each distance threshold, along with the number of routes (out of 100) where the threshold was eventually exceeded.

Distance Threshold (km)	Average Time to Threshold (min)	Number of Routes Exceeding the Threshold
5	10.60	98
10	20.78	96
20	43.72	92
50	112.58	82
100	296.51	75

This trend illustrates the compounding nature of predictive error in long-term autoregressive forecasting. While most routes maintain relatively low spatial error in the early stages of prediction, sustained accuracy becomes increasingly challenging to uphold over time. The decrease in the number of routes exceeding larger thresholds reflects either the natural capping of some route lengths or improved long-term performance on a subset of routes. This analysis provides a quantitative view of trajectory stability, helping identify thresholds of operational acceptability for different maritime applications. Finally, these findings reinforce the necessity of incorporating error mitigation strategies or hybrid correction mechanisms in real-world deployments where prolonged accuracy is critical.

6.3 Discussion

This study has rigorously evaluated the predictive capabilities of three recurrent neural network architectures (i.e., LSTM, Bi-LSTM, and Bi-GRU) for vessel trajectory forecasting, with an emphasis on both

short-term (single-step) and long-term (multi-step) prediction horizons. By systematically varying the input sequence length and employing both ground-truth-driven and autoregressive inference strategies, the analysis offers several key insights for the development of robust data-driven maritime trajectory forecasting models.

Among the tested architectures, the Bi-LSTM consistently delivered the highest overall performance. Particularly when trained with longer input sequences (e.g., 40 points), the Bi-LSTM demonstrated strong accuracy in predicting spatial (latitude, longitude) and dynamic (SoG, CoG) vessel attributes. Its ability to model bidirectional temporal dependencies allows it to extract richer contextual information from sequence data, leading to more coherent and temporally stable predictions. This finding corroborates previous studies underscoring the strength of bidirectional architectures in complex sequential tasks [40, 79, 89].

While the LSTM model achieved generally good results, it did not match the Bi-LSTM in terms of consistency or precision, particularly at longer input lengths. The Bi-GRU model, despite its computational efficiency, exhibited highly unstable behavior when trained on extended sequences, with significant degradation in prediction quality. This may be attributed to its simpler gating mechanism, which, while faster to train, is potentially more vulnerable to gradient instability or overfitting in long-horizon contexts.

The shift to a multi-step long-term prediction framework, where vessel positions are predicted recursively for 10, 30, and 60 minutes into the future, provides a more realistic and operationally relevant assessment of model performance. The results indicate a clear pattern of error accumulation: while predictions at 10 minutes remain spatially close to the ground truth, performance deteriorates at 30 and 60 minutes, particularly in directional metrics such as CoG. This is a known limitation of autoregressive models, where small prediction errors compound over successive recursive inputs [34].

This trend is clearly visualized in Figure 6.3, which depicts model predictions at all three horizons across representative vessel trajectories. The 10-minute predictions align closely with actual routes, while the 30- and 60-minute forecasts progressively diverge; though still maintaining general route structure and directional intent. Such visualizations reinforce the model's practical potential: despite increased uncertainty, Bi-LSTM preserves usable trajectory information over extended intervals without further AIS input.

Our final experiment assessed the Bi-LSTM model's ability to perform long-term trajectory prediction from a single 40-point input, without additional observational updates. While the model demonstrated strong short-term to mid-term accuracy, its performance declined steadily over time due to cumulative error propagation. Deviations became more pronounced beyond 30 minutes, particularly in directional and spatial metrics. Threshold-based analysis further revealed that although many routes remained accurate within short ranges, fewer sustained performance as the horizon extended. These results highlight the model's effectiveness in short-range forecasting and its limitations over extended predictions, underscoring the need for strategies to mitigate long-term drift.

These findings underscore both the promise and limitations of current recurrent approaches in long-range trajectory forecasting. To address these challenges, future work should focus on improving the model's sensitivity to context-dependent maneuvers, potentially through the integration of auxiliary inputs such as heading change rates, sea state conditions, or proximity to navigational turning points. Additionally,

hybrid models that blend data-driven learning with rule-based maneuver detection may offer better robustness in handling abrupt course changes.

To further extend long-term predictions, future work should explore model architectures explicitly designed to mitigate recursive error propagation. Approaches such as scheduled sampling, noise-aware training, or sequence-level loss functions could offer improvements. In particular, attention-based architectures like the Transformer [93] show promise in this domain. Unlike recurrent models, Transformers can attend selectively to relevant temporal contexts without relying on iterative state updates, offering a pathway to more stable and accurate long-term predictions.

6.4 Conclusions

This work contributes to the understanding of recurrent architectures for vessel trajectory prediction by providing empirical evidence and practical insights for model design and deployment. Bi-LSTM models outperform LSTM and Bi-GRU, particularly with longer input sequences, highlighting the strength of bidirectional architectures in capturing complex spatiotemporal dependencies. While longer input sequences improve predictive accuracy, they also increase model complexity and training challenges, especially for Bi-GRU, underscoring the importance of careful model selection and tuning. Our findings also show that short-term prediction evaluations can overestimate real-world performance, while long-term recursive forecasting can lead to error accumulation and reduced accuracy over time. Despite these challenges, Bi-LSTM maintains reasonable trajectory coherence over extended horizons, supporting its applicability for operational maritime applications within defined spatial and temporal bounds. Future research should explore architectures like Transformers and training strategies that mitigate recursive errors to enhance long-term forecasting capabilities. The present investigation into Bi-LSTM-based vessel trajectory forecasting builds upon the methodological foundations established in the earlier chapters on AIS data management and ETA prediction. The AIS framework, incorporating fuzzy matching techniques, provided a reliable basis for high-quality, temporally aligned vessel movement datasets, an essential prerequisite for both arrival time estimation and spatial trajectory modelling. The ETA prediction study demonstrated the operational benefits of machine learning in enhancing port scheduling efficiency and reducing uncertainty in maritime logistics. Extending this focus, the current work shifts from temporal arrival estimates to the modelling of complete spatiotemporal vessel trajectories, enabling long-term forecasting of vessel positions.

7 Summary

This PhD dissertation presents a layered and empirically grounded investigation into intelligent maritime systems, unified by the strategic use of AIS data to address core operational challenges in vessel monitoring, data quality enhancement, ETA prediction, and trajectory forecasting. The research unfolds across four interrelated chapters, advancing from infrastructure-level data handling to high-level predictive analytics, while offering both methodological innovations and practical implications throughout the maritime information pipeline.

The work begins with the development of a scalable and distributed architecture for real-time AIS data processing (Chapter 3). This system addresses the technical complexities inherent in handling high-frequency AIS messages, including issues such as message duplication, data loss, and temporal inconsistencies. Its operational deployment in Cyprus, processing over one billion AIS messages, validates its practical viability and establishes a high-quality data backbone for the remaining studies. However, while the system ensures reliable data ingestion and preprocessing, it does not directly support predictive tasks. Moreover, its performance may vary in regions with limited AIS infrastructure or in contexts with significantly different transmission characteristics.

Building on this foundation, Chapter 4 addresses the semantic inconsistencies within one of the most problematic AIS fields—the manually entered destination field. A domain-specific fuzzy matching framework is proposed, utilizing structured transformation rules and a similarity function tailored to maritime port naming conventions. This method successfully normalized a large set of free-text entries, achieving a high match rate on real-world data from the Eastern Mediterranean. An adaptive lookup mechanism further enhances the system's scalability by storing validated matches for future reuse. Despite these strengths, the approach is limited by its dependency on a precompiled reference table, and its performance may degrade in contexts where port aliases, abbreviations, or spelling variations deviate significantly from known patterns. Additionally, the algorithm's reliance on rule-based transformations makes it susceptible to edge cases not covered by the predefined logic.

With a robust and cleaned AIS dataset in place, Chapter 5 focuses on the development of machine learning models for vessel ETA prediction. This component of the research explores six different algorithms, demonstrating that ensemble-based models (particularly Random Forest and XGBoost) consistently outperform others across a range of performance metrics. The study underscores the operational value of accurate ETA predictions for port logistics, traffic management, and emissions reduction. A route-based data partitioning strategy is adopted to enhance model generalization and avoid overfitting to specific vessel trajectories. Nevertheless, certain limitations persist. The models exhibit reduced performance when deployed in maritime regions with previously unseen routes, pointing to the need for adaptive retraining strategies. Additionally, the study does not yet integrate external factors such as weather, port congestion, or human behavioral patterns, which could further enhance predictive accuracy.

Finally, Chapter 6 investigates vessel trajectory prediction using deep learning architectures. A comparative analysis of LSTM, Bi-LSTM, and Bi-GRU models is conducted to assess their performance in short- and long-term trajectory forecasting tasks. The results demonstrate that Bi-LSTM models provide the most robust performance due to their ability to capture bidirectional temporal dependencies. The find-

ings also emphasize the importance of sequence length in improving spatial accuracy. However, the study reveals a critical limitation in long-term forecasting: the recursive nature of RNN-based models leads to error accumulation, resulting in spatial drift and directional inaccuracies. These limitations suggest the need for exploring Transformer-based architectures or hybrid models that can integrate spatial features, external variables, and attention mechanisms to enhance long-horizon predictions.

Taken together, these four studies form an integrated research trajectory, moving from scalable data collection and semantic cleaning to predictive modeling and operational analytics. Each chapter addresses a distinct challenge while building upon the outputs of the previous one, showcasing the importance of high-quality, structured AIS data in enabling reliable and intelligent maritime decision-making. Despite the achievements, several limitations are acknowledged across the studies, including algorithmic generalization, edge case handling, and model robustness under real-world conditions. Future research should consider adaptive systems that combine rule-based and data-driven approaches, incorporate external contextual factors, and leverage emerging deep learning architectures to advance the state of maritime analytics.

8 Conclusions

This PhD Dissertation significantly contributes to maritime informatics by presenting a unified research program centered on AIS data analytics for operational decision-making in port and vessel management. The four studies presented offer complementary innovations across the data lifecycle:

- **An Intelligent Framework for Vessel Traffic Monitoring Using AIS Data [57]** developed a scalable real-time processing framework for AIS data that successfully manages high data velocity and message inconsistencies. Its deployment in Cyprus highlights its readiness for real-world application and its value in supporting downstream maritime analytics.
- **Employing Fuzzy Matching for Cleaning Manual AIS Entries [94]** addressed the challenge of cleaning the unstructured and error-prone destination field in AIS messages. The proposed fuzzy matching algorithm, combining nine domain-specific rules with adaptive learning, achieved a 91.6% matching rate on over 2.7 million records. It demonstrates how integrating expert knowledge and approximate string matching can transform noisy AIS fields into structured information, thus enhancing the reliability of maritime datasets for predictive modeling.
- **Enhancing Prediction Accuracy of Vessel Arrival Times Using Machine Learning [95]** introduced a robust methodology for ETA prediction using classical and ensemble machine learning algorithms. Route-based data splitting and feature optimization led to high prediction accuracy, particularly for Random Forest and XGBoost models. The study also emphasized the environmental and economic benefits of accurate ETA forecasting, especially in reducing port idle times and emissions.
- **Vessel Trajectory Prediction with Deep Learning: Temporal Modeling and Operational Implications (under submission)** explored deep learning approaches to vessel trajectory forecasting using LSTM variants. It confirmed the superiority of Bi-LSTM in capturing spatiotemporal patterns, but also identified limitations in long-horizon recursive forecasting, pointing toward future directions involving attention-based or hybrid architectures.

Across these studies, the thesis establishes a vertically integrated framework for maritime intelligence:

1. Real-time, scalable data acquisition and processing;
2. Adaptive and rule-based cleaning of AIS destination fields;
3. Operationally impactful ETA prediction with classical ML;
4. Temporally aware deep learning for vessel path forecasting.

Building upon the foundations established in this dissertation, several promising directions for future research can further enhance the accuracy, adaptability, and operational relevance of intelligent maritime systems.

One critical avenue involves the integration of external variables—such as meteorological data (e.g., wind, wave height, sea currents), geopolitical factors (e.g., conflict zones, piracy risk), and economic signals (e.g., port congestion levels, global trade indices)—into predictive frameworks. The current models primarily rely on AIS-derived features, but incorporating such contextual information can sig-

nificantly improve the robustness and generalization of ETA and trajectory predictions. For instance, vessel behavior may change drastically under adverse weather conditions or regional disruptions, which are not captured in AIS data alone. Future research should explore multi-source data fusion techniques that dynamically integrate these exogenous variables into learning architectures.

Another area for advancement is enhancing AIS data cleaning and anomaly detection capabilities within preprocessing pipelines. While the current work addresses semantic inconsistencies in the destination field, other AIS fields—such as speed, course, and position—can be prone to spoofing, GPS drift, or human errors. Future efforts could employ machine learning or probabilistic models to detect and correct such anomalies, improving the reliability of downstream predictive tasks. Furthermore, combining rule-based logic with probabilistic graphical models or unsupervised learning approaches (e.g., isolation forests, autoencoders) could enable scalable and automated anomaly detection across large streaming datasets.

The deployment of online and adaptive learning systems represents another critical direction. Static models trained on historical data may degrade performance over time due to concept drift or changes in maritime traffic patterns. To address this, future work should explore frameworks that support real-time model updates through incremental or online learning. Techniques such as reinforcement learning, federated learning, or streaming gradient descent could be employed to continuously adapt predictive models to evolving maritime conditions without requiring full retraining on the entire dataset. This is particularly relevant for high-frequency AIS data environments where patterns shift dynamically.

To overcome the limitations associated with recursive sequence prediction in RNN-based models, future research should also investigate Transformer-based and hybrid deep learning architectures. Transformers have shown remarkable success in handling long-range dependencies in natural language processing and time-series forecasting. Applying them to vessel trajectory prediction may mitigate issues like error accumulation and directional drift observed in autoregressive models. Additionally, hybrid architectures that combine Convolutional Neural Networks (CNN) for spatial encoding, attention mechanisms for temporal weighting, and external variable embeddings could offer more flexible and interpretable forecasting pipelines.

Finally, the application of Large Language Models (LLMs) opens up novel and underexplored possibilities in the maritime domain. LLMs could be leveraged to interpret unstructured AIS-related metadata (e.g., voyage plans, captain's notes, port announcements), automatically annotate semantic patterns in AIS messages, or assist in generating synthetic data for model training. Moreover, they can be used to translate and normalize diverse linguistic representations of destination names or vessel logs across languages and dialects. Beyond data processing, LLMs may play a role in enhancing maritime decision support systems, offering natural language interfaces for querying predictions, summarizing vessel behavior, or generating risk alerts in human-readable form.

Together, these future directions point toward a more integrated, context-aware, and intelligent maritime analytics ecosystem. By extending the methodological innovations of this dissertation to incorporate external knowledge, adaptive systems, and advanced model architectures, future research can contribute to safer, more efficient, and environmentally sustainable maritime operations.

In conclusion, this PhD Dissertation delivers both foundational insights and practical advancements in maritime machine learning and data-driven systems. By integrating scalable infrastructure, intelligent data cleaning, and advanced predictive models, it lays a comprehensive roadmap for the development of intelligent, sustainable, and adaptive decision-support frameworks within the maritime domain.

BIBLIOGRAPHY

[1] D. H. Bailey, Shiphandling for the Mariner, 4th ed. Cornell Maritime Press, 2000.

[2] "Review of Maritime Transport 2023, UNCTAD," Septemebr 2023, https://unctad.org/publication/review-maritime-transport-2023.

[3] S. Aslam, M. P. Michaelides, and H. Herodotou, "A survey on computational intelligence approaches for intelligent marine terminal operations," IET Intelligent Transport Systems, 2024.

[4] ——, "Internet of ships: A survey on architectures, emerging applications, and challenges," IEEE Internet of Things journal, vol. 7, no. 10, pp. 9714–9727, 2020.

[5] X. Tian, R. Yan, S. Wang, and G. Laporte, "Prescriptive analytics for a maritime routing problem," Ocean & Coastal Management, vol. 242, p. 106695, 2023.

[6] B. Lin, M. N. B. Zabit, and S. Huang, "Development of chinese danmaku video sites based on self-directed learning (cdsdl) model for undergraduate students in tam model," International Journal of Computing and Information Technology, vol. 3, no. 1, pp. 1–1, 2023.

[7] M. Lind, M. Michaelides, R. Ward, and R. T. Watson, Maritime informatics. Springer, 2021.

[8] X. Fu, Z. Xiao, H. Xu, V. Jayaraman, N. B. Othman, C. P. Chua, and M. Lind, "Ais data analytics for intelligent maritime surveillance systems," in Maritime Informatics. Springer, 2020, pp. 393–411.

[9] M. Vodas, K. Bereta, D. Kladis, D. Zissis, E. Alevizos, E. Ntoulias, A. Artikis, A. Deligiannakis, A. Kontaxakis, N. Giatrakos et al., "Online distributed maritime event detection & forecasting over big vessel tracking data," in 2021 IEEE International Conference on Big Data (Big Data). IEEE, 2021, pp. 2052–2057.

[10] G. A. Vouros, A. Vlachou, G. M. Santipantakis, C. Doulkeridis, N. Pelekis, H. V. Georgiou, Y. Theodoridis, K. Patroumpas, E. Alevizos, A. Artikis et al., "Big data analytics for time critical mobility forecasting: Recent progress and research challenges." in EDBT, 2018, pp. 612–623.

[11] G. M. Santipantakis, A. Vlachou, C. Doulkeridis, A. Artikis, I. Kontopoulos, and G. A. Vouros, "A stream reasoning system for maritime monitoring," in 25th International Symposium on Temporal Representation and Reasoning (TIME 2018). Schloss Dagstuhl-Leibniz-Zentrum fuer Informatik, 2018.

[12] P. Bernabé, A. Gotlieb, B. Legeard, D. Marijan, F. O. Sem-Jacobsen, and H. Spieker, "Detecting intentional ais shutdown in open sea maritime surveillance using self-supervised deep learning," IEEE Transactions on Intelligent Transportation Systems, vol. 25, no. 2, pp. 1166–1177, 2023.

[13] C. Maganaris, E. Protopapadakis, and N. Doulamis, "Outlier detection in maritime environments using ais data and deep recurrent architectures," in Proceedings of the 17th International Conference on PErvasive Technologies Related to Assistive Environments, 2024, pp. 420–427.

[14] M. Liang, L. Weng, R. Gao, Y. Li, and L. Du, "Unsupervised maritime anomaly detection for intelligent situational awareness using ais data," Knowledge-Based Systems, vol. 284, p. 111313, 2024.

[15] G. Pallotta, M. Vespe, and K. Bryan, "Vessel pattern knowledge discovery from ais data: A framework for anomaly detection and route prediction," Entropy, vol. 15, no. 6, pp. 2218–2245, 2013.

[16] P. Sheng and J. Yin, "Extracting shipping route patterns by trajectory clustering model based on automatic identification system data," Sustainability, vol. 10, no. 7, p. 2327, 2018.

[17] R. P. Jain, E. F. Brekke, and A. Rasheed, "Unsupervised clustering of marine vessel trajectories in historical ais database," in 2022 25th International Conference on Information Fusion (FUSION). IEEE, 2022, pp. 1–6.

[18] P. Silveira, A. Teixeira, and C. G. Soares, "Use of ais data to characterise marine traffic patterns and ship collision risk off the coast of portugal," The Journal of Navigation, vol. 66, no. 6, pp. 879–898, 2013.

[19] S. Hornauer and A. Hahn, "Towards marine collision avoidance based on automatic route exchange," IFAC Proceedings Volumes, vol. 46, no. 33, pp. 103–107, 2013.

[20] J.-F. Balmat, F. Lafont, R. Maifret, and N. Pessel, "A decision-making system to maritime risk assessment," Ocean Engineering, vol. 38, no. 1, pp. 171–176, 2011.

[21] H. Zheng-wei, Y. Fan, and L. Li-rong, "Ship safe navigation depth reference map based on ais data," Journal of Traffic and Transportation Engineering, vol. 18, no. 4, pp. 171–181, 2018.

[22] N. B. Abdallah, C. Iphar, G. Arcieri, and A.-L. Jousselme, "Fixing errors in the ais destination field," in Oceans 2019-Marseille. IEEE, 2019, pp. 1–5.

[23] M. Tichavska, F. Cabrera, B. Tovar, and V. Araña, "Use of the automatic identification system in academic research," in International Conference on Computer Aided Systems Theory. Springer, Cham, 2015, pp. 33–40. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-319-22120-8_4

[24] W. M. Wijaya and Y. Nakamura, "Predicting ship behavior navigating through heavily trafficked fairways by analyzing ais data on apache hbase," in 2013 First International Symposium on Computing and Networking. IEEE, 2013, pp. 220–226.

[25] K. Park, S. Sim, and H. Bae, "Vessel estimated time of arrival prediction system based on a path-finding algorithm," Maritime Transport Research, vol. 2, p. 100012, 2021.

[26] I. Parolas, "Eta prediction for containerships at the port of rotterdam using machine learning techniques," Port Rotterdam Authority, 2016.

[27] T. Ogura, T. Inoue, and N. Uchihira, "Prediction of arrival time of vessels considering future weather conditions," Applied Sciences, vol. 11, no. 10, p. 4410, 2021.

[28] A. Alessandrini, F. Mazzarella, and M. Vespe, "Estimated time of arrival using historical vessel tracking data," IEEE Transactions on Intelligent Transportation Systems, vol. 20, no. 1, pp. 7–15, 2018.

[29] S. El Mekkaoui, L. Benabbou, and A. Berrado, "Predicting ships estimated time of arrival based on ais data," in Proceedings of the 13th International Conference on Intelligent Systems: Theories and Applications, 2020, pp. 1–6.

[30] J.-H. Yoon, D.-H. Kim, S.-W. Yun, H.-J. Kim, and S. Kim, "Enhancing container vessel arrival time prediction through past voyage route modeling: A case study of busan new port," Journal of Marine Science and Engineering, vol. 11, no. 6, p. 1234, 2023.

[31] H. Arbabkhah, A. Sedaghat, M. Jafari Kang, and M. Hamidi, "Automatic identification system-based prediction of tanker and cargo estimated time of arrival in narrow waterways," Journal of Marine Science and Engineering, vol. 12, no. 2, p. 215, 2024.

[32] M. H. Tun, G. S. Chambers, T. Tan, and T. Ly, "Maritime port intelligence using ais data," in RNSA Security Technology Conference. Research Network for a Secure Australia (RNSA), 2007, pp. 33–43.

[33] B. J. Rhodes, N. A. Bomberger, M. Seibert, and A. M. Waxman, "Maritime situation monitoring and awareness using learning mechanisms," in MILCOM 2005-2005 IEEE Military Communications Conference. IEEE, 2005, pp. 646–652.

[34] H. Li, H. Jiao, and Z. Yang, "Ais data-driven ship trajectory prediction modelling and analysis based on machine learning and deep learning methods," Transportation Research Part E: Logistics and Transportation Review, vol. 175, p. 103152, 2023.

[35] S. Capobianco, L. M. Millefiori, N. Forti, P. Braca, and P. Willett, "Deep learning methods for vessel trajectory prediction based on recurrent neural networks," IEEE Transactions on Aerospace and Electronic Systems, vol. 57, no. 6, pp. 4329–4346, 2021.

[36] E. Chondrodima, N. Pelekis, A. Pikrakis, and Y. Theodoridis, "An efficient lstm neural network-based framework for vessel location forecasting," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 8, pp. 9470–9481, 2021.

[37] S. Capobianco, N. Forti, L. M. Millefiori, P. Braca, and P. Willett, "Recurrent encoder-decoder networks for vessel trajectory prediction with uncertainty estimation," IEEE Transactions on Aerospace and Electronic Systems, vol. 59, no. 3, p. 2554 – 2565, 2023.

[38] X. Wang and Y. Xiao, "A deep learning model for ship trajectory prediction using automatic identification system (ais) data," Information, vol. 14, no. 4, p. 212, 2023.

[39] X. Wu, J. Chen, C. Xion, D. Liu, X. Wan, and Z. Chen, "Vessel trajectory prediction method based on the time series data fusion model," Promet - Traffic and Transportation, vol. 36, no. 6, p. 1160 – 1175, 2024.

[40] C.-H. Yang, C.-H. Wu, J.-C. Shao, Y.-C. Wang, and C.-M. Hsieh, "Ais-based intelligent vessel trajectory prediction using bi-lstm," IEEE Access, vol. 10, pp. 24 302–24 315, 2022.

[41] G.-H. Shin and H. Yang, "Vessel trajectory prediction at inner harbor based on deep learning using ais data," Journal of Marine Science and Engineering, vol. 12, no. 10, 2024.

[42] "Libais AIS decoding library libais description," https://pypi.org/project/libais/, accessed: 2024-09-17.

[43] C. Pani, P. Fadda, G. Fancello, L. Frigau, and F. Mola, "A data mining approach to forecast late arrivals in a transhipment container terminal," Transport, vol. 29, no. 2, pp. 175–184, 2014.

[44] M. Michaelides, H. Herodotou, M. Lind, and R. Watson, "Port-2-port communication enhancing short sea shipping performance: The case study of cyprus and the eastern mediterranean," Sustainability, vol. 11, p. 1, 03 2019.

[45] N. Evmides, S. Aslam, A. Televantos, A. Karagiannis, A. Paraskeva, M. Michaelides, and H. Herodotou, "Employing fuzzy matching for cleaning manual ais entries," in Proc. of World of Shipping Portugal-An International Research Conference on Maritime Affairs, 2021.

[46] S. Chaudhuri, K. Ganjam, V. Ganti, V. Narasayya, and T. Vassilakis, "Fuzzy lookup and fuzzy grouping in sql server integration services," Microsoft Corporation, Tech. Rep., 2005, http://msdn2.microsoft.com/en-us/library/ms345128.aspx.

[47] N. Aydin, H. Lee, and S. A. Mansouri, "Speed optimization and bunkering in liner shipping in the presence of uncertain service times and time windows at ports," European Journal of Operational Research, vol. 259, no. 1, pp. 143–154, 2017.

[48] S. Aslam, M. P. Michaelides, and H. Herodotou, "Berth allocation considering multiple quays: A practical approach using cuckoo search optimization," Journal of Marine Science and Engineering, vol. 11, no. 7, p. 1280, 2023.

[49] R. Meijer, "Eta prediction: Predicting the eta of a container vessel based on route identification using ais data," Ph.D. dissertation, TU Delft Delft, The Netherlands, 2017.

[50] Y. Xu, Q. Chen, and X. Quan, "Robust berth scheduling with uncertain vessel delay and handling time," Annals of Operations Research, vol. 192, pp. 123–140, 2012.

[51] C. Bierwirth and F. Meisel, "A survey of berth allocation and quay crane scheduling problems in container terminals," European Journal of Operational Research, vol. 202, no. 3, pp. 615–627, 2010.

[52] F. Rodrigues and A. Agra, "Berth allocation and quay crane assignment/scheduling problem under uncertainty: A survey," European Journal of Operational Research, vol. 303, no. 2, pp. 501–524, 2022.

[53] M. P. Michaelides, H. Herodotou, M. Lind, and R. T. Watson, "Port-2-port communication enhancing short sea shipping performance: The case study of cyprus and the eastern mediterranean," Sustainability, vol. 11, no. 7, p. 1912, 2019.

[54] C. I. Valero, Á. Martínez, R. Oltra-Badenes, H. Gil, F. Boronat, and C. E. Palau, "Prediction of the estimated time of arrival of container ships on short-sea shipping: A pragmatical analysis," IEEE Latin America Transactions, vol. 20, no. 11, pp. 2354–2362, 2022.

[55] T. F. Schindler, J.-H. Ohlendorf, and K.-D. Thoben, "Towards vessel arrival time prediction through a deep neural network cluster," in International Conference on Dynamics in Logistics. Springer, 2024, pp. 160–170.

[56] P. Wenzel, R. Jovanovic, and F. Schulte, "A neural network approach for eta prediction in inland waterway transport," in International Conference on Computational Logistics. Springer, 2023, pp. 219–232.

[57] N. Evmides, L. Odysseos, M. P. Michaelides, and H. Herodotou, "An intelligent framework for vessel traffic monitoring using ais data," in 2022 23rd IEEE International Conference on Mobile Data Management (MDM). IEEE, 2022, pp. 413–418.

[58] J. Li, K. Cheng, S. Wang, F. Morstatter, R. P. Trevino, J. Tang, and H. Liu, "Feature selection: A data perspective," ACM computing surveys (CSUR), vol. 50, no. 6, pp. 1–45, 2017.

[59] E. Flapper, “Eta prediction for vessels using machine learning,” B.S. thesis, University of Twente, 2020.

[60] L. Kolley, N. Rückert, M. Kastner, C. Jahn, and K. Fischer, “Robust berth scheduling using machine learning for vessel arrival time prediction,” Flexible Services and Manufacturing Journal, vol. 35, no. 1, pp. 29–69, 2023.

[61] J. F. Griffiths and D. M. Driscoll, Survey of climatology. CE Merrill Publishing Company, 1982.

[62] J. Yu, G. Tang, X. Song, X. Yu, Y. Qi, D. Li, and Y. Zhang, “Ship arrival prediction and its value on daily container terminal operation,” Ocean Engineering, vol. 157, pp. 73–86, 2018.

[63] O. Bodunov, F. Schmidt, A. Martin, A. Brito, and C. Fetzer, “Real-time destination and eta prediction for maritime traffic,” in Proceedings of the 12th ACM international conference on distributed and event-based systems, 2018, pp. 198–201.

[64] E. K. Ayesu, “Does shipping cause environmental emissions? evidence from african countries,” Transportation Research Interdisciplinary Perspectives, vol. 21, p. 100873, 2023.

[65] X. Chen, S. Dou, T. Song, H. Wu, Y. Sun, and J. Xian, “Spatial-temporal ship pollution distribution exploitation and harbor environmental impact analysis via large-scale ais data,” Journal of Marine Science and Engineering, vol. 12, no. 6, p. 960, 2024.

[66] P. Serra and G. Fancello, “Towards the imo’s ghg goals: A critical overview of the perspectives and challenges of the main options for decarbonizing international shipping,” Sustainability, vol. 12, no. 8, p. 3220, 2020.

[67] M. Riveiro, G. Pallotta, and M. Vespe, “Maritime anomaly detection: A review,” Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, vol. 8, no. 5, p. e1266, 2018.

[68] F. Farahnakian, F. Farahnakian, J. Sheikh, P. Nevalainen, and J. Heikkonen, “Short and long term vessel movement prediction for maritime traffic,” Lecture Notes in Computer Science, vol. 14599 LNCS, p. 62 – 80, 2024.

[69] Y. Shin, N. Kim, H. Lee, S. Y. In, M. Hansen, and Y. Yoon, “Deep learning framework for vessel trajectory prediction using auxiliary tasks and convolutional networks,” Engineering Applications of Artificial Intelligence, vol. 132, 2024.

[70] R. W. Liu, K. Hu, M. Liang, Y. Li, X. Liu, and D. Yang, “Qsd-lstm: Vessel trajectory prediction using long short-term memory with quaternion ship domain,” Applied Ocean Research, vol. 136, 2023.

[71] M. M. Alam, G. Spadon, M. Etemad, L. Torgo, and E. Milios, “Enhancing short-term vessel trajectory prediction with clustering for heterogeneous and multi-modal movement patterns,” Ocean Engineering, vol. 308, 2024.

[72] B. Xu and L. P. Perera, “A data-driven approach to vessel trajectory prediction for safe autonomous ship operations,” ResearchGate, 2018. [Online]. Available: https://www.researchgate.net/publication/327107528_A_Data-Driven_Approach_to_Vessel_Trajectory_Prediction_for_Safe_Autonomous_Ship_Operations

[73] R. W. Liu, M. Liang, J. Nie, X. Deng, Z. Xiong, J. Kang, H. Yang, and Y. Zhang, "Intelligent data-driven vessel trajectory prediction in marine transportation cyber-physical system," in IEEE Congress on Cybermatics: 2021 IEEE International Conferences on Internet of Things, iThings 2021, IEEE Green Computing and Communications, GreenCom 2021, IEEE Cyber, Physical and Social Computing, CPSCom 2021 and IEEE Smart Data, SmartData 2021, 2021, p. 314 – 321.

[74] T. Liu, X. Xu, Z. Lei, X. Zhang, M. Sha, and F. Wang, "A multi-task deep learning model integrating ship trajectory and collision risk prediction," Ocean Engineering, vol. 287, p. 115870, 2023. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S0029801823022540

[75] Y. Lei, Y. Zhang, and J. Liu, "Research on ship trajectory prediction method based on difference long short-term memory neural network," MDPI Journal of Marine Science and Engineering, 2020. [Online]. Available: https://www.mdpi.com/2077-1312/11/9/1731

[76] X. Zhang, X. Fu, Z. Xiao, H. Xu, and Z. Qin, "Vessel trajectory prediction in maritime transportation: Current approaches and beyond," IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 11, pp. 19 980–19 998, 2022.

[77] D. Zissis, E. K. Xidias, and D. Lekkas, "Real-time vessel behavior prediction," Evolving Systems, vol. 7, pp. 29–40, 2016.

[78] Y. Suo, W. Chen, C. Claramunt, and S. Yang, "A ship trajectory prediction framework based on a recurrent neural network," Sensors, vol. 20, no. 18, p. 5133, 2020.

[79] W. Li, Y. Lian, Y. Liu, and G. Shi, "Ship trajectory prediction model based on improved bi-lstm," ASCE-ASME Journal of Risk and Uncertainty in Engineering Systems, Part A: Civil Engineering, vol. 10, no. 3, p. 04024033, 2024. [Online]. Available: https://ascelibrary.org/doi/10.1061/AJRUA6.RUENG-1234

[80] X. Wang, H. Liu, and Q. Li, "A ship trajectory prediction framework based on a recurrent neural network," PMC, 2020. [Online]. Available: https://pmc.ncbi.nlm.nih.gov/articles/PMC7570964/

[81] J. Chen, Y. Wang, and X. Zhao, "A data-driven approach to vessel trajectory prediction for safe autonomous ship operations," ResearchGate, 2019. [Online]. Available: https://www.researchgate.net/publication/327107528_A_Data-Driven_Approach_to_Vessel_Trajectory_Prediction_for_Safe_Autonomous_Ship_Operations

[82] Y. Li, Q. Yu, and Z. Yang, "Vessel trajectory prediction for enhanced maritime navigation safety: A novel hybrid methodology," Journal of Marine Science and Engineering, vol. 12, no. 8, p. 1351, 2024. [Online]. Available: https://www.mdpi.com/2077-1312/12/8/1351

[83] R. Yan, S. Wang, L. Zhen, and G. Laporte, "Emerging approaches applied to maritime transport research: Past and future," Communications in Transportation Research, vol. 1, p. 100011, 2021. [Online]. Available: https://www.sciencedirect.com/science/article/pii/S2772424721000111

[84] A. Harati-Mokhtari, A. Wall, P. Brooks, and J. Wang, "Ais: Data reliability and human error implications," Journal of Navigation, vol. 60, no. 3, pp. 373–389, 2007.

[85] D. Zissis, D. Lekkas, and M. Papadopoulou, "A machine learning approach to anomaly detection for vessel trajectories," Expert Systems with Applications, vol. 66, pp. 120–140, 2016.

[86] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” Neural computation, vol. 9, no. 8, pp. 1735–1780, 1997.

[87] Y. Bengio, P. Simard, and P. Frasconi, “Learning long-term dependencies with gradient descent is difficult,” IEEE Transactions on Neural Networks, vol. 5, no. 2, pp. 157–166, 1994.

[88] K. Zhang, W. Sun, and P. He, “Maritime traffic prediction using lstm networks: A case study on ais data,” Transportation Research Part C: Emerging Technologies, vol. 150, p. 104276, 2023.

[89] Z. Wang, L. Huang, and J. Yu, “Short-term trajectory prediction of maritime vessels using bi-lstm networks,” IEEE Transactions on Intelligent Transportation Systems, vol. 23, no. 4, pp. 3105–3114, 2022.

[90] N. Evmides, S. Aslam, T. T. Ramez, M. P. Michaelides, and H. Herodotou, “Enhancing prediction accuracy of vessel arrival times using machine learning,” Journal of Marine Science and Engineering, vol. 12, no. 8, 2024. [Online]. Available: https://www.mdpi.com/2077-1312/12/8/1362

[91] A. Troupiotis-Kapeliaris, C. Kastrisios, and D. Zissis, “Vessel trajectory data mining: a review,” IEEE Access, 2025.

[92] T. Eiter and H. Mannila, “Computing discrete fréchet distance,” Technische Universitat Wien, Tech. Rep. CD-TR 94/64, 1994.

[93] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention is all you need,” Advances in neural information processing systems, vol. 30, 2017.

[94] N. Evmides, S. Aslam, A. Televantos, A. Karagiannis, A. Paraskeva, M. Michaelides, and H. Herodotou, “Employing fuzzy matching for cleaning manual ais entries,” in Proceedings of the 2021 World of Shipping Portugal: An International Research Conference on Maritime Affairs. Carcavelos, Portugal: World of Shipping Portugal, Jan. 2021, presented at Hotel Riviera, Carcavelos, 28–29 January 2021.

[95] N. Evmides and Others, “Enhancing prediction accuracy of vessel arrival times using machine learning,” Journal of Marine Science and Engineering, vol. 12, no. 8, p. 1362, 2024. [Online]. Available: https://www.mdpi.com/2077-1312/12/8/1362