Technische Universiteit
Eindhoven
University of Technology

Jheronimus Academy of Data Science
MSc in Data Science for Business and Entrepreneurship

Contextual Maritime Anomaly Detection

Master Thesis

Huub Van de Voort

Supervisors:
Indika PK Weerasingha Dewage
Fedor Baart
Rogier Brussee

's-Hertogenbosch, July 2025

Abstract

The maritime sector is facing increasing operational complexity and risk as shipping activity increases, and is subject to both geopolitical and environmental disruptions. Current deep learning approaches for maritime anomaly detection inadequately address the role of weather-induced behavior, resulting in false alarms that limit their usability in real-world scenarios. This thesis investigates the integration of meteorological data within a multi-model deep learning anomaly detection framework to more effectively distinguish between normal weather-induced vessel movements and artificial trajectory anomalies. Experiments were conducted using the United States West Coast AIS dataset (January–September 2023), augmented with Fifth generation ECMWF Reanalysis (ERA5) meteorological reanalysis data. Evaluation was done with synthetic anomalies, to rigorously evaluate the effects of model complexity and weather variable incorporation. The results show that increasing model complexity enhances model specificity at the potential expense of sensitivity, and that meteorological data integration markedly improves the correct classification of weather-induced vessel behaviors—particularly during severe weather events—while overall discriminative performance remains stable for all types of synthetic anomalies. The improvement in specificity is most pronounced for heading and speeding anomalies, while shift deviation anomalies remain difficult to separate from genuine weather-induced behavior. This indicates that the impact of weather features on anomaly detection varies depending on the extent to which synthetic anomalous behavior represents behavior induced by weather conditions. These findings highlight the need for more robust validation methods, including the use of real-world anomaly cases, to ensure reliable assessment and continued improvement of maritime anomaly detection systems. Furthermore, integrating meteorological data significantly enhances detection accuracy during severe weather events, which can help industry practitioners and policymakers reduce false alarms and improve operational decision-making in challenging maritime environments.

Preface

I am sincerely grateful for the opportunity to complete this thesis project in collaboration with Fedor Baart (Rijkswaterstaat) and Indika PK Weerasingha Dewage (JADS). I greatly respect their immense knowledge, patience, and generous investment of time. I would also like to thank dr. Rogier Brussee (JADS) for his valuable time and interesting discussions throughout this process. My heartfelt thanks go to my father, Maarten, for his care and genuine interest in the progress of this thesis.

I am particularly grateful of being able to conduct research in a field that closely aligns with my personal interests. From a young age, I participated in sailing schools and eventually became an instructor myself. The instructors always taught us to steer clear of large vessels, but I must confess, I did not always listen.

Designing and implementing a multi-modal deep learning architecture proved to be a significant challenge; however, I am proud of the progress made and the results achieved in this project.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Research Context

Global economic trends display significant influence on the maritime shipping industry [23], sustaining moderate yet continuous growth in shipping activities over the past decade—a trend expected to persist in the foreseeable future [42]. This growth has resulted in an increased interdependence between maritime trade and national economic prosperity [16]. Nevertheless, the maritime sector is characterized by substantial complexity [4] and is highly susceptible to external factors. For instance, global disruptions such as the Panama Canal drought, the war in Ukraine, and United States trade restrictions caused redistribution of maritime traffic patterns.

In addition, changing oceanographic and meteorological conditions, including heightened wave activity and stronger winds [4], further amplify risks to maritime safety by increasing the likelihood of infrastructure damage and operational incidents [33]. Under severe weather, vessels may experience loss of control and causing uncontrollable drift, exacerbated by the large size and limited maneuverability of modern cargo ships [43]. Such incidents can lead to severe consequences, including personal injury, damage to maritime infrastructure, financial loss, environmental hazards, and even the loss of life [10]. These risks underscore the necessity for a detailed contextual analysis of the relationship between vessel behavior and meteorological conditions [12, 43].

Recent advances in Maritime Situational Awareness (MSA) systems have enhanced the capability to prevent or mitigate these adverse outcomes [25, 20]. In particular, real-time anomaly detection models hold promise for the early identification of distress situations [18]. However, prevailing state-of-the-art approaches [20, 18, 27] often disregard the impact of weather conditions on vessel movement patterns [24]. This oversight diminishes system usability and reliability, as evidenced by elevated false alarm rates [33] caused by weather-induced behavioral patterns [33, 31]. In response, this thesis explores the integration of meteorological data within a multi-model deep learning framework, with the objective of decreasing false alarm rates. All experiments were conducted using the United States West Coast AIS dataset spanning January to September 2023, with meteorological data sourced from ERA5 reanalysis, including wind components, significant wave height, wave period, and wave direction.

1.2 Research Relevance

1.2.1 Practical Relevance

False alarms impose significant operational challenges to Vessel Traffic Monitoring Staff (VTMS) [33]. This study addresses the problem by evaluating whether temporally aware models can

enhance robustness against false positive alerts triggered by weather-induced, yet operationally normal, vessel movements. Temporally aware systems [18] facilitate for early detection of potential abnormalities and contribute to providing a solution to another significant challenge found in the allocation of Search and Rescue Resources (SAR) [4]. Enhanced early warning capabilities are anticipated to contribute to more effective incident response and optimized resource distribution.

1.2.2 Theoretical Relevance

This work provides empirical evidence for the impact of meteorological data with respect to reducing false alarms in a multi-model deep learning framework. We show successful integration of temporally diverse data sources by incorporation in a temporally and spatially aware architecture. We leverage structural validation methods and test the significance of results through statistical testing. Finally, we provide validated perspectives on the influence of weather severity and synthetically created anomalies.

1.3 Research Questions

In this thesis, the main research question is defined as follows: *How can the integration of meteorological information augment the effectiveness of deep learning approaches in distinguishing weather-induced vessel movements from artificial maritime trajectory anomalies?* In relation to providing an answer to the main research question, the following sub-research questions are defined as follows:

RQ1: How does the number of Gaussian Mixture Model (GMM) components contribute to distinguishing weather-induced behavioral patterns from artificial anomalies?

To address the this research question, this work implements and compares the original model architecture as presented by [18] with an enhanced version that incorporates meteorological statistics within the dynamic pattern clustering component. In this work, we extend the dimensionality of the data from which patterns are learned. Since the capacity of the model to learn vessel behavior patterns strongly depends on the ability of the clustering component to identify groups in the training data, we conduct experiments to determine the optimal model configuration for our research objective.

RQ2: To what extent does incorporating meteorological variables influence the ability to distinguish between weather-induced behavioral patterns from artificial maritime trajectory anomalies?

To answer the second research question, we compare the baseline implementation with the weather-enhanced model. Specifically, we compare our baseline implementation with our weather enhanced version and statistically validate if vessel behavior caused by occurring meteorological conditions is contextualized better.

RQ3: How does weather integration affect the model's ability to correctly identify different types of artificial anomalies while avoiding false positives from weather-induced behavioral patterns?

This thesis analyzes how our model distinguishes weather-induced vessel behavior from three types of artificial maritime anomalies [18] - shift deviation, abnormal heading, and abnormal speeding — by injecting synthetic anomalies with varying parameters of spatial deviation (d), anomaly ratio (r), and proportion of affected waypoints (ρ).

RQ4: How does the integration of meteorological data affect model performance in distinguishing weather-induced behavioral patterns from artificial maritime trajectory anomalies across varying weather severity conditions?

To answer the fourth research question, this work categorizes trajectories by weather severity using aggregated meteorological statistics and evaluates model performance across different weather conditions.

1.4 Thesis Structure

The remainder of this thesis is organized as follows: Chapter 2 reviews provides a perspective on the problem domain and relevant literature; Chapter 3 details the methodology and dataset; Chapter 4 presents experimental results; and Chapter 5 discusses the findings and implications. Chapter 6 concludes and outlines directions for future research.

Chapter 2

Background And Related Work

2.1 Background

This chapter provides a synthesis of the key literature relevant to maritime anomaly detection, with an emphasis on foundational concepts and recent advancements in the field. The subsequent sections are organized to first introduce essential background knowledge, followed by a discussion of technical methodologies relevant to the detection of anomalies in maritime domains. Finally, we conclude chapter by elaborating the literature gap foundational to this research.

2.1.1 Maritime Shipping

Maritime transportation serves as a cornerstone of the global economy, facilitating the majority of international trade and underpinning the processes of globalization. Estimates indicate that between 75% and 90% of global trade, both in terms of volume and financial value, is transported by sea [16, 42]. Consequently, maritime shipping is integral to the functioning of globalized trade, manufacturing supply chains, and the economic stability of nations [16].

The correlation between national economic prosperity and access to maritime trade has been well-established, underscoring the importance of efficient port infrastructure and sustained shipping connectivity [16]. These factors highlight the necessity for robust maritime safety and monitoring systems to protect critical economic interests [20]. Inadequate safety measures render national economies vulnerable to significant risk, as disruptions in maritime shipping can have far-reaching consequences.

Marine casualties present substantial threats, including the loss of human life, environmental degradation, and considerable financial repercussions [4]. For instance, from 2014 to 2023, a total of 26,595 marine incidents were reported, resulting in 650 fatalities [42]. To address these challenges, both the International Maritime Organization (IMO) and national Maritime Safety Administrations (MSAs) have issued comprehensive safety regulations and operational guidelines aimed at minimizing the occurrence and impact of maritime accidents [23].

2.1.2 AIS Data

To improve nautical safety, the IMO (International Maritime Organization) established both traffic regulations and the obligation for professional ships to be equipped with AIS (Automatic Identification System), which significantly improved Maritime Situational Awareness (MSA) [25].

AIS systems are equipped with a GPS and a transponder, allowing real-time monitoring and historic analysis of vessel movements [33]. AIS data is transposed non-encrypted via VHF (Very High Frequency) radio signals, making it publicly accessible. AIS systems typically send out

dynamic information such as position, speed and heading with a temporal frequency of approximately 3-10 seconds. In addition, static information is sent out at a lower frequency and involves the dimensions of the ship, the port of departure, and the port of arrival [33].

Inherent to AIS data is questionable reliability due to the limitations of the technology of onboard AIS systems and their dependency on human efforts. The stability of VHF networks degrades when weather conditions such as fog, snow, or rain occur [19]. The material, length, and height of the antenna of the transponder also influence the distance VHF signals can travel. Besides technological limitations, the navigational status, origin, and destination attributes within AIS signals require human input, and this is often neglected [33]. However, a solution to this challenge is to deduce these attributes from other attributes [19].

Due to the reliability challenges of AIS data, publications within the maritime anomaly detection domain provide considerable efforts discussing potential solutions which can involve: interpolation techniques [24], cleaning methods [19], data enhancement [36, 1].

Due to the high frequency at which AIS data points are recorded [24] and the slow movements of vessels, the volume of AIS data becomes large [33]. This brings additional challenges for utilizing it in other applications. Within anomaly detection approaches, techniques to overcome this challenge are compression or sampling [24]. To illustrate, compressing is adopted in work by [36], describing trajectories as statistical aggregations over position, speed, and course attributes. On the other hand, sampling using interpolation techniques is adopted in [54, 19, 25, 18], reducing the volume of the data but also ensuring temporal consistency.

2.1.3 Meteorological Conditions in Relation to Maritime Navigation

The ocean is a large water mass which is influenced by external conditions such as wind, current, and wave height [4, 24, 12]. The hydrodynamic properties of water allow for transportation using the ocean as a means of transportation. Due to the relation between sea state and external weather conditions [12], navigational decisions are strongly influenced by these external conditions [55, 43].

To be more precise, external meteorological conditions significantly influence vessel behavior through various force vectors acting on the vessel. As illustrated in Figure 2.1, wind and current forces create complex interactions with a vessel's actual course. The wind force (θ_w), (shown by its direction in Figure: 2.1), acts primarily on the vessel's superstructure, while ocean currents, with their own directional component (θ_c), affect the submerged hull [55]. These external forces result in deviations from the intended course, manifested as leeway and drift angle (γ), which represents the angular deviation between the ship's heading (ψ) and its actual path over ground (course over ground, Ψ). The distinction between speed over ground (v_{SOG}) and actual vessel velocity components (longitudinal speed u and lateral speed v) is crucial for navigation [55], as these parameters directly influence the vessel's trajectory and potential corrective actions manifesting in navigational decisions.

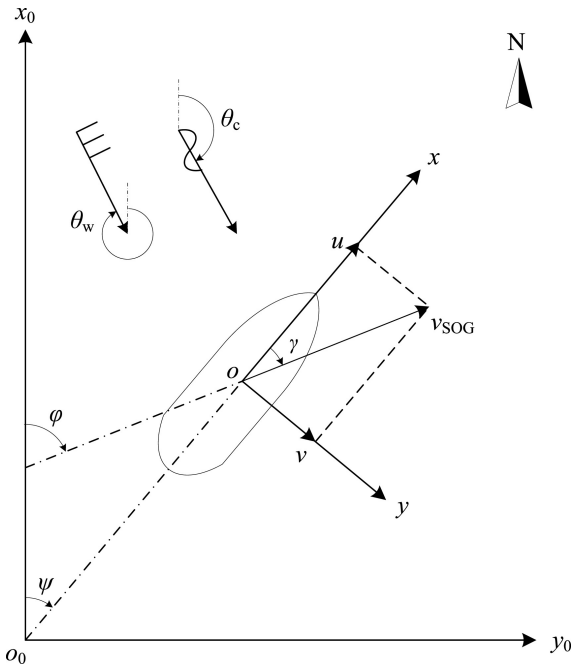

Figure 2.1: The influence (γ) of wind (θ_w) and ocean currents (θ_c) on actual course over ground (ϕ). Adapted from [55].

The relationship between the risk of maritime accidents and external environmental conditions such as weather and ocean state is well established in literature by [4, 1, 12]. These studies have systematically evaluated the predictive value of meteorological variables in the context of maritime incidents, confirming that weather-influenced forces—such as waves, wind, and ocean currents—have a substantial impact on vessel behavior and controllability [55, 24]. This is reflected in navigational decision-making [43, 7], which is strongly influenced by prevailing weather conditions [33]. To illustrate, captains may reduce vessel speed or alter the vessel's course to mitigate risks during rough weather [43, 46].

Recent work has demonstrated the combined use of meteorological and Automated Identification System (AIS) data to estimate collision risks with improved accuracy [4, 1, 12]. Among the various features related to meteorology are wind speed, sea level pressure, visibility, cloud cover, and moon phase identified as the most significant predictors for different types of maritime accidents [4, 1, 12]. Wind speed is particularly noteworthy for its role in wave generation, which, in turn, results in a substantial effect on the probability of collisions and contact accidents. This predictive value is most marked at wind speeds exceeding 35 knots, a threshold at which wave height increases considerably, especially in open sea conditions [4, 1, 12]. The heightened risk associated with rough seas is thus closely tied to elevated wind speeds.

Sea level pressure represents another critical feature in predictive modeling. Sudden sea level pressure drops are indicative of deteriorating weather conditions. The temporal resolution at which sea level pressure is measured is crucial, as higher temporal granularity enables more precise detection of rapid changes associated with hazardous weather [4]. The significance of temporal resolution is further underscored in the integration of weather data with AIS data, as discussed in [24]. When merging meteorological and AIS datasets, contrasting recording frequencies of these data sources introduce challenges that require careful temporal alignment and manipulation.

Cloud cover has also been found to be a significant predictor, particularly due to the dynamics of cloud formation and movement. Since land areas heat more rapidly than ocean surfaces, clouds are predominantly generated over land and subsequently transported toward coastal re-

gions. Given that much maritime infrastructure is concentrated near the shore, increased cloud cover in these areas correlates with a greater risk of collisions [4]. Additionally, the phase of the moon exerts a discernible influence on maritime accident risk. When the sun and moon are aligned during new moon or full moon phases, the resulting gravitational forces produce spring tides, characterized by larger tidal variations than those observed at other times. These heightened tidal conditions have been statistically associated with increased accident risk [4].

Taken together, these findings underscore the impact of meteorological and astronomical factors on maritime safety and highlight the importance of their inclusion in predictive risk models. Incorporating such variables not only enhances the accuracy of accident risk forecasts but also provides critical contextual information that can inform real-time operational decisions, as recommended by [36], the introduction of heterogeneous sources of information, such as meteorologic observations, might result in achieving more accurate detection of abnormal ship behavior and optimize model predictions.

2.1.4 Anomaly Detection

In this section, general definitions of anomalies and the task of detecting them are discussed and later put into perspective in maritime shipping. In [5], anomalies and the task of detecting them are defined by the following:

> *"Anomalies are patterns in data that do not conform to a well-defined notion of normal behavior. ... Anomaly detection refers to the problem of finding patterns in data that do not conform to expected behavior. These nonconforming patterns are often referred to as anomalies, outliers, discordant observations, exceptions, aberrations, surprises, peculiarities, or contaminants in different application domains."*

The value of anomalies is found in the fact that they contain critical and useful information within various application domains [5]. The importance of anomaly detection in maritime shipping stems from its vital role in ensuring safety, security, and efficiency in global maritime operations [33].

The detection of maritime anomalies is becoming more important for shipping due to its potential to improve Maritime Situational Awareness (MSA) [33, 20]. MSA is defined as the comprehensive understanding of what is happening in the maritime environment and the implications of that information for current and future operations [25]. In the maritime context, MSA is essential for safe navigation and primarily focuses on obstacle detection, collision risk assessment, and the prediction of close-range encounter situations with other vessels, which are considered the primary obstacles in maritime operations [25]. Due to the continued increase in the density of the traffic of vessels, the maritime environment becomes more complex, creating a challenge for human operators to manually monitor and identify potential abnormal situations[20]. MSA systems can automatically identify deviations from normal maritime patterns. This automated capability is essential because it provides early warning systems that can detect potential collision risks, suspicious activities, or safety hazards before they escalate into close-range encounter situations, thereby supporting proactive decision-making rather than reactive responses to emergencies [25].

Within the literature, the most frequently detected maritime anomalies are positional anomalies, contextual anomalies, kinematic anomalies, complex anomalies, and data-related anomalies [35]. Positional anomalies occur when a vessel appears in an unusual location. For example, a voyage passing through an anchorage area. Contextual anomalies are similar to positional anomalies but involve contextual information such as the type of vessel and the geographic environment. For example, a tanker following a ferry route. Kinematic anomalies are defined by the movements of a ship. Anomalous movements can be U-turns or moving sideways. Complex anomalies that require an ensemble of detectors to capture specific behaviors. Examples of this type are drug smuggling

or deliberately disabling the onboard AIS system. Lastly, data-related anomalies refer to detecting noise or inconsistencies in AIS data. For example, an incorrect registration of the vessel's position.

The European Maritime Safety Agency (EMSA) serves as the European Union's primary maritime authority, dedicated to promoting safe, secure, green, and competitive maritime operations across EU waters and beyond. The EMSA acknowledges the importance of MSA [8] and provides a framework for categorizing anomalous behaviors for Automated Behavior Monitoring (ABM) algorithms.

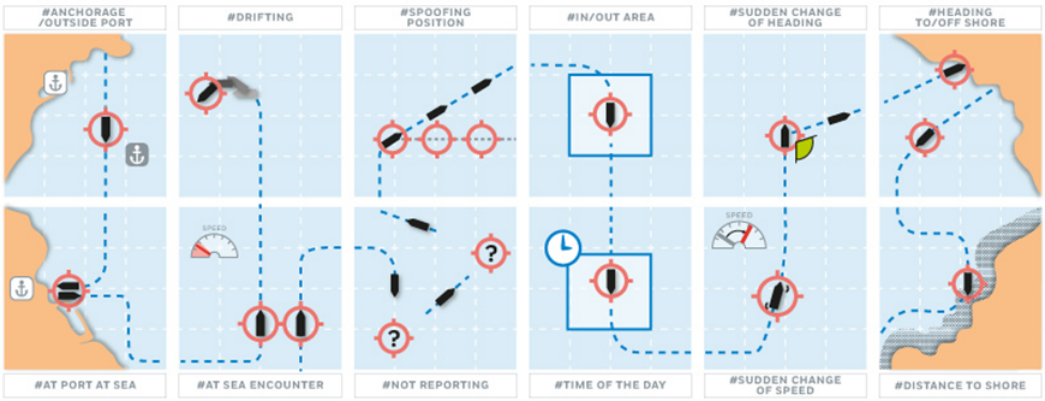

Figure 2.2: Anomalous Behaviours as defined by EMSA, adapted from [8].

2.2 Related Work

Within the literature, various approaches exist for detecting anomalous patterns in vessel behaviors. Based on the approaches sub-taxonomy as given in [33], their main categories are: *Data-driven*, *Knowledge-driven*, or *Hybrid*. Data-driven approaches for maritime anomaly detection leverage machine learning algorithms to identify patterns from historical vessel movement data, then detect deviations from these learned normal behaviors [33]. Knowledge-driven approaches are based on expert domain knowledge encoded through rules, ontologies, or predefined patterns [33]. These methods match new vessel data against predefined rules to detect specific anomalies of interest. Hybrid approaches combine data-driven and knowledge-driven methods to leverage the strengths of both paradigms [33].

2.2.1 Learning Based Anomaly Detection in the Maritime Domain

The scope of this work limits itself to Data-Driven Machine Learning approaches. These approaches are further categorized into *Statistical* and *Machine Learning* approaches. Statistical approaches build statistical models that represent normal vessel behavior using historical data, and then apply statistical inference tests to evaluate new trajectories. Machine Learning approaches train models to represent normal vessel behavior patterns and match new data against these learned models [27]. As mentioned by [33], Deep Learning techniques have shown particular promise in this domain. In the following sections, works that involve Data-Driven Machine Learning Based approaches are discussed in more detail.

Machine Learning Based

Clustering methods are semi-supervised or unsupervised approaches to detect anomalies [5] and can be used to detect anomalous trajectories in a larger collection of trajectories by labeling tra-

jectories belonging to sparse clusters as anomalous and trajectories belonging to dense clusters as normal [36, 33]. As noted in [27], Density-Based Spatial Clustering of Applications with Noise (DBSCAN) is a commonly used clustering algorithm for maritime anomaly detection, due to its applicability for datasets with noise [9].

In [17] DBSCAN was used for spatio-temporal vessel trajectory clustering. The method transforms trajectory clustering into point clustering by using Merge Distance (MD) to measure trajectory similarities and Multidimensional Scaling (MDS) to map these similarities to low-dimensional spatial points. Afterwards, DBSCAN was used to cluster the mapped points. This approach successfully identifies traffic flow patterns, customary routes, and anomalous trajectories, including vessels with opposite courses, while achieving lower computational complexity compared to traditional trajectory clustering methods. In [36], a comparable method was proposed in which DBSCAN was used to group precomputed motion parameters of trajectories.

In [3] DBSCAN was improved ship trajectory clustering. The authors addressed DBSCAN's traditional parameter selection challenges by implementing an adaptive threshold selection method that combines similarity distribution analysis with an improved K-Adaptive Nearest Neighbors (KANN) approach to automatically determine optimal clustering parameters. They reduced computational complexity by pre-calculating all trajectory similarities using Fast-DTW and storing results in a matrix to avoid repeated calculations during clustering. This improved DBSCAN successfully identified 11 distinct trajectory clusters from Baltic Sea AIS data.

Clustering-based approaches face several fundamental limitations in maritime anomaly detection. A primary challenge is the underlying assumption that normal data forms dense clusters while anomalies remain isolated or in sparse clusters [5]; an assumption that fails when anomalous vessels exhibit similar behaviors and form tight clusters themselves[17]. To illustrate, the normal shipping route might be blocked by an obstacle for a longer time, causing ships to deviate from the route.

Moving on, the computational complexity of anomaly detection using AIS Data [33] presents another significant barrier to clustering, as many algorithms require quadratic time complexity for pairwise distance calculations, making them impractical for real-world purposes due to the volume of AIS data [17, 33].

Furthermore, these methods suffer from sensitivity to parameter selection; algorithms like K-means and DBSCAN require careful tuning of critical parameters (e.g., number of clusters K, or Eps and MinPts), where *"simplistic approaches to parameter selection may not always yield the best clustering results" [3].* This manual parameter tuning often leads to unstable clustering results and high false alarm rates when thresholds derived from clustering are applied to anomaly detection [33].

In addition, the spatio-temporal dimension of maritime trajectory data poses additional challenges when used with traditional clustering algorithms designed for point data. As mentioned by [17], AIS-based trajectory data cannot be directly used for trajectory clustering, as vessel trajectories consist of complex sequences with spatio-temporal characteristics. While methods using Dynamic Time Warping (DTW) or waypoint-based clustering can group trajectories by spatial similarity, they do not account for the temporal aspect, potentially missing crucial patterns that distinguish normal from anomalous behavior.

Finally, clustering techniques may also suffer from the "curse of dimensionality" in high-dimensional spaces, where distance measures lose their effectiveness in distinguishing between normal and anomalous instances [45]. Clustering approaches result in systems with limited universal applicability, as models trained on specific geographic regions or vessel types require extensive recalibration for deployment in new contexts [35].

Deep Learning Based

AIS data exhibits both high-volume characteristics and complex pattern formations due to geographical factors, temporal correlations, and human factors [27, 20, 33, 4, 1]. Next, no ground truth datasets exist and due to that, supervised anomaly detection methods are inapplicable [27]. For those reasons, Deep Learning based methods for anomaly detection are commonly used because of their autonomous ability to learn complex patterns from data [18, 27, 20, 51]. Deep learning methods provide a solution to the volume and complexity challenges of AIS data [33], translating trajectories into a latent space. The latent space represents a compressed variant of the original trajectory, in which the most important patterns of a trajectory are presumed to be emphasized [25].

Proven methods for anomaly detection involve autoencoders for learning normal trajectory patterns and identifying deviations, LSTM networks for capturing temporal dependencies in vessel movement sequences, and CNNs for processing AIS-derived navigation images to detect collision risks [55]. These models excel at handling the sequential nature of maritime data while automatically learning complex spatio-temporal patterns that traditional rule-based systems cannot capture effectively [2, 18].

As the common approach for maritime anomaly detection involves creating a normalcy model, computing reconstruction error, and using that to define abnormal behavior [33, 36], predictive models are commonly used [52].

2.2.2 Explanability

Although the use of deep learning techniques supports overcoming the challenges inherent to AIS data by representing high-dimensional trajectories in compressed latent space, the model outcomes are often not interpretable by humans [33, 27, 20], making their deployment in real-world scenarios infeasible [55, 56].

2.2.3 Research Gaps and Goals

While established frameworks, such as EMSA's ABM categorization, provide structured approaches to identifying known anomalous behaviors, the true potential lies in leveraging deep learning models that operate independently of these predefined taxonomies. Unlike rule-based systems constrained by existing classifications, neural networks may identify patterns and deviations that may represent entirely novel forms of anomalous maritime behavior—behaviors that current literature and policy frameworks have yet to recognize or categorize.

By incorporating meteorological variables alongside AIS trajectory data, these models can capture the complex interplay between environmental conditions and vessel behavior, potentially revealing new types of weather-influenced anomalies that traditional classification schemes have overlooked.

This capability to discover unknown unknowns positions data-driven anomaly detection as not merely a tool for identifying established threat categories, but as a means of expanding our fundamental understanding of what constitutes anomalous behavior in the maritime domain.

Gap For Incorporation of Meteorological Data

The integration of environmental context represents a critical gap in current maritime anomaly detection methodologies. While existing approaches have made significant advances in detecting trajectory anomalies through spatial-temporal analysis, they fundamentally overlook the influence

of meteorological conditions on vessel behavior [33, 36, 54]. Current state-of-the-art methods, including unsupervised deep learning frameworks such as [18, 20], focus primarily on identifying deviations from historical movement patterns and spatial clustering behaviors, yet fail to account for the influence of the meteorologic context during normal maritime operations [55, 43]. This limitation results in incorrect classification of weather-induced behaviors as anomalous, while potentially missing genuine anomalies that occur during adverse weather conditions [25, 31].

Anomaly detection research within the remote sensing domain by [21] relates to the importance of incorporating contextual information. Traditional remote sensing anomaly detection approaches demonstrated the importance of contextual analysis in Earth observation data. Spatial-temporal context alone proved insufficient for accurate anomaly identification capabilities without considering underlying environmental processes. The maritime domain presents analogous challenges, where vessel trajectory patterns are inherently influenced by meteorological factors such as wind speed and wave height [4, 1, 12, 24] making weather-agnostic anomaly detection fundamentally incomplete [36]. Furthermore, the dynamic nature of maritime weather conditions creates varying operational contexts that require adaptive anomaly detection thresholds [21], as vessel behavior that appears normal in calm conditions may be highly anomalous during severe weather, and vice versa [26]. This gap becomes particularly pronounced when considering the practical deployment of anomaly detection systems in real maritime operations, where false positives from weather-induced movements can overwhelm operators and reduce system usability [50, 40, 31].

Chapter 3

Methodology

In this thesis, the CRISP-DM framework for data mining was followed [49]. The CRISP-DM framework is particularly well-suited for this project, as it supports its intensive domain knowledge requirements and the integration of several heterogeneous data sources, requiring advanced data processing techniques. Next, since this project is conducted in collaboration with *Rijkswaterstaat*, the use of this framework is beneficial since it satisfies both academic and business requirements. In the following sections, the steps taken for this project are presented in the context of the generic phases of the CRISP-DM framework, excluding the deployment phase. Finally, CRISP-DM's industry-standard approach contributes to reporting in a manner that is both transparent and reproducible, a desirable characteristic due to the collaborative nature of this project.

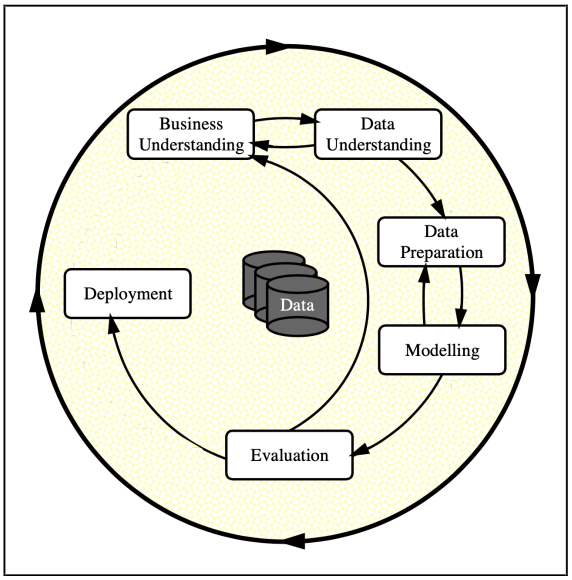

Figure 3.1: Phases of the CRISP-DM Process Model for Data Mining [49]

3.1 Business Understanding

Maritime anomaly detection contributes to improving Maritime Situational Awareness (MSA) [20, 33]. Anomaly detection advancements in this field result in systems with enhanced capabilities for identifying vessel behaviors that deviate from typical or expected patterns [36]. More practical, a country has a coast guard and their purpose is to ensure security at the maritime level by offering support in fatal situations. MSA systems aid this purpose by acting proactively; De-

tecting potential incidents prematurely can initiate coastguards' actions to avoid material damage or the loss of human life [20].

In the initial phase other of fatal situations, there exists an information asymmetry whereby maritime authorities remain unaware of the developing situation until vessels formally report distress conditions. Upon notification, rescue operations are initiated, and emergency response vessels are dispatched to the incident location. The temporal interval between incident occurrence and official response deployment represents a critical dispatch window during which emergency conditions can escalate from minor operational difficulties to major life-threatening situations [33]. MSA systems help prevent such situations by shortening the dispatch window.

The process of providing support in fatal situations is guided by maritime traffic monitoring staff [15, 34]. They are tasked with maintaining supervision over the situation, making responsible decisions. MSA systems support these tasks. Thus, maritime traffic monitoring staff might benefit from these systems. Due to human involvement, they must strike a balance between complexity and explainability. If an operator's decision is based on those tools, they need to be able to justify their decisions. In conclusion, a successful MSA system is both easy to comprehend and helps shorten the dispatch window.

3.2 Data Understanding

To include data describing the external metocean state in the dataset, individual AIS data points are enriched with features for wind, ocean current, and waves. In the following section, the source and characteristics of the different datasets used for this purpose are described.

3.2.1 AIS Data

Each AIS data point constitutes a transmitted message from a vessel containing standardized attributes, including timestamp, vessel identifier (MMSI), latitude, longitude, speed over ground (SOG), and course over ground (COG). The primary function of AIS data within maritime operations is to enhance navigational safety through real-time communication of positional and kinematic parameters between vessels. Given the limited maneuverability of large vessels and their operation under conditions of restricted visibility, AIS facilitates collision avoidance through continuous positional awareness [27].

In addition to its initial purpose, AIS data can be used to achieve various goals. For example, in [1], it was used to predict the risk of maritime accidents for insurance policies. For this research, the purpose is similar in the sense that it also involves improving nautical safety by detecting anomalous vessel behavior. However, it differs since the scope is not limited to individual vessels.

The dataset containing AIS data are sourced from work by [18] and contains processed vessel trajectory data. This dataset was originally derived from the public Vessel Traffic Data repository maintained by the U.S. Coast Guard and is spatially divided for the West Coast and the Gulf of Mexico. Before preprocessing, the complete dataset contains 34,209,386 individual AIS messages, which are aggregated into 17,217 unique vessel trajectories of which 7,340 belong to the West Coast. The data spans the period from January 2023 to September 2023.

3.2.2 Meteorological Data

The ocean state has different characteristics depending on the waves, current, and wind [12]. Since vessels use the ocean as their infrastructure, its state has a moderating effect on how a vessel behaves [4]. Sea states with high waves might cause ships to follow different trajectories than when the sea is calm [24]. For example, in [7] it was found that captains choose trajectories closer when

rough sea states occur. Because of this, anomaly detection models that do not incorporate these external conditions might predict a vessel's normal, adaptive maneuvers as anomalous [36, 33].

Figure 3.2 demonstrates this effect by comparing the spread of the standard deviation of Speed Over Ground (SOG) between two groups of trajectories. The boxplot indicates that trajectories occurring during high wave conditions exhibit significantly greater variation in speed compared to those in calmer seas. This wider spread might indicate that wave height influences ship velocity patterns.

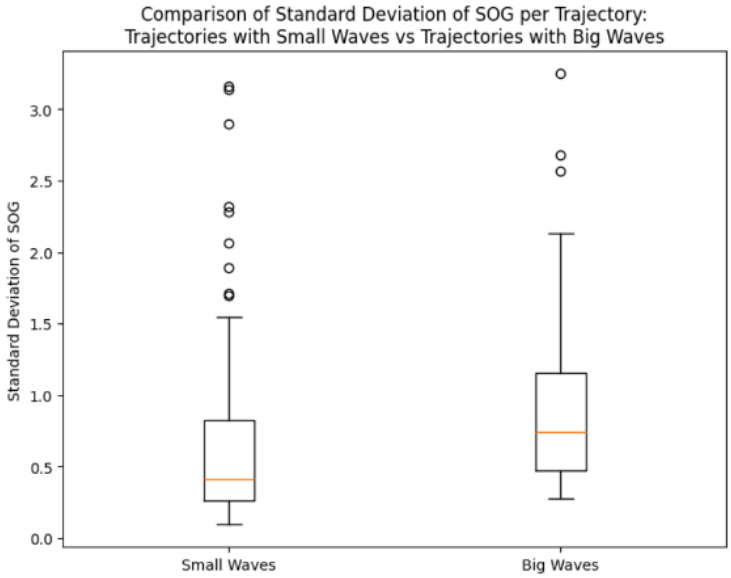

Figure 3.2: Comparison of trajectory groups based on wave conditions. Trajectories are categorized into small wave and large wave groups based on the maximum wave height during transit.

To enrich AIS-based vessel trajectories, meteorological variables describing prevailing weather and sea conditions were sourced from the ERA5 reanalysis, provided by the Copernicus Climate Data Store. ERA5, developed by the European Centre for Medium-Range Weather Forecasts (ECMWF), supplies globally consistent and comprehensive estimates of atmospheric, ocean-wave, and land-surface parameters from 1940 to the present. It assimilates observations from various platforms—including satellites, ships, buoys, and ground stations—using advanced physical modeling and data assimilation techniques, thereby producing robust reconstructions of past environmental conditions [13].

For this research, the "ERA5 hourly data on single levels" product was used, which offers variables on a regular latitude-longitude grid (0.25° for atmospheric and 0.5° for ocean-wave variables). Although this dataset offers various variables, literature showed that the following surface and near-surface parameters proved to have the most significant effect on vessel behaviour [1, 4, 24, 12]: Eastward component of wind at 10 meters above sea level, Northward component of wind at 10 meters above sea level, Significant wave height, Mean wave period, Mean wave direction.

The spatial and temporal resolution of the ERA5 data was aligned with the coverage of the relevant AIS datasets from both the U.S. West Coast and the North Sea, allowing precise matching between ship movements and environmental conditions. This integration enables a robust analysis of how metocean factors may influence maritime anomaly detection.

Figure 3.3 shows meteorological variables plotted over a trajectory. For wave height, the trend is slightly increasing. This might be explained by the vessel passing through the Golden Gate Bridge (Latitude: 38, Longitude: -122.4) and setting course for open waters. As a result, the wave height increases.

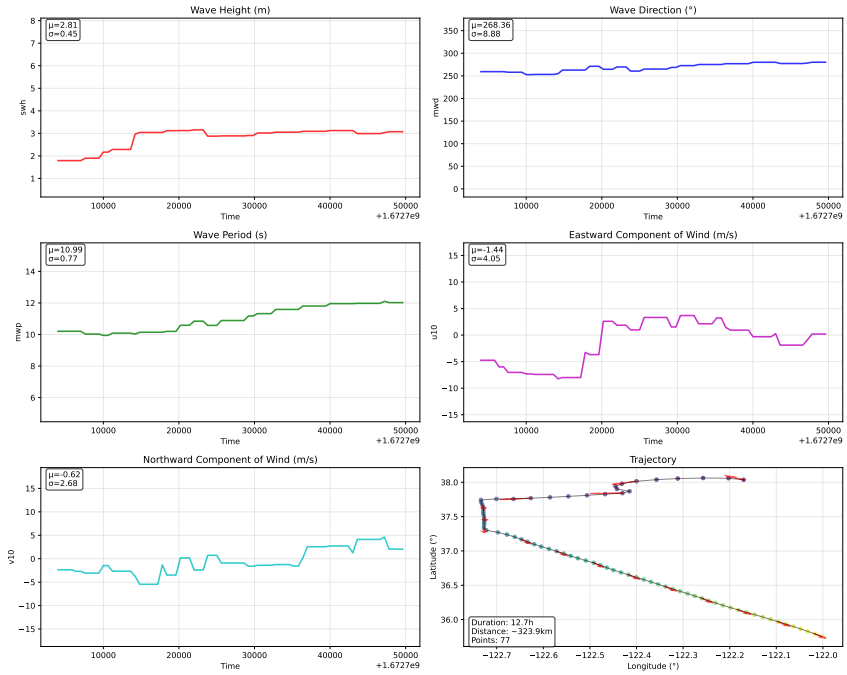

Figure 3.3: Variables originated from the ERA5 dataset over the time span of a trajectory.

3.3 Data Preparation

The AIS-based trajectory dataset utilized in this study was obtained in its preprocessed form from [18]. For a detailed description of the procedures used to convert raw AIS data into trajectory data, the reader is referred to [18]. In this section, we describe the additional processing steps applied to the preprocessed dataset to integrate meteorological data with the vessel trajectories. Particular emphasis was placed on ensuring that these procedures are reproducible.

3.3.1 Feature Engineering

The preprocessed trajectories were provided as matrices, where the columns were scaled according to the feature spread. Based on the provided data processing source code, the features were rescaled based on their original values.

3.3.2 Joining AIS Data With Meteorological Measurements

The process of linking individual AIS data points to metocean conditions is described in the following section and involves two main stages: creating a standardized metocean grid and performing spatial-temporal joins between AIS data points and weather data.

In the first stage, a spatial grid was constructed to standardize the varying resolutions of different metocean variables from the ERA5 dataset. The grid generation process is initialized by defining the study area boundaries [18] and creating spatial tiles using the web Mercator projection formulas at a zoom level of 11. This zoom level was selected to approximate the spatial resolution

of the ERA5 wind data (0.25° × 0.25°). Since wave data in ERA5 has a coarser resolution (0.5° × 0.5°) compared to atmospheric variables, forward filling was applied to match the finer grid resolution.

Figure 3.4 shows the spatial dimension of the ERA5 data points mapped to the tile grid (green). For each grid tile, the nearest ERA5 data point is identified using spatial joins. This process creates a comprehensive dataset, where each grid tile contains time-series meteorological data for the entire study period.

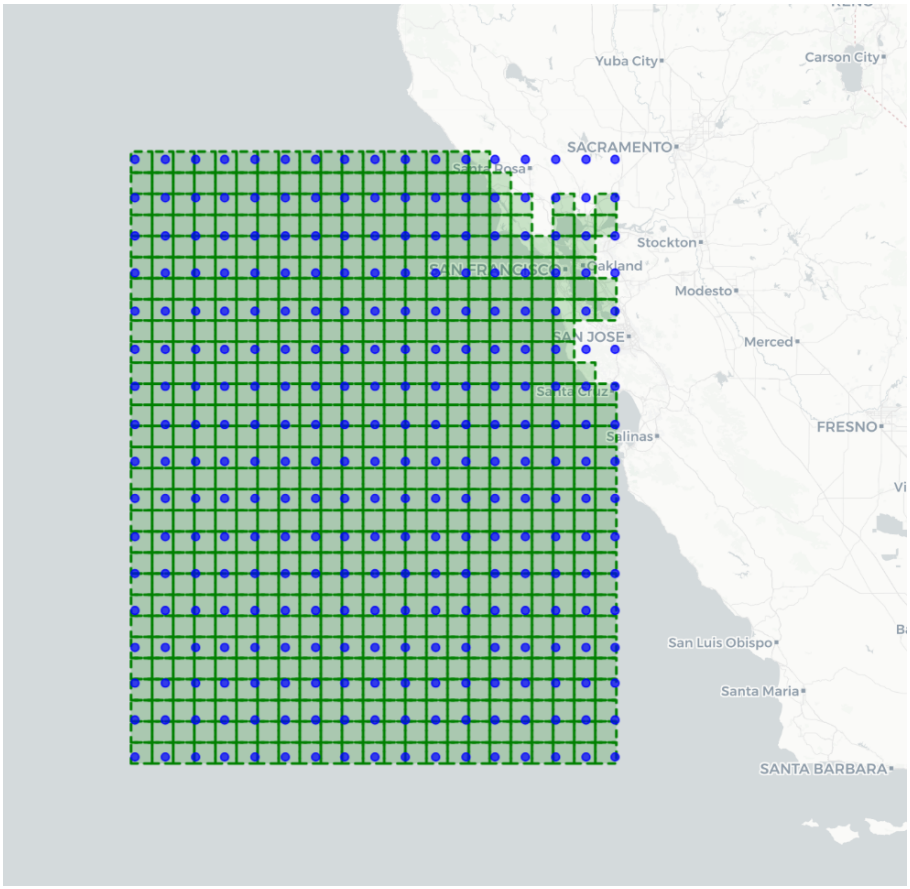

Figure 3.4: The ERA5 datapoints (blue) plotted on the spatial tile grid

In the second stage, individual AIS data points are linked to metocean conditions through a combination of spatial and temporal joins. Each AIS message, containing latitude, longitude, SOG, COG, and timestamp, is first assigned to a grid tile using spatial containment queries. The timestamp of each AIS message is then used to extract the corresponding hourly weather conditions from the assigned tile's metocean time series. This results in each AIS data point being augmented with information on wave height, wave direction, wave period, and wind speed.

The final joined dataset preserves all original AIS trajectory information while adding five additional metocean variables that represent the environmental conditions at each vessel's location and time. This integrated dataset enables analysis of ship behavior under varying weather conditions and supports maritime safety applications.

3.3.3 Dataset Partitioning

The original data partitioning strategy in [18] employs a chronological splitting approach that divides the dataset into three temporally distinct periods: a training set comprising data from January to July 2023 (7 months), a validation set containing August 2023 data (1 month), and a test set using September 2023 data (1 month). This temporal partitioning methodology serves two primary purposes: first, it prevents temporal information leakage between future and historical observations, thereby ensuring robust evaluation of the model's generalization capability to unseen time periods; second, it simulates realistic deployment scenarios where models are trained on historical data and subsequently applied to future observations.

In accordance with the unsupervised learning framework, the original partitioning strategy includes only normal ship trajectories in the training and validation sets, while anomalous behaviors are artificially injected exclusively into the test set through controlled perturbations representing shift deviation, abnormal heading, and abnormal speed patterns. This approach reflects practical deployment conditions where labeled anomalous trajectory data is typically scarce or unavailable during the training phase.

For the present investigation, the partitioning strategy was modified to address the specific research objective of evaluating the integration of meteorological features. Rather than maintaining the chronological division, the normal trajectories from the original training period (January-August 2023) were randomly redistributed between training and validation sets while preserving the temporal integrity of the test set. This methodological adaptation serves to isolate the effect of meteorological features by ensuring that both training and validation sets contain representative samples distributed across all temporal periods and corresponding weather conditions. By eliminating potential seasonal bias that could confound the evaluation of meteorological feature contributions [12, 1, 4], this approach provides a more controlled experimental framework for investigating the effects of weather data integration. While this modification deviates from the original STAD temporal partitioning designed for time-series generalization assessment, it establishes a methodologically sound foundation for examining the specific research question regarding the impact of meteorological data on maritime trajectory anomaly detection performance.

3.4 Modeling

The movement of a vessel is determined by its direction and speed; however, its actual behavior depends on its spatial location, maritime traffic, and external conditions. What constitutes abnormal behavior is thus context-dependent [33, 35]. For that reason, this research extends the methodology proposed by [18]. The original approach learns vessel kinematics through a deep learning model based on the transformer architecture [44], guided by spatially dependent patterns captured by a Gaussian Mixture Model (GMM). We investigate whether the inclusion of meteorological variables will enhance anomaly detection capabilities.

This enhanced version of [18] produces anomaly scores that reflect vessel movement patterns and contextual factors. This methodology was chosen for its ability to combine the strengths of both clustering and deep learning anomaly detection methods, capturing complex spatio-temporal dependencies and learning underlying clustering patterns of shipping routes.

3.4.1 Data Representation

Processed AIS-based trajectories are mapped onto a 2D grid, transforming continuous data into discrete representations. Similar data representation approaches can be found back in literature [20, 7, 43, 21, 54] and potentially serve as solution to the volume and complexity challeges of AIS-data [33]. The key advantages within this approach are: creating a more granular search space, which is beneficial for deep learning optimization, handling circular variables, such as course over

ground [54, 18, 20], and allowing for simplified coupling of AIS data to external sources [43]

Trajectories are organized as time-ordered sequences $T = s_1, s_2, \ldots, s_m$, where each point s_t captures the vessel's status through four attributes: latitude, longitude, SOG, and COG. To handle irregular AIS transmission intervals, linear interpolation was used to impose temporal consistency between data points, yielding standardized sequences $\mathbf{X} = [x_1, x_2, \ldots, x_T]$ [18].

For model training, each trajectory is segmented into overlapping windows of $w = 10$ time steps. The features of each window are discretized into bins and transformed into one-hot encoded feature vectors, $\mathbf{F}_t$. Following the resolutions by [18], latitude and longitude are both binned at a spatial resolution of 0.01° (yielding 400×400 bins), SOG is binned in 1-knot increments (30 bins), and COG is discretized into 5° intervals (72 bins). The number of bins for COG is chosen to be divisible by 12, as the 12-hour divisions of a clock are used to represent compass bearings.

3.4.2 Model Architecture

The architecture consists of three interconnected components that work synergistically to detect trajectory anomalies: (1) offset reconstruction-based representation learning, (2) dynamic pattern clustering, and (3) sliding anomaly scoring. Figure 3.5 provides a visualization of the architecture as used in [18]. This multi-component design captures temporal dependencies and spatial clustering characteristics while accounting for environmental conditions.

A fundamental assumption of the approach demonstrated in [18] is its treatment of ship trajectory prediction as a multi-modal problem. This implies that the conditional probability distribution, $p(t|1:T-1)$ (vessel position given trajectory T), should be modeled as multi-modal rather than uni-modal, based on the observation that ship movements are composed of multiple possible behaviors including forward motion, left turns, and right turns. This multi-modal assumption implies that at any given time step t, vessel may follow several viable trajectory patterns rather than converging toward a single expected path (uni-modal). Consequently, the model employs a classification-based formulation for the training objective (see Section: 3.4.5 using cross-entropy loss instead of traditional mean squared error reconstruction loss. In the following sections, the model architecture (see: Figure 3.5) is explained in detail, with special attention given to reproducibility.

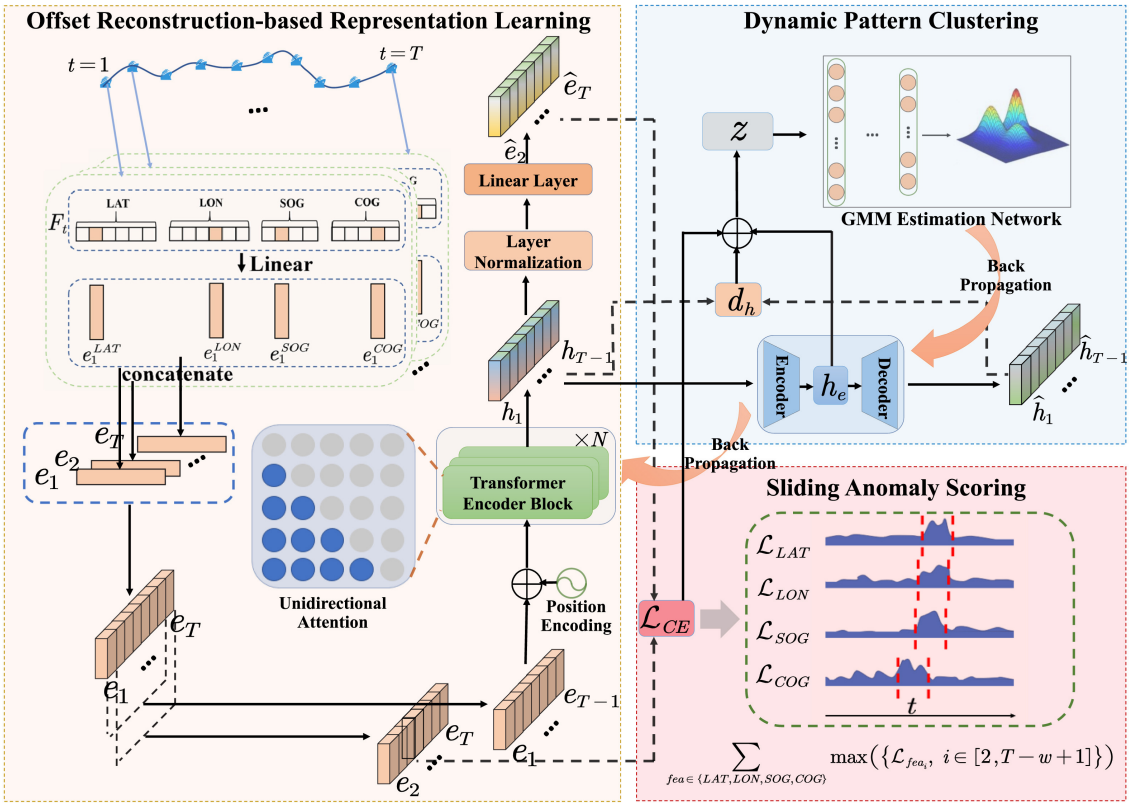

Figure 3.5: Model Architecture proposed by [18].

Offset-Based Representation Learning

The sequence-based component employs the transformer encoder architecture to transform each trajectory window to a latent representation. Importantly, causal self-attention is utilized, ensuring that predictions for the subsequent time step are influenced only by the previous, a critical requirement for real-time anomaly detection.

Transformer Encoder This implementation exclusively employs the encoder of the transformer architecture. This design choice reflects the specific objective of learning trajectory window representations and capturing forward temporal dependencies within fixed-length trajectory segments, which differs fundamentally from sequence-to-sequence applications that require token-by-token output generation, such as machine translation tasks [51]. The encoder is well-suited for this problem, as it enables the parallel processing of trajectory windows and captures both spatial and temporal dependencies. In contrast, the decoder is typically required for tasks that involve interpretable model outputs, such as autoregressive generation or decoding structured outputs, which is unnecessary for predicting subsequent movements based entirely on the input window. By using only the encoder, this implementation leverages the attention mechanism, achieves computational efficiency [44], and aligns with the requirements of real-time anomaly detection.

Sliding Window Training The model processes embedded trajectory windows to predict subsequent movements learned through offset reconstruction. The input embeddings for each feature are concatenated and enhanced with positional encoding before being passed through the transformer encoder layers. The input to the transformer encoder component is an embedded window $w_t = \{e_1, \ldots, e_{10}\}$. The model is trained to predict the next window $w_{t+1} = \{e_2, \ldots, e_{11}\}$. During training, batch sizes vary depending on the number of possible windows per trajectory. The number of possible windows (N_w) per trajectory (T) is given by the following formula:

$$N_w = (|T| - 10)$$

Trajectories are temporally split, resulting in durations varying between 4 and 24 hours and interpolated to ensure temporal consistency with a frequency of 10 minutes [18]. Resulting in trajectories varying between 24 and 144 data points.

Applying the formula will result in batches varying between 14 and 134 windows. Importantly, since we require sequence w_t to predict the next sequence w_{t+1} for a trajectory of length $|T|$ the last sequence prediction is limited to using $w_{(|T|-1)}$ to predict $w_{|T|}$. In practical terms, data leakage is prevented by excluding the last trajectory window from the sequences used for making predictions. To summarize, batches represent N_w trajectory windows w, containing $\{w_1, \ldots, w_{|T|}\}$ windows.

Embedding Layers As mentioned before, AIS-based features (latitude, longitude, SOG, and COG) are represented as one-hot encoded vectors, with sizes corresponding to the same binning resolution as used in [18]. Each feature vector ($\mathbf{e}_t^{feature}$) is embedded using a separate equally sized embedding and then concatenated to form a single embedding vector ($\mathbf{e_t}$), representing a single sliding window (w_t).

$$\mathbf{e}_t = [\mathbf{e}_t^{LAT}; \mathbf{e}_t^{LON}; \mathbf{e}_t^{SOG}; \mathbf{e}_t^{COG}]$$

The implementation in this work represents an architectural variation from the approach used in [18]. One-hot encoded vectors are mapped through individual embedding layers and then concatenated as opposed to concatenating the one-hot encoded vectors before passing them through a high-dimensional linear layer.

Positional Encoding The positional encoding implementation follows the standard sinusoidal approach introduced by [44]. For a given position *pos* in the sequence and dimension i within the embedding vector ($\mathbf{e_t}$), the positional encoding is computed as:

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

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

where d_{model} represents the model dimension, equivalent to $4 \cdot d_{embed}$ in our embedding approach, *pos* is the position index ranging from 0 to $w-1$, and i is the dimension index within the embedding vector. The encoding alternates between sine and cosine functions for even and odd dimensions, respectively.

Dynamic Pattern Clustering

The purpose of the Dynamic Pattern Clustering component (see: Figure 3.5) is to group latent representations of trajectories exhibiting similar behavioral patterns [18]. In the following paragraphs, its internal mechanisms are explained in more detail.

Auto Encoder To avoid high-dimensional representations [45], an auto-encoder model is used for compressing latent trajectory window representations. Auto encoders are unsupervised neural networks composed of an encoder and a decoder and have applications in different domains [28]. The encoder maps input data into a latent variable space, and the decoder maps these latent variables back into the input space, aiming to reconstruct the original input [41, 2].

In this implementation the autoencoder is employed to compress the dimensionality of the latent representation of the trajectory $\{h_1 \dots h_{|T|-1}\}$ (see Figure: 3.5) to the autoencoder latent variable space ($\mathbf{h_e}$). The encoder component is a three-layer linear feedforward neural network that progressively compresses the input representation $\mathbf{h_e}$. The decoder mirrors the encoder architecture but operates in reverse, progressively expanding the compressed trajectory representation back to the original input dimensions $\{\hat{h_1} \dots \hat{h}_{|T-1|}\}$.

The input dimensionality of the autoencoder depends on the model dimensionality of the transformer encoder ($4 \cdot d_{embed}$). The transformer encoder predicts two-dimensional trajectory window representations with dimensionality: $(|w|, 4 \cdot d_{embed})$. The input to the autoencoder is one-dimensional and is equal to $(|w| \cdot 4 \cdot d_{embed})$.

Gaussian Mixture Model Gaussian Mixture Models (GMMs) are a class of probabilistic models that represent data as a combination of multiple Gaussian distributions, each with their own mean and covariance structure. This method provides a flexible approach to modeling complex data distributions that may not conform to a single, simple shape. By assuming that each data point is generated from one of several underlying Gaussian distributions, GMMs can capture the presence of subpopulations within heterogeneous datasets. The model parameters are typically estimated using the Expectation-Maximization (EM) algorithm [32], which iteratively refines the estimates of each component's parameters while assigning data points to the most probable distributions. This ability to model data as a weighted sum of Gaussian components makes GMMs a powerful tool for clustering, density estimation, and anomaly detection in a variety of applications [32].

In this work, the Gaussian Mixture Model component is implemented not as a separately trained module using EM, but as a differentiable neural network. All GMM parameters, including mean values, covariances, and mixture weights, are treated as learnable parameters, jointly optimized with the transformer and autoencoder through gradient-based backpropagation [47].

The implementation of Gaussian Mixture Models through deep neural networks, as demonstrated by [47], involves replacing the traditional iterative EM algorithm with a multi-layer neural network architecture that directly estimates GMM parameters. In their Deep Gaussian Mixture Model (DGMM) framework, a specialized estimation network employs multiple fully connected layers to learn the mixture weights, means, and covariances of Gaussian components from compressed latent representations. Rather than relying on the computationally intensive EM algorithm, the network architecture enables end-to-end gradient-based optimization where GMM parameters are

continuously refined through backpropagation. This approach offers the advantage of handling high-dimensional data more effectively. The resulting system maintains the probabilistic foundations of traditional GMMs while leveraging the representational power and computational efficiency of deep learning architectures.

Building upon this deep neural network approach to Gaussian Mixture Models, [18] demonstrate the practical application of integrated GMM estimation networks in the maritime domain through their deep learning framework. Their implementation follows the DGMM paradigm [47] by incorporating a deep GMM estimation network. In [18], the GMM estimation network serves as a training constraint that guides the transformer encoder to learn meaningful trajectory representations by enforcing clustering patterns in the latent space, rather than operating independently during inference.

In [18], the input to the GMM estimation network is defined as follows: $\mathbf{z} = [\mathbf{h_e}; \mathcal{L}_{\text{TE}}; d_{\text{AE}}]$ where h_e is the compressed latent trajectory representation, d_h is the auto-encoder reconstruction error (see: Section 3.4.5), $\mathcal{L}_{\text{TE}}$ is the cross-entropy loss (see: Section 3.4.5).

This research's key contribution lies in extending the GMM formulation to incorporate meteorological variables as opposed to enriching individual data points within trajectories, which is found in the temporal inconsistency between these two data sources [24]. In addition, preliminary analysis showed little variation within the data over the duration of trajectories (See: Figure 3.3). Thus, we assume that this implementation ensures the retention of meteorological information per trajectory while minimizing computational complexity [24], thereby enabling the model to distinguish between environmentally influenced behavior changes and true anomalies. For each trajectory, the mean and standard deviation for each of the metocean variables are computed as follows:

$$\mathbf{m}_{\mu,\sigma} = [\mu^{\mathbf{meteo}}, \sigma^{\mathbf{meteo}}]$$

where i represents the i-th meteorological variable, $\mu^{\mathbf{meteo}}$ and $\sigma^{\mathbf{meteo}}$ represent the mean and standard deviation of each meteorological variable over the trajectory. Meteorological statistics are concatenated with the original formulation to form an enriched feature vector:

$$\mathbf{z}_{weather} = [\mathcal{L}_{\text{TE}}; \mathbf{h_e}; d_{\text{AE}}, \mathbf{m}_{\mu,\sigma}]$$

3.4.3 Inference Anomaly Scoring

During training, the model optimizes the composite loss function $\mathcal{L}_{\text{TE}}$ to learn normal trajectory patterns and environmental correlations. However, for testing, a different scoring mechanism is employed. The testing anomaly score utilizes a sliding window approach that captures local deviations rather than global trajectory characteristics. For each trajectory window, the model computes attribute-specific reconstruction errors for latitude, longitude, SOG, and COG. The anomaly score for testing S is calculated as [18]:

$$S = \sum_{f \in LAT,\ LON,\ SOG,\ COG}^{f_i} \max(\mathcal{L}_{f_i} : i \in [2, T-w])$$

where $\mathcal{L}_f^{(i)}$ represents the cross-entropy loss for feature f at time step i, and w is the sliding window length. This formulation identifies the maximum reconstruction error within each sliding window for each trajectory attribute, effectively highlighting localized anomalous segments that might otherwise be masked by averaging across the entire trajectory.

3.4.4 Model Integration

The integrated model produces anomaly scores addressing three key questions: Is the vessel's movement pattern normal? Is this behavior expected in this location? Is this behavior appropriate given current weather conditions? The final anomaly score is computed by combining the transformer's reconstruction error with the probabilistic cluster assignment from the GMM, providing a weighted metric that reflects both sequential prediction errors and deviations from learned cluster distributions.

3.4.5 Training Objective

During training, the model minimizes a composite loss function ($\mathcal{L}_{\text{STAD}}$) that combines temporal and spatial objectives and is given by:

$$\mathcal{L}_{\text{STAD}} = \mathcal{L}_{\text{TE}} + \lambda_1 E_{\text{GMM}} + \lambda_2 d_{\text{AE}} + \lambda_3 p_{\text{GMM}}$$

where $\mathcal{L}_{\text{STAD}}$ is the transformer encoder loss, E_{GMM} is the negative log-likelihood of the trajectory belonging to the Gaussian components, d_{AE} is the auto-encoder reconstruction loss, and p_{GMM} is the GMM penalty term regularizing covariance values. The implication of each of the components of this function will be elaborated on in the following sections.

Transformer Encoder Loss

($\mathcal{L}_{\text{TE}}$): $\mathcal{L}_{\text{TE}}$ relates to the ability of the model to reconstruct the kinematic behavior of a vessel trajectory. Computed over the sum of cross-entropy losses across predicted trajectory features (latitude, longitude, SOG, and COG) for each window in the trajectory:

$$\mathcal{L}_{\text{TE}} = \sum_{f \in \{\text{lat, lon, sog, cog}\}} \text{CE}(f_{\text{logits}}, f_{\text{true}})$$

Autoencoder Reconstruction Loss

(d_{AE}): Measures the squared error between the latent representation of the next windows in a trajectory predicted by the transformer encoder and the reconstructed latent trajectory representations by the autoencoder:

$$d_{\text{AE}} = (h_i - \hat{h}_i)^2$$

Energy

(E_{GMM}) is computed as the negative log-likelihood under a learned Gaussian Mixture Model (GMM) for the latent representations. Energy is optimized for during training and its value represents the ability of the GMM to learn the distributions of the latent representations representing the behavioral vessel patterns [18]. relation to this implementation and refer the reader to [47] for more specific details regarding the mathematical formulations.

$$E_{\text{GMM}}(\mathbf{z}) = -\log\left(\sum_{c=1}^{C} \phi_c \mathcal{N}(\mathbf{z} \mid \mu_c, \Sigma_c)\right)$$

where $\mathbf{z}$ is the concatenated input vector comprising the compressed autoencoder representation (h_e), reconstruction error (d_h), and cross-entropy loss $\mathcal{L}_{\text{TE}}$; C represents the number of Gaussian components; ϕ_c, μ_c, and Σ_c are the mixture weights, means, and covariances respectively.

Penalty Term

(p_{GMM}): Regularizes component variances, expressed as the sum of reciprocals of the diagonal elements (plus ε for stability) from all GMM covariance matrices:

$$p_{\text{GMM}} = \sum_{k=1}^{K} \sum_{j=1}^{D} \frac{1}{\sigma_{k,jj} + \epsilon}$$

The loss corresponding to each of the components of the model is weighted by their respective hyper-parameters ($\lambda_1, \lambda_2, \lambda_3$), the total loss is minimized with respect to model parameters using gradient-based optimization throughout training.

3.4.6 Experimental setup

The rationale behind choosing the values for the model's hyper-parameters is elaborated on in the following. It should be noted that only hyper-parameters that are not explicitly listed in [18] are mentioned here. While [18] provided the model architecture, key training hyper-parameters were not reported. The key hyperparameters to this approach are elaborated on in the following paragraphs while their corresponding values used for conducting the experiments in Chapter 3 are listed in the Appendix 6.3.4.

Model Dimensions As the input data is transformed to one-hot encoded vectors, each vector is mapped via an embedding layer. Then, each embedding vector is concatenated so that the trajectory is represented as an embedding vector with a dimensionality of $4 \cdot$ *embedding dimension* and serves as the input for the transformer encoder.

Weight Optimization In the training context, AdamW [22], a variant of Adam that decouples L2 regularization and weight decay, is widely used for training Transformers. It generally converges faster, is more memory-efficient, and more accurate than Adam [56]. In addition, the size of a batch corresponds to the number of windows within a trajectory, leading to variability in batch size throughout training, which can potentially impact the stability of the learning process. We note that batch size variability may introduce noisier gradient estimates and unstable parameter updates, especially for smaller batch sizes [14, 37].

Learning Rate The use of the AdamW optimizer shows optimal performance with the use of scheduled learning rates multiplier with the cosine annealing strategy outperforming the others [56]. For that reason, we adopted this configuration.

GMM components The number of GMM components represents the number of groups that can be identified from the latent trajectory representations. As extensively tested in [18], the optimal number of components has been identified. For an initial comparison in this research, the same value was used; however, since the dimensionality of the trajectory representations increases after introducing weather statistics, additional experiments should be conducted to find the optimized number of components.

3.5 Evaluation

3.5.1 Artificial Anomalies

As noted by [33, 36], a significant challenge within the field of anomaly detection is the scarcity of labeled data, rendering it an unsupervised classification problem. A common approach is to introduce synthetic anomalies by perturbing or removing a sequence of data points of original trajectories [18, 20]. In this research, trained models are validated in a similar sense. We adopt three types of anomalous trajectories [18] by altering latitude, longitude, and SOG attributes and injecting them into the test set. In this way, models can be validated under controlled conditions

[39], enabling the structured evaluation of the model's effectiveness.

In this work, the three types of anomalies under investigation are: shift deviation, abnormal heading, and abnormal speeding [18]. In relation to the grid representation of the AIS-based trajectories , the following anomaly injection parameters are used [18]: d, the distance measured in grid cells that the trajectory deviates from its original course, r the ratio of anomalous trajectories within the test set and ρ, the number of data points within the trajectory for which the features are perturbed. Although the method for permuting trajectories is not explicitly described in [18], we reconstructed their approach for illustrative purposes. In the following paragraphs, we elaborate on the types of synthetic anomalies used for validation.

An *shift deviation* occurs when a ship steers towards one side abruptly while later continuing its original course; this behavior will be visible in a shift of its position with respect to its original trajectory. This type of anomaly is introduced by shifting either the lateral or longitudinal values by a fixed amount [18]. Figure 3.6 shows an example of how this anomaly injection method alters the original trajectory.

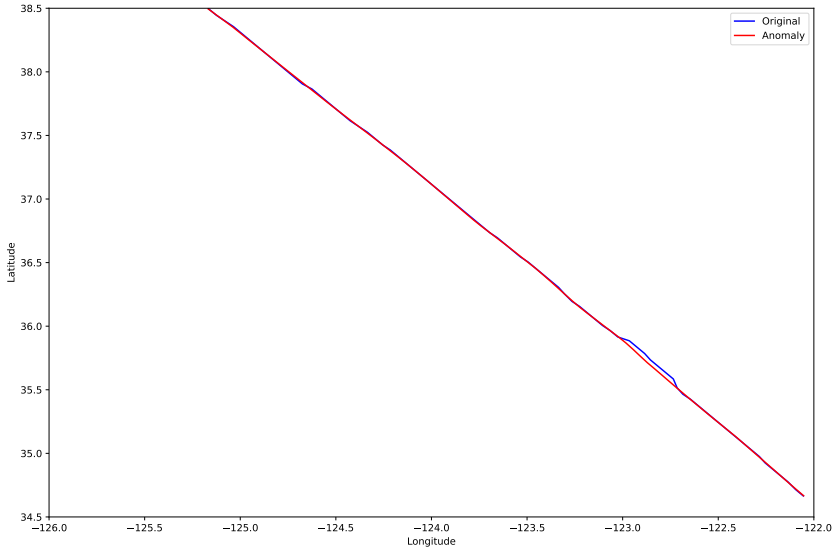

Figure 3.6: Illustration of a shift deviation anomaly. The original trajectory is perturbed by shifting ρ consecutive data points with a deviation of d grid cells. The precise spatial extent of this deviation is determined by the grid resolution used for mapping the AIS data points.

Abnormal heading is displayed through sways during the course of the vessel. This anomaly is introduced by perturbing both its position and its course over ground (COG) attributes. Figure 3.7 displays how this anomaly is imposed upon trajectories in more detail.

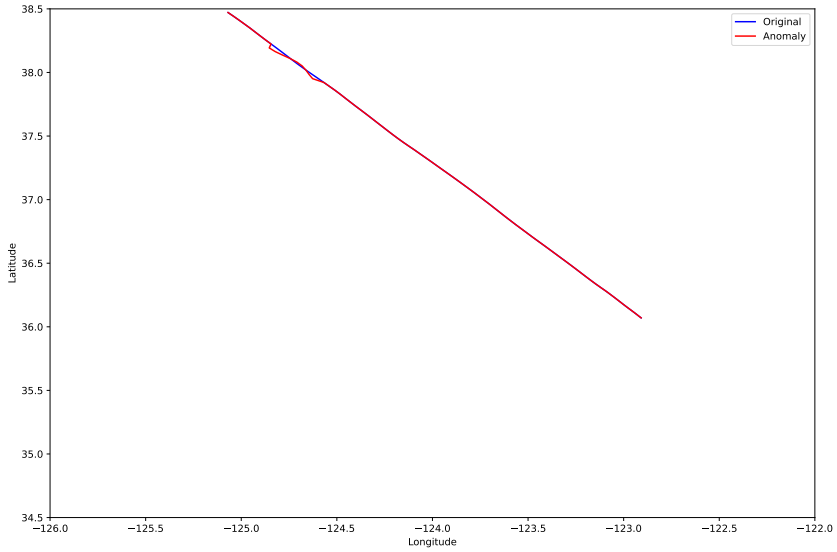

Figure 3.7: Illustration of an abnormal heading anomaly. In this example, the original trajectory is perturbed by superimposing a sine wave with an amplitude of d grid cells onto a subset of ρ trajectory points. The true spatial displacement is determined by the resolution of the grid to which the trajectory points are mapped. In addition to spatial alterations, the course over ground (COG) attribute of the affected points is perturbed by adding deviations sampled from a normal distribution with a maximum of 10 degrees.

Abnormal speeding can indicate problems with the propulsion system of the vessel and is imposed by adding Gaussian noise to the original velocity values. As this type of anomaly does not directly affect the vessel's position, parameter d is not applicable.

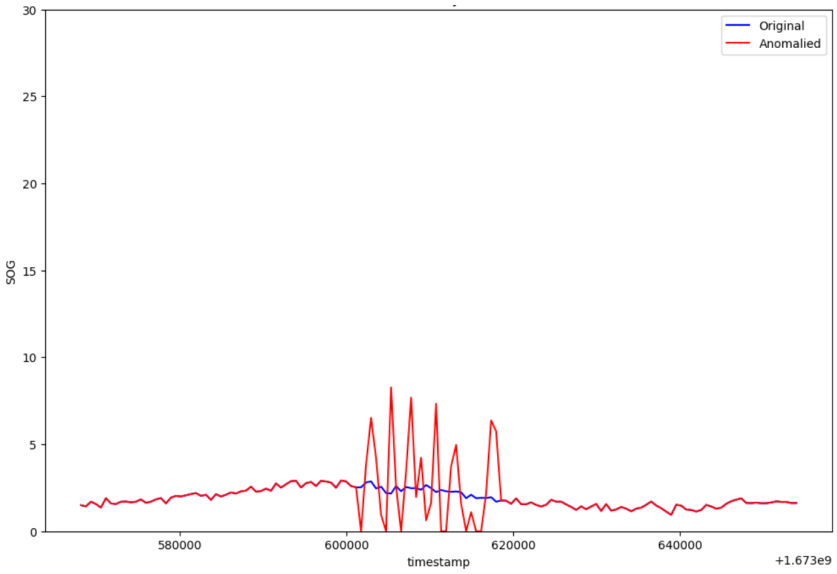

Figure 3.8: Illustration of a abnormal speeding anomaly. The original trajectory is perturbed imposing velocity deviations of ρ consecutive data points. The deviations are sampled from a normal distribution representing a deviation of maximal 3 knots on the SOG attribute of the trajectory.

These synthetic anomalies relate to the the framework of anomalous behaviors as defined by the EMSA [8] in the following way: shift deviation corresponds to *Sudden Change of Heading* (Figure 2.2 first row, fifth column), abnormal heading corresponds to *Drifting* (Figure 2.2 first row, second column), and abnormal speeding corresponds to *Sudden Change of Speed* (Figure 2.2 second row, fifth column)

Although evaluation with synthetic anomalies allows for controlled validation, its representativeness concerning real-world anomalies is questioned [33]. Actual anomalous behaviors are complex and depend on external factors, such as the geographical environment and interactions with other maritime traffic [27]. The method used in this work has limitations in terms of the variety and representativeness of anomalies. It enables quantitative model evaluation and addresses the issue of limited labeled data.

3.5.2 Evaluation Metrics

The model validation strategy is designed to systematically address each research question through carefully selected metrics that directly measure the model's ability to distinguish between weather-induced vessel movements and artificial maritime trajectory anomalies (see Chapter: 1). The validation framework employs threshold-dependent and threshold-independent metrics to provide a comprehensive assessment of model performance in relation to this work. We first contextually define each metric, then apply it to the associated research question.

Threshold Dependent Metrics

As described in Section 3.4.3, this modeling approach yields anomaly scores per trajectory based on reconstruction loss. For this reason, the computation of classification metrics depends on the threshold used to mark trajectories as anomalous. Since these metrics are highly sensitive to the chosen threshold, methodologies for setting the threshold are often criticized [28] or neglected [18]. The chosen threshold is presented strictly as a heuristic for academic assessment against our stated research objectives, readers should refrain from interpreting it as a validated threshold for real-world implementation of maritime anomaly detection systems.

Two primary approaches for threshold selection were considered: (1) the statistical approach using $\mu + \alpha \cdot \sigma$ (mean plus α standard deviations of the negative class), and (2) the performance-based Youden J statistic [53], $J = Sensitivity + Specificity - 1$ (also known as: bookmarker informedness within machine learning domains [6]). The statistical approach is commonly used in anomaly detection literature [28]. However, it assumes normal distribution of anomaly scores for trajectories marked as normal (negative instances) thus, its focus is limited to controlling false positive rates (specificity). Although this aligns with this research, it might introduce bias towards in relation to detecting positive instances (sensitivity). Therefore, we employ the Youden J statistic to compute the threshold for determining threshold-dependent evaluation metrics; it provides a balanced optimization that considers both classes simultaneously.

The threshold corresponding to the maximum Youden Index (c^*) represents the point on the ROC curve farthest from the diagonal line of no discrimination between classes, thus optimizing the overall classification accuracy when equal importance is assigned to sensitivity (true positive rate) and specificity (true negative rate) [38], providing balance between detecting anomalies and minimizing false alarms.

From a practical perspective, for each individual test set we determine the threshold independently in the following way; we use the model to predict anomaly scores per trajectory (see: Section 3.4.3); then, we note the true positive rate (sensitivity) and the true negative rate (specificity) varying the decision boundary (a possible threshold), we then apply the formula for computing J [53] given each threshold. Finally, we determine the value for the threshold (c^*) by based on the maximum value for J max_J. This allows for mapping the loss scores per trajectory to class label

prediction. We then compute the classification metrics, how they relate to our problem domain is described in the following paragraphs.

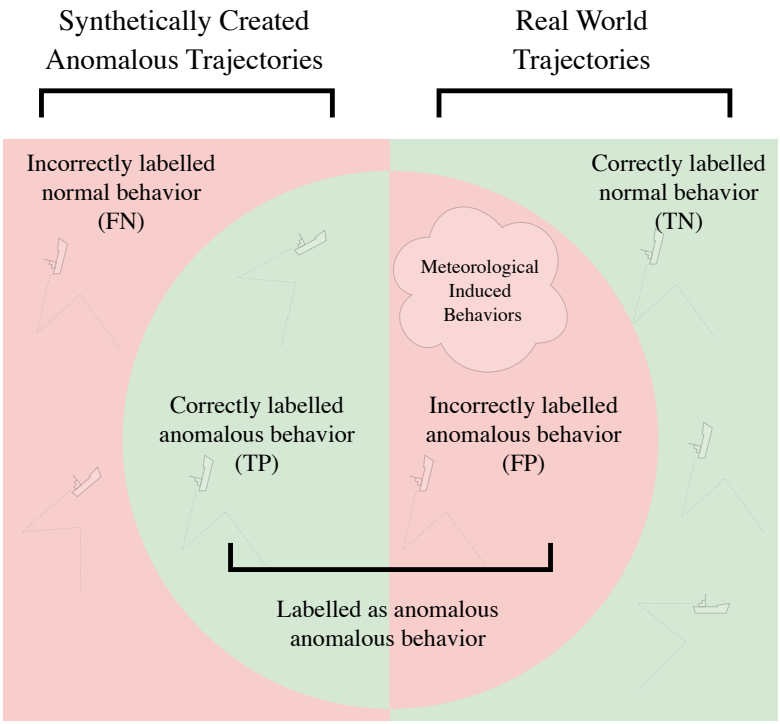

Figure 3.9: Classification Metrics in Relation to this Thesis.

From a geometric perspective, Figure 3.9 depicts a rectangle overlain by a circle. The rectangle represents the entirety of behaviors classified as normal, while the circle encompasses vessel behaviors labeled as anomalous. From a chromo-geometric standpoint, the primary focus of this study corresponds to the red overlapping region of the circle, which includes trajectories identified as anomalous by the model. It is hypothesized that behavioral patterns are substantially influenced by the severity of prevailing meteorological conditions. As these meteorological variables fall outside the scope of the model, behaviors induced by such conditions are likely to be incorrectly flagged as anomalous.

Figure 3.9 shows the confusion matrix metrics. The rectangle signifies all behaviors classified as normal, encompassing true negatives (TN) and false negatives (FN), while the circle comprises those behaviors labeled by the model as anomalous, including both true positives (TP) and false positives (FP). The red overlapping region within the circle specifically highlights trajectories deemed anomalous by the model, which, according to our hypothesis, may include behaviors actually induced by prevailing meteorological conditions. These meteorologically induced behaviors constitute a subset of the false positives (FP), as the model is not designed to account for external meteorological influences. As such, an increased incidence of FP—reflected as a decrease in specificity—serves as a direct indicator of the model's tendency to misclassify meteorologically influenced normal behaviors as anomalous. By monitoring the specificity metric, which quantifies the proportion of correctly identified normal behaviors (TN) among all actual normal cases (TN + FP), we assess the impact of meteorological conditions on the model's classification accuracy, thereby validating our hypothesis.

Threshold Independent Metrics

Operating Characteristics Curve (ROC) [11] plots the True Positive Rate (TPR) against the False Positive Rate (FPR) across all possible decision thresholds, where TPR represents the proportion of correctly identified anomalies (referred to as sensitivity) and FPR represents the proportion of normal instances incorrectly classified as anomalies (referred to as 1-specificity). The ROC Area Under the Curve (AUC) quantifies a model's discriminative ability between normal and anomalous classes, with values ranging from 0 to 1, where 0.5 indicates no discriminative ability and 1 represents perfect classification [28]. However, it should be noted that ROC AUC is not a good indicator for real-world performance, since it handles normal and artificial trajectories indifferently [28].

3.5.3 Evaluation Of Statistical Significance

To quantitatively assess the impact of meteorological data integration on model performance, this study conducts a statistical comparison between the baseline (AIS-only) model and the weather-enhanced model. We employ McNemar's test to evaluate the significance of differences in model specificity, focusing on the rates of true negatives (correctly identified normal trajectories) and false positives (misclassified weather-induced behaviors) between the two models. McNemar's test is well-established for evaluating paired nominal data where each instance is mutually exclusive to two classification outcomes [29], making it particularly suitable for comparing two models on the same set of samples.

Specificity is a key metric in this context, as a high specificity implies a reduced likelihood[1] of misclassifying weather-induced vessel movement as anomalous, thereby minimizing false positives. Table 3.1 provides the format for reporting the statistical significance of baseline to weather-enhanced model comparisons.

Reporting the statistical significance of results aligns with recommendations in the maritime anomaly detection literature [36] especially due to the lack of labeled benchmark datasets [33]. Omitting this type of reporting reduces the generalization of results [18, 27]. We emphasize the its importance for this work, especially in relation to the representation of the test set used for evaluation (see: 3.3.3). Differences between train-test set representations can cause bias in relation to the results, a potential mitigation strategy was adopted in [12] by conducting significance tests at the 1% level instead of the more conventional 5% level. Due to similar validity threads in this thesis, we will adopt the same mitigation strategy.

In this thesis the key metric is specificity (the proportion of true negatives among all actual normal cases) for reducing false alarms, especially in the presence of confounding contextual factors such as weather [5, 43]. By applying the McNemar test to paired outcomes (specificity baseline vs. specificity weather enhanced model) we can formally test whether the observed reduction in false positives to meteorological data augmentation—is statistically significant, thereby rigorously addressing the primary research question concerning the distinguishability between artificial anomalous behavior and weather-induced behaviors capacity of the proposed model.

Table 3.1: Output Format for Reporting Statistic Significance using McNemar's Statistical test

d, *ρ*,	**True Negatives**	**False Positives**
Weather Model	a	b
Baseline Model	c	d
McNemar statistic, *p-value*		

[1] In this context, a "reduced likelihood" of misclassification means that when specificity increases, there is a lower probability that the model will incorrectly identify weather-induced vessel movements as artificial anomalies, thereby decreasing false alarms.

3.5.4 Metrics per Research Question

The main research question of this thesis is as follows:

> *How can the integration of meteorological information augment the effectiveness of deep learning approaches in distinguishing weather-induced vessel movements from artificial maritime trajectory anomalies?*

Building on research that has shown contextual verification can reduce false alarms [31], we chose to evaluate our meteorological integration approach through positive performance metrics, specifically measuring the system's ability to correctly identify weather-induced movements (specificity) rather than focusing solely on false alarm reduction. Table 3.2 provides an over per research question of the key metric used for reporting in the results section. In the paragraphs that follow, the rationale for the metrics associated per research question is further elaborated on.

Table 3.2: Summary of Key Metrics per Research Question

Research Question	Key Metric(s)
RQ1	ROC AUC
RQ2	ROC AUC, Sensitivity, Specificity
RQ3	ROC AUC, Sensitivity, Specificity, McNemar's Test (Specificity)
RQ4	ROC AUC, Sensitivity, Specificity

Optimal Model Configuration

To provide an answer to **RQ1**, we utilize the Area Under the Receiver Operating Characteristic Curve (ROC AUC), sensitivity, and specificity. This experiment determines the configuration of the model's hyperparameters used in the research questions that follow RQ1. ROC AUC provides a balanced perspective on model distinguishability between both normal and anomalous trajectories, preventing the outcome of this experiment from being positively biased towards the detection of either normal trajectories or anomalous trajectories. In addition, we observe sensitivity (true positive rate) in relation to specificity (true negative rate) to assess the optimal configuration of the number of Gaussian Mixture Model components in relation to RQ1.

Impact of Meteorological Variables

Based on our findings for RQ1, we set up an experiment comparing the baseline model with our weather-optimized model. Preliminary statistical analysis of weather conditions over trajectories showed variations in vessel behavior depending on weather conditions. For that reason, we believe weather conditions can help explaining vessel behaviors. We examine whether deep learning models can distinguish between vessel behavior caused by meteorological conditions and artificially created anomalous behavior. We compare model predictions for our baseline and weather-enhanced model by reporting on specificity, sensitivity, and ROC AUC.

Impact of Meteorological Data Integration

In our methodology, we define three types of synthetic anomalies - shift deviation, abnormal heading, and abnormal speeding - for each type, we investigate if our model distinguishes better between weather-induced behavior and artificially created anomalous behavior. We provide an answer to RQ3 by comparing changes relative to our baseline configuration and report on specificity, sensitivity and ROC AUC. In addition, for this experimental setup we report the statistical significance of the observed differences for the specificity metric. We compare true negative predictions with false positive predictions for the same test configuration using the McNemar test [29].

Weather Severity

Finally, we assume that the severity of meteorological conditions influence vessel behavioral patterns [1]. For that reason, we compare model predictions by grouping trajectories based on the severity of the prevailing meteorological conditions. We report specificity, sensitivity and ROC AUC to provide an answer to our research questions.

Chapter 4

Results

This chapter presents the experimental evaluation of the weather-enhanced model compared to the baseline model. We systematically address each research question through a comprehensive analysis attributed by statistical significance tests. This chapter is structured as follows. Section 4.1 explores RQ1 by evaluating ROC AUC, sensitivity, and specificity metrics of the enhanced model; while varying the number of Gaussian Mixture Model Components (C). Section 4.2 addresses RQ2 by comparing baseline and meteorological data-enhanced models across sensitivity and specificity metrics; Section 4.3 examines RQ3 through detailed anomaly type-specific analysis of shift deviation, abnormal heading, and abnormal speeding by examining ROC AUC, sensitivity, and specificity evaluation metrics; and Section 4.4 investigates RQ4 by evaluating model resiliency under varying weather severity conditions; Experiments utilize sensitivity and specificity as the primary evaluation metrics, supplemented by ROC Area Under the Curve to provide balanced performance assessment with respect to normal and anomalous vessel trajectories.

This research leverages the experimental setup as provided in [18], we report on test configuration parameters - d, r, and ρ - used for the test configurations to compute evaluation metrics. As mentioned in Chapter 3, for abnormal speeding anomalies, the parameter d is not applicable. To avoid ambiguous reporting on test configurations, we indicate the inclusion of abnormal speeding anomalies by mentioning $d = 0$.

- **RQ1:** How does the number of Gaussian Mixture Model (GMM) components contribute to distinguishing weather-induced behavioral patterns from artificial anomalies?
- **RQ2:** To what extent does incorporating meteorological variables influence the ability to distinguish between weather-induced behavioral patterns from artificial maritime trajectory anomalies?
- **RQ3:** How does weather integration affect the model's ability to correctly identify different types of artificial anomalies while avoiding false positives from weather-induced behavioral patterns?
- **RQ4:** How does the integration of meteorological data affect model performance in distinguishing weather-induced behavioral patterns from artificial maritime trajectory anomalies across varying weather severity conditions?

4.1 RQ1: Impact of Gaussian Mixture Components

This experiment evaluates trained models, varying the number of Gaussian components, to assess both training convergence and the model's ability to distinguish between artificially created anomalous behavior and meteorologically influenced behavior. This provides empirical guidance for determining the optimal model configuration when incorporating weather data. The results of

this experiment provide a foundation for the experimental setup of subsequent experiments.

The clustering component of the model requires determining the optimal number of Gaussian components for the GMM estimation network. While in [18] the optimal number of components was found to be 20 as optimal for their original dataset, the present study incorporates additional weather data features, thereby increasing the dimensionality of the input space from which clusters are formed. This dimensional expansion may alter the optimal clustering structure and necessitate a different number of components to capture the underlying trajectory patterns effectively.

4.1.1 Experimental Setup

The number of Gaussian components affects the model's ability to represent the diversity of normal trajectory behaviors. Insufficient components may result in oversimplified representations that fail to capture trajectory variability, while excessive components may lead to overfitting and reduced anomaly detection performance [18]. Therefore, systematic evaluation of component numbers (C) is essential to optimize model performance for the enhanced feature space.

For this experiment, the model architecture of the clustering component was adjusted according to the number of Gaussian components to be evaluated. The following explains this in more detail:

$$\text{hidden dimension GMM} = \left\lfloor \frac{(\text{C} - 20)}{20} \times 32 + 32 \right\rfloor$$

With C representing the number of Gaussian components. In preliminary experiments, the optimal value of C was adopted from [18], and meteorological data was excluded. Based on this configuration, we found that with a hidden GMM dimension of 32, the model's learning characteristics remained stable. For that reason, 32 was chosen as a reference point for training with meteorological data. Table 4.1 shows the values tested for C associated with the adjusted size of the hidden dimension of the GMM. Finally, to ensure fair comparison, all feature-enhanced models are trained for 100 epochs.

Table 4.1: The number of Gaussian Components (C) and the adjusted hidden dimension layer size of the Gaussian Mixture Model

Number of Gaussian Components (C)	Hidden Dimensions of GMM
1	1
5	8
10	16
20	32
30	48
40	64
50	80
100	160

4.1.2 Training Dynamics

Figure 4.1 shows the learning curves of the models trained with varying numbers of components. Models trained with $C \in 1, 5, 10, 20, 30, 40, 50$ start to display converging patterns around 50 training epochs, which validates the choice to train models for 100 epochs for this experiment. Important to mention here is that instead of measuring the training loss per epoch, the training loss was measured by disabling regularization mechanics and calculating training loss per epoch

with the model in evaluation mode (Dashed lines in Figure 4.1). This effectively excludes the implicit effect of dropout on the training loss score [48]. Due to small batch sizes, dropout introduces implicit noise to the gradient updates.

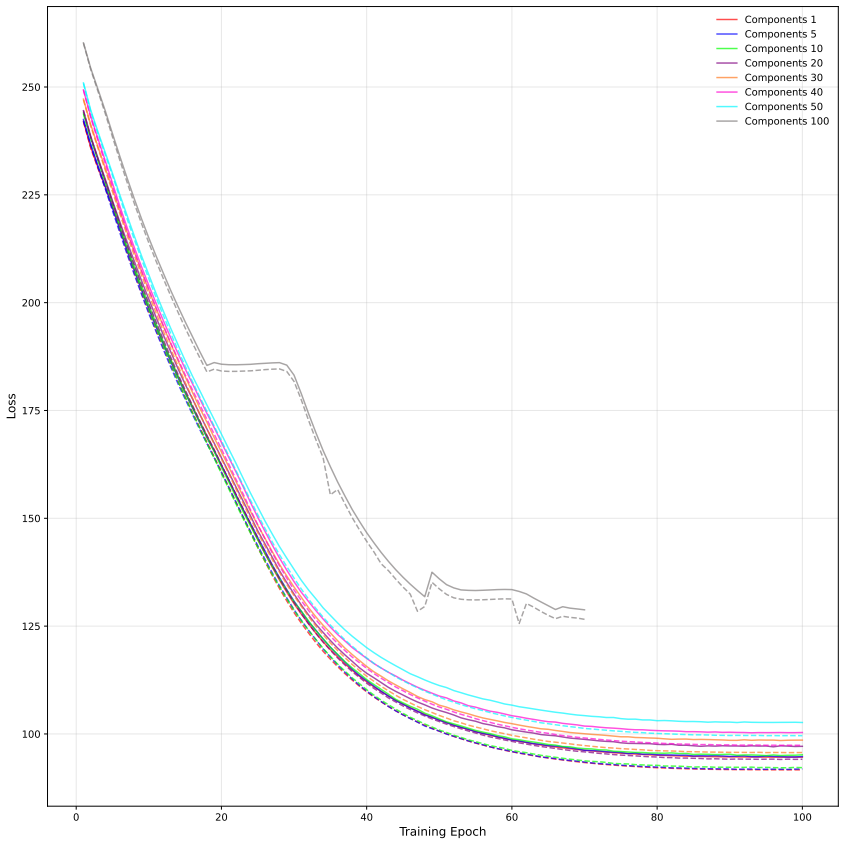

Figure 4.1: Loss Curves for Models Trained with increasing values for the number of Gaussian Components. Validation Loss (solid) and True Loss (dashed). During the training run for $C = 50$, 2 training epochs resulted in numerical instability, resulting in large negative values for validation loss. For the creation of this plot, the values for these training epochs were removed from the data.

[48] tested the effect of implicit noise on training loss due to dropout regularization by sampling (k) dropout masks and measuring the variation of training loss. This experiment demonstrated that with larger values of k an additional regularization effect was introduced beyond the explicit modification of the expected training objective. This implicit regularization arises from the variance in gradient estimates caused by random dropout sampling. [48] demonstrate that this gradient noise encourages convergence to flatter minima that generalize better, and that removing this noise, as done by averaging over multiple dropout samples (k) can degrade performance. The experiment as demonstrated by [48] relates to this research in the following way: during training we noticed the average training loss scores reported higher than the evaluation loss scores. As found in [48], a potential explanation for this observation might be related to gradient estimates derived from training loss scores, including regularization factors. For that reason, we report training loss (see: dashed curve in Figure 4.1) by measuring loss (see Section: 3.4.5) in evaluation mode for the training set, isolating the explicit learning dynamics from this implicit regularization effect, allowing for a clearer assessment of the model's actual fitting behavior without the confounding influence of dropout's stochastic gradient perturbations. However, gradient updates are based on

regularization included training loss scores.
Figure 4.2 provides a high-resolution perspective on the same learning curves shown in Figure 4.1, showing that by measuring training loss (dashed line) by disabling dropout, the validation loss is higher than the training loss (dashed line).

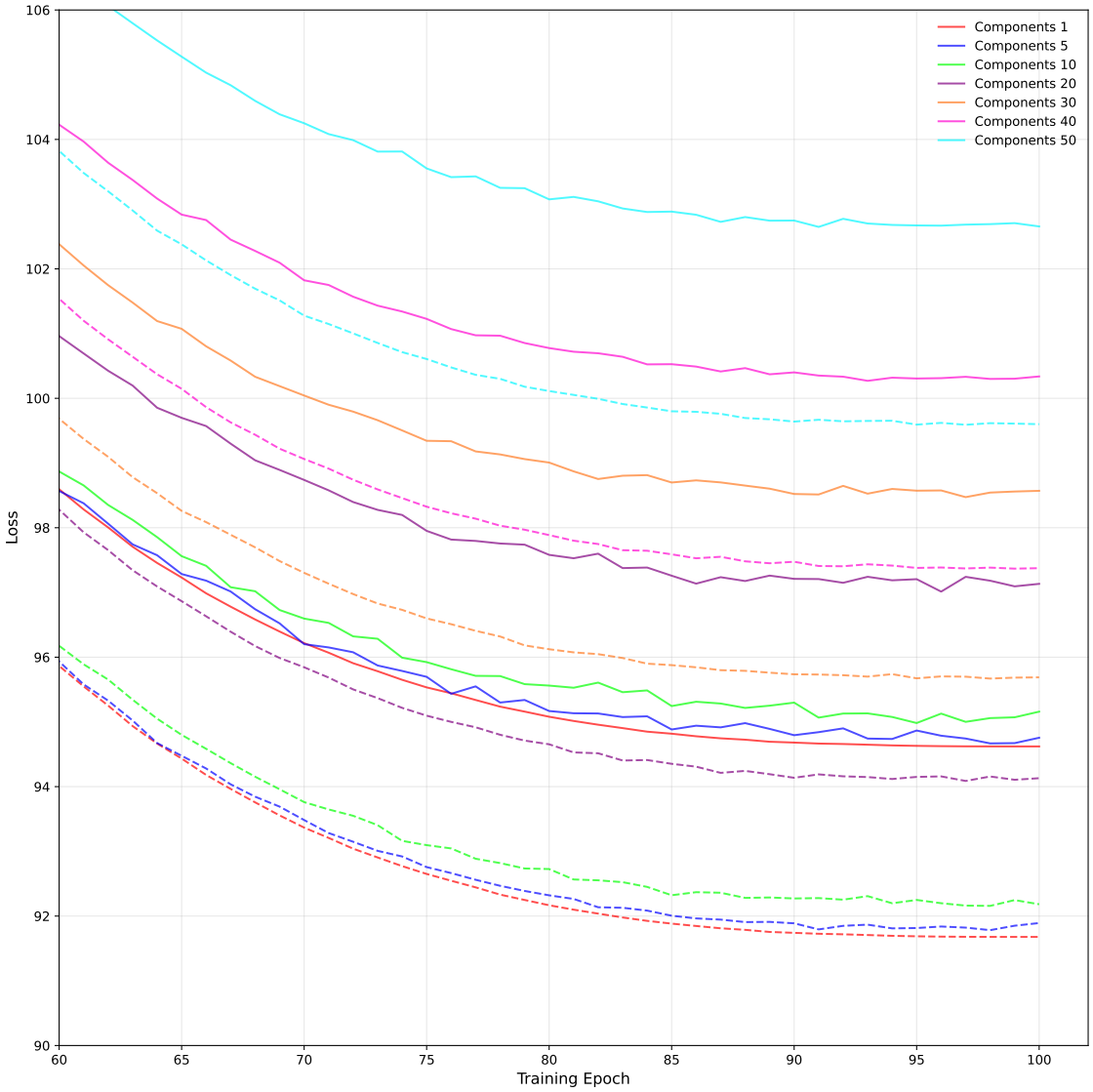

Figure 4.2: Zoomed Loss Curves for Models Trained with increasing values for the number of Gaussian Components. Validation Loss (solid) and True Loss (dashed)

During model training, minimization of the composite loss function $\mathcal{L}_{\text{STAD}}$ is performed by optimizing the set of learnable model parameters, as described in Section 3.4.5. Figure 4.3 presents the average per-epoch values for the energy term (E_{GMM}) across different values of the GMM component count parameter $C \in \{10, 20, 30, 40\}$. The results demonstrate that the average energy value decreases as C increases. Substantial variability in energy values is observed between epochs across all tested values of C.

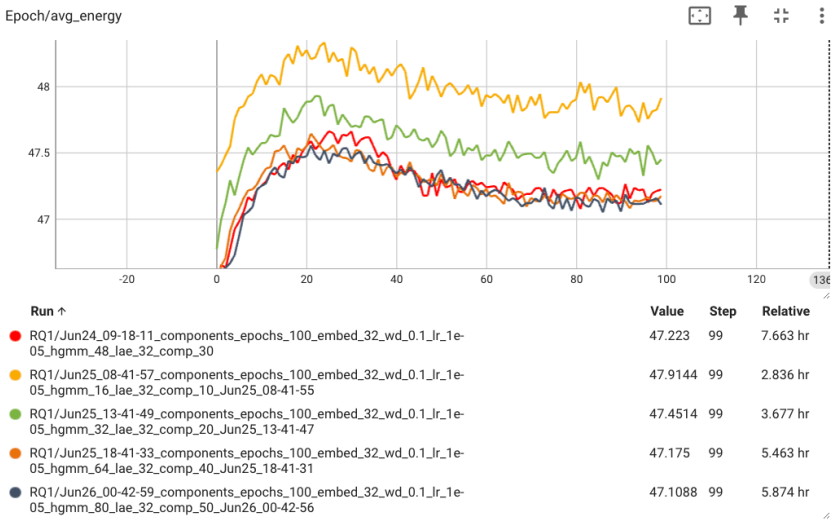

Figure 4.3: Average Epoch Training Energy for $C \in \{10, 20, 30, 40\}$

4.1.3 Performance Comparison

Figure 4.4 presents the evaluation metrics—ROC AUC, sensitivity, and specificity—as a function of the number of Gaussian components (C). To ensure balanced representation across anomaly types, each test configuration is held constant at the following parameter settings: spatial deviation $d \in \{0, 1\}$, anomalous trajectory ratio $r = 0.1$, and proportion of perturbed points per trajectory $\rho = 0.05$. For each value of C, one test configuration is evaluated per anomaly type, maintaining a fixed anomaly-to-normal ratio of 0.1. This configuration is designed to maximally challenge the model's discriminative capability by making the artificial anomalies particularly difficult to detect. Specifically, the minimal values for d and ρ correspond to less pronounced anomaly patterns, thereby increasing the task difficulty. As noted in [18], greater values of d accentuate the anomaly pattern, increasing the reconstruction error during offset reconstruction; conversely, small d values reduce anomaly distinctiveness. For each anomaly type per value of C, the resulting metric scores are averaged and displayed in Figure 4.4. The results in Figure 4.4, detailed in Table 4.2, indicate that as the number of GMM components increases, the ROC AUC score remains relatively stable. Sensitivity decreases with an increasing number of components, whereas specificity tends to increase, reflecting a trade-off between these two metrics as model complexity grows.

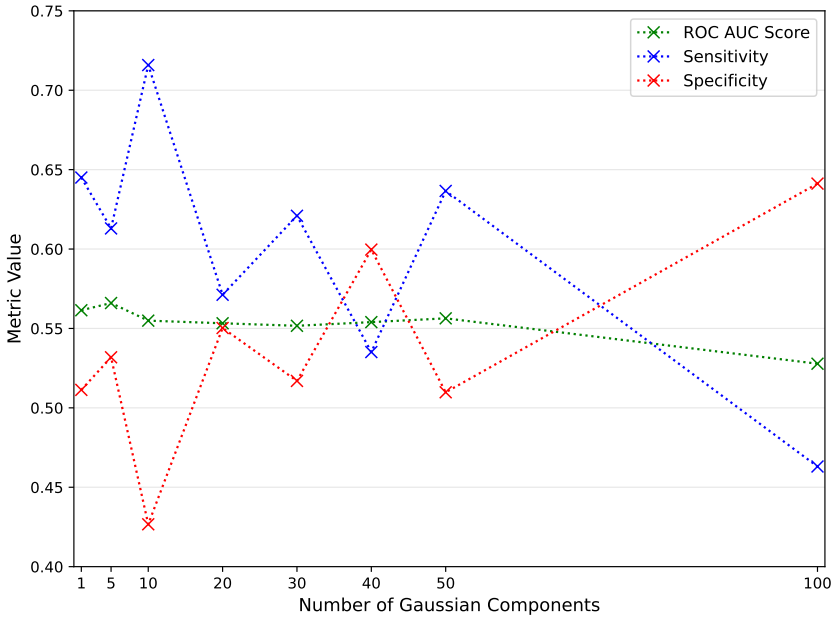

Figure 4.4: Evaluation metrics (ROC AUC, sensitivity, and specificity) as a function of the number of Gaussian mixture model components (C), evaluated under test parameters $d \in \{0, 1\}$, $r = 0.1$, and $\rho = 0.05$.

Table 4.2: Evaluation Metrics for Varying Number of Gaussian Mixture Model Components (C).

C	ROC AUC Score	Sensitivity	Specificity
1	0.5615	0.6450	0.5113
5	**0.5660**	0.6130	0.5318
10	0.5549	**0.7159**	0.4267
20	0.5533	0.5712	0.5503
30	0.5517	0.6209	0.5170
40	0.5540	0.5351	0.5997
50	0.5564	0.6366	0.5098
100	0.5278	0.4631	**0.6412**

4.2 RQ2: Impact of Meteorological Variables

This experiment assesses the impact of incorporating meteorological variables into trajectory anomaly detection models. To isolate the effect of meteorological data, we compared two model configurations: a baseline model utilizing solely AIS-derived features, and an enhanced model with additional meteorological statistics incorporated into its GMM component.

4.2.1 Experimental Setup

The baseline experiment validates the effectiveness of integrating metocean features by comparing two identical models - one following the original architecture without weather data, and an enhanced version that incorporates meteorological statistics within the GMM component. Both models were trained for 250 epochs. Both models share identical architectures and were trained for 250 epochs with consistent hyperparameters, detailed in Appendix A. The number of Gaussian

components (C) was set to 20 for the baseline model, and to 30 for the weather-enhanced model, based on optimal configuration findings from Section: 4.1. We scaled the hidden dimension of the Gaussian Mixture model according to the values noted in Table 4.1.

To further examine the role of C in relation to meteorological features, we attempted to train the baseline model with $C = 30$ (without meteorological information), but this configuration exhibited numerical instability and failed to converge. We conducted different experiments for $d \in \{0, 1\}$ and $d \in \{0, 2\}$ because of its significant effect on the overall discriminative ability of the model [18]. Table 4.3 summarizes the experimental setups and their interpretations.

Table 4.3: Overview of Experimental Setups, Parameters, and Interpretations.

Setup	d	r	ρ	**Interpretation**
A	0, 1	0.1	0.05	Minimum spatial deviation; minimal proportion of anomalous points per perturbed trajectory; maximum ratio of anomalous trajectories to emphasize the positive class.
B	0, 2	0.1	0.05	Differs from Setup A by applying the maximal spatial deviation value (d)

To ensure a fair comparison, both models were trained for 250 epochs, and the reported results reflect the model weights that achieved the lowest validation loss during training. The total training duration was 8.7 hours for the baseline model and 18.5 hours for the weather-enhanced model, with the increased time attributable to higher model complexity. The remainder of this section reports the sensitivity and classification performance for each experimental configuration.

4.2.2 Training Dynamics

Figure 4.5 presents the training loss curves for both the baseline and weather-enhanced models evaluated across 250 training epochs. Both models exhibit a rapid decrease in training loss during the initial 60 epochs, after which the curves tend to plateau, indicating stabilization and convergence of the training process.

A consistent difference in the absolute training loss is observed: the weather-enhanced model maintains higher loss values across all epochs compared to the baseline model. The difference between the two models is established within the initial phase of training and persists until training completion. Both models attain stable loss values after approximately 125 epochs, without exhibiting signs of divergence or late-stage instability.

The training objective, as defined in Section 3.4.5, comprises the sum of several components: transformer encoder loss ($\mathcal{L}_{\text{TE}}$), Gaussian Mixture Model energy term (E_{GMM}), autoencoder reconstruction loss (d_{AE}), and a penalty term on the Gaussian components (p_{GMM}), each weighted by their corresponding hyperparameters. The overall loss magnitude reported in Figure 4.5 reflects the aggregate of these terms under each architecture configuration.

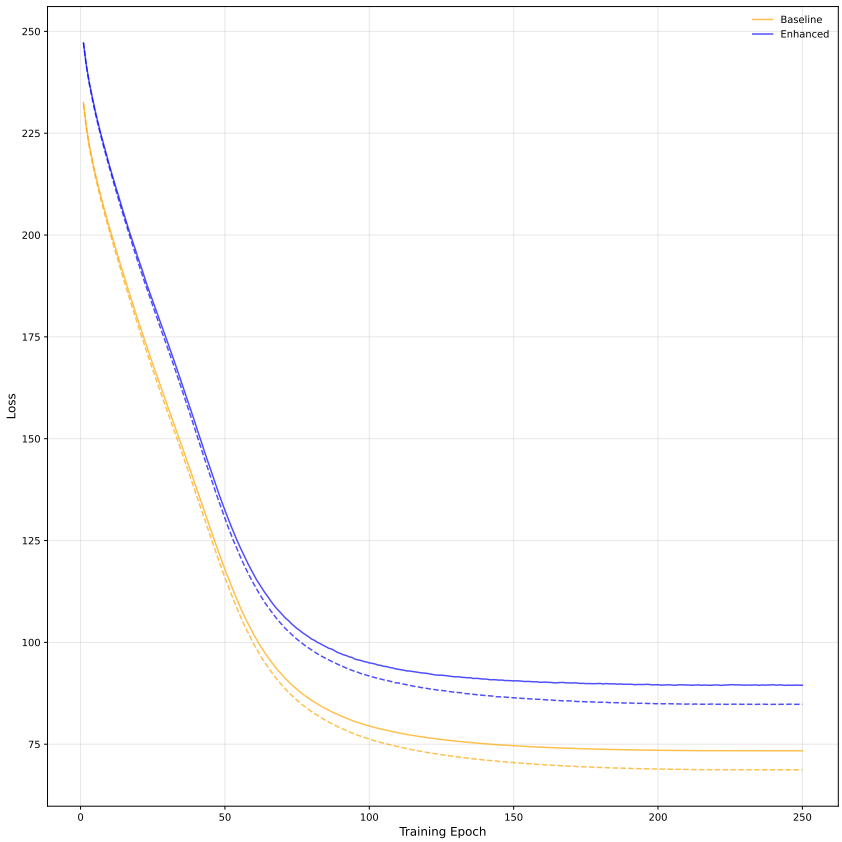

Figure 4.5: Training loss curves for the baseline (blue) and meteorological data-enhanced (orange) models over 250 training epochs. The baseline model achieves consistently lower loss values, while the weather-enhanced model exhibits a systematically higher loss trajectory, both converging after approximately 60 epochs.

The relative magnitude of the training losses is reported here for completeness and to ensure transparency about the learning dynamics of both architectures. Subsequent sections evaluate model generalization using ROC AUC sensitivity and specificity to determine the practical effects of these observed training behaviors.

4.2.3 Performance Comparison

Model performance was evaluated under both experimental setups (A and B) by reporting specificity, sensitivity, and ROC AUC scores. Detailed results for both baseline and weather-enhanced models are summarized in Tables4.4 and4.5.

Performance Under Setup A

Table 4.4: Comparison of Baseline and Weather-Enhanced Model Performance for setup A ($d \in \{0, 1\}$, $r = 0.1$, and $\rho = 0.05$)

Metric	Baseline	Weather-Enhanced	Difference
Specificity	0.5411	0.5783	0.0372
Sensitivity	0.6290	0.6014	-0.0276
ROC AUC Score	0.5947	0.5851	-0.0096

For Setup A, the weather-enhanced model achieved a higher specificity (0.5783) compared to the baseline (0.5411), reflecting an increase of 0.0372. This indicates improved ability to correctly identify normal (weather-induced) trajectories. However, the sensitivity of the weather-enhanced model (0.6014) was slightly lower than that of the baseline (0.6290), representing a decrease of 0.0276. The ROC AUC score for the weather-enhanced model (0.5851) was lower than the baseline (0.5947), with a difference of -0.0096.

Performance Under Setup B

Table 4.5: Performance Comparison Between Baseline and Weather-Enhanced Models for Test Configuration B ($d \in \{0, 2\}$, $r = 0.1$, and $\rho = 0.05$)

Metric	Baseline	Weather-Enhanced	Difference
Specificity	0.4522	0.5096	0.0574
Sensitivity	0.7654	0.7037	-0.0617
ROC AUC Score	0.6228	0.6067	-0.0161

Under Setup B, we increased the spatial deviation of the shift deviation and abnormal heading anomalies (d). The weather-enhanced model again demonstrated higher specificity (0.5096 versus 0.4522), with an improvement of 0.0574 over the baseline. Similar to Setup A, this gain in specificity was accompanied by a reduction in sensitivity (0.7037 for the weather-enhanced model compared to 0.7654 for the baseline, a difference of -0.0617). The ROC AUC score for the weather-enhanced model (0.6067) was also slightly lower compared to the baseline (0.6228), with a difference of -0.0161.

Summary of Observations

Across both experimental setups, integrating meteorological variables led to a consistent increase in model specificity, indicating that the weather-enhanced model is more effective at correctly distinguishing weather-induced vessel behavior from artificial anomalies. However, this improvement in specificity is offset by a reduction in sensitivity and a moderate decline in overall discriminative capacity as measured by the ROC AUC score.

4.3 RQ3: Impact of Meteorological Data Integration

To address RQ3, we investigate whether integration of meteorological data affects the model's capability to distinguish between weather-influenced vessel behaviors and artificial anomalies, considering different types of artificial anomalies. The following results present a detailed comparison for each anomaly type.

4.3.1 Experimental Setup

The experimental setup for RQ3 follows that described in Section 4.2, utilizing the baseline (AIS-only) and weather-enhanced (AIS + meteorological variables) models. For this investigation, models were evaluated according to anomaly type, with experimental parameters set as $d \in \{0, 1\}$, $r = 0.1$, and $\rho \in \{0.05, 0.1, 0.2\}$. Performance metrics were averaged across test sets with different values of ρ for each anomaly type and configuration.

4.3.2 Performance Comparison

Table 4.6: Weather-Enhanced Model minus Baseline Comparison by Anomaly Type ($d \in \{0, 1\}$, $r = 0.1$, and $\rho = 0.05$).

Anomaly	ROC AUC	Sensitivity	Specificity
Shift deviation	-0.0044	0.1375	-0.1132
Abnormal heading	-0.0082	-0.1585	0.1676
Abnormal speeding	-0.0161	-0.0617	0.0574

Table 4.6 provides a comparison of the average change in our validation metrics for each anomaly type when meteorological data is integrated. We observe the most significant increases in the model's ability to classify normal trajectories for abnormal heading and abnormal speeding anomalies. We also note the opposite effect for shift deviation anomalies, meaning that the enhanced model performs better when detecting artificial anomalies than detecting normal behavior. An overview of the results by anomaly type is presented in the following sections.

Anomalies of Type 1: Shift Deviation

Table 4.7 details the differences in ROC AUC, sensitivity, and specificity between the weather-enhanced and baseline models for shift deviation anomalies. Results indicate that the weather-enhanced model's highest relative improvement in sensitivity occurs when the spatial deviation parameter d is larger, although the impact of the number of perturbed trajectory points ρ is noteworthy since it generally shows specificity improvements when $\rho \in \{0.1\}$ but specificity scores generally decline when $\rho \in \{0.05, 0.2\}$. However, specificity consistently decreased in several configurations, suggesting that while the weather-enhanced model was better at detecting artificial anomalies under some conditions, it also misclassified a slightly higher proportion of weather-induced normal behavior as anomalous in others. Average ROC AUC differences were near zero, indicating little overall change in the overall discrimination capability for shift deviation anomalies.

Table 4.7: Difference for ROC AUC, Sensitivity, and Specificity between Weather Enhanced Model and Baseline Model for Shift Deviation Anomalies

d	r	ρ	ROC AUC	Sensitivity	Specificity
1	0.05	0.05	0.0114	0.0513	-0.0426
		0.1	0.0005	-0.0769	0.0749
		0.2	-0.0110	0.0000	-0.0397
1	0.1	0.05	-0.0044	0.1375	-0.1132
		0.1	0.0080	-0.0864	0.0833
		0.2	-0.0044	0.0633	-0.0640
2	0.05	0.05	-0.0079	-0.0750	0.0530
		0.1	0.0241	-0.0750	0.1475
		0.2	0.0118	0.0513	0.0155

Anomalies of Type 2: Abnormal Heading

Table 4.8 summarizes results for abnormal heading anomalies. Incorporation of meteorological data produced consistently positive (though modest) gains in specificity, indicating improved identification of weather-induced normal behaviors. Specificity changes varied across conditions, sometimes showing notable increases (e.g., $d \in \{0, 1\}$, $r = 0.1$, and $\rho = 0.05$), while in other cases small improvements were observed. These results suggest that weather integration tends to enhance the model's ability when predicting normal trajectories as opposed to predicting artificial anomalies of type for abnormal heading, although it should be noted that larger increases generally come

at the expense of detecting true artificial anomalies (sensitivity decrease). ROC AUC differences were generally small, showing that overall discrimination capacity remained stable.

Table 4.8: Difference for ROC AUC, Sensitivity, and Specificity between Weather Enhanced Model and Baseline Model for Type 2 Anomalies

d	r	ρ	**ROC AUC**	**Sensitivity**	**Specificity**
1	0.05	0.05	0.0021	-0.1220	0.1063
		0.1	-0.0181	-0.0244	0.0166
		0.2	0.0270	-0.0244	0.0192
1	0.1	0.05	-0.0082	-0.1585	0.1676
		0.1	0.0042	0.0000	0.0324
		0.2	0.0105	0.1951	-0.1622
2	0.05	0.05	0.0020	0.0488	-0.0487
		0.1	0.0149	-0.0732	0.1140
		0.2	-0.0173	-0.0732	0.0282

Anomaly Type 3: Abnormal Speeding

Table 4.9 presents the results for abnormal speeding. Here, the impact of weather data integration was mixed. For lower values ρ, the weather-enhanced model demonstrated decreased ROC AUC and sensitivity, largely offset by increased specificity.

This pattern indicates the model became more conservative, with a tendency to misclassify artificially created anomalous trajectories as normal. This generally attributes a decreased ability to classify artificially created anomalies due to lower sensitivity scores. For lower ratios of perturbed points per artificial anomalous trajectory (ρ), the weather-enhanced model sometimes outperformed the baseline in specificity, though occasionally at the expense of ROC AUC.

Table 4.9: Difference for ROC AUC, Sensitivity, and Specificity between Weather Enhanced Model and Baseline Model for Type 3 Anomalies

r	ρ	**ROC AUC**	**Sensitivity**	**Specificity**
0.05	0.05	-0.0034	-0.5750	0.6067
	0.1	-0.0177	-0.1750	0.1669
	0.2	0.0101	0.2000	-0.1591
0.1	0.05	-0.0161	-0.0617	0.0574
	0.1	-0.0130	0.0000	-0.0260
	0.2	0.0191	-0.0123	0.0342

In summary, the results show that weather-enhanced models score higher for specificity, especially for heading and speed anomalies

4.3.3 Summary Performance per Anomaly Type

Across all anomaly types, results show that integration of meteorological data generally leads to moderate improvements in specificity (true negative rate), meaning the weather-enhanced model more accurately classified weather-induced vessel deviations as normal. However, this was frequently accompanied by reduced sensitivity for certain anomaly injection configurations, indicating a higher rate of missed artificial anomalies compared to the baseline model. ROC AUC changes were typically small, pointing to a limited overall shift in the overall models' capacity to discriminate between classes.

4.3.4 Statistical Tests

Following the methodology outlined in Chapter 3, we assess the statistical significance of observed differences in specificity between the baseline and weather-enhanced model per anomaly type. We report on significance by employing the McNemar test[29]. The McNemar test is applied to contingency tables constructed from the true negative and false positive predictions for each anomaly type under identical test configurations. Tables 4.10, 4.11 and 4.12 show the results per anomaly type.

Table 4.10: McNemar's test results ($d = 1$, $\rho = 0.05$ and $r = 0.1$) for detecting shift deviation anomalies.

Model	**True Negatives**	**False Positives**
Weather-Enhanced	437	305
Baseline	521	221
McNemar statistic: 221.0, *p-value*: 2.72×10^{-17}		

Table 4.11: McNemar's test results ($d = 1$, $\rho = 0.05$ and $r = 0.1$) for detecting abnormal heading anomalies.

Model	**True Negatives**	**False Positives**
Weather-Enhanced	471	269
Baseline Model	347	393
McNemar statistic: 393.0, *p-value*: 8.77×10^{-3}		

Table 4.12: McNemar's test results ($\rho = 0.05$ and $r = 0.1$) for abnormal speeding anomalies.

Model	**True Negatives**	**False Positives**
Weather-Enhanced	373	359
Baseline Model	331	401
McNemar statistic: 373.0, *p-value*: 0.332		

The results show that statistically significant differences for the 1% level in specificity were observed for the shift deviation and abnormal heading anomaly types (p-values: 2.72×10^{-17} and 8.77×10^{-3}, respectively). No statistically significant difference in specificity was observed for the abnormal speeding anomaly type (p-value: 0.332).

4.4 RQ4: Model Performance Under Varying Weather Conditions and Environmental Scenarios

To address RQ4, we evaluated the weather-enhanced model's ability to distinguish weather-induced movements from artificial maritime trajectory anomalies under varying weather severity conditions. This analysis examines whether meteorological integration supports stable discriminative classification capabilities across above and below average severity of weather conditions.

4.4.1 Experimental Setup

To investigate the relation between the ability of the classification model to distinguish weather-induced behavior from artificial anomalies under differing meteorological conditions, test sets were

configured with parameters $d \in \{0, 1\}$, $r = 0.1$, and $\rho = 0.05$. This corresponds to anomalous trajectories with the smallest spatial deviation setting and the smallest ratio of perturbed points to the total number of points within the same trajectory. The trajectories in the test sets were partitioned into two groups according to the median value of the sum of the maximum wave height and the maximum wind speed recorded for each trajectory. The rationale for this grouping is consistent with the findings of [24, 1], who reported on wind speed and wave height as being the principal environmental variables influencing vessel behavior. It is noted that additional variables with potential influence on vessel behavior are beyond the scope of both this thesis and the studies by [24, 1]. The distribution for the sum of maximum wave height and wind speed per trajectory is presented in Figure 4.6. Evaluation metrics were independently calculated for each weather severity group within the same test set, facilitating an evaluation of detection performance as a function of environmental severity.

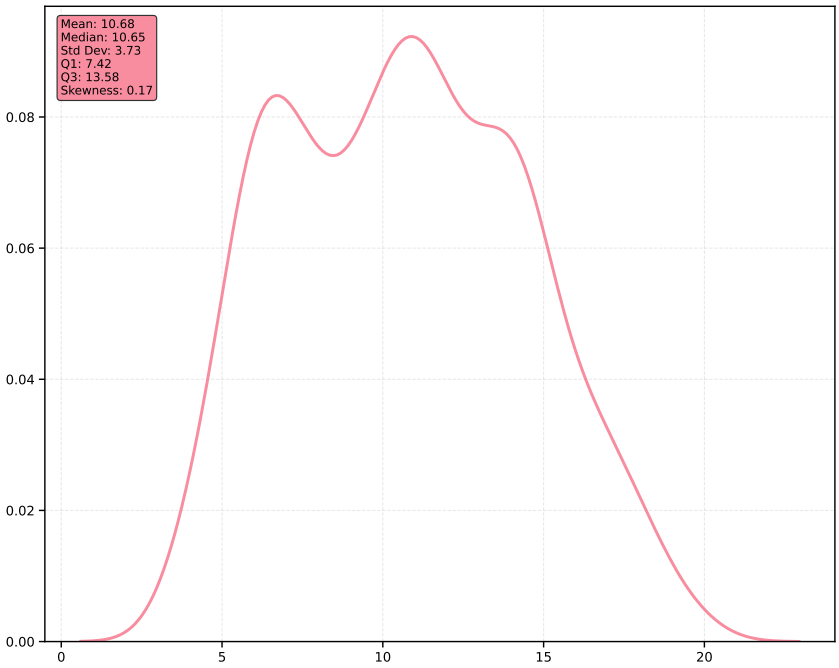

Figure 4.6: Kernel Density Estimation Plot for the Sum of the Maximum Values for Significant Wave Height (m) and Wind Speed (m/s) of Each Trajectory in the Test Set.

The distribution of the sum of maximum wave height and wind speed exhibits significant positive skewness (skewness coefficient = 0.17), as confirmed by a skewness test ($p = 0.0485$, $\alpha = 0.05$).

Thus, we partitioned trajectories based on the median value to create two distinct groups: one representing typical operational conditions and another capturing trajectories exposed to more challenging environmental circumstances that are likely to induce significant deviations in vessel behavior.

Figure 4.7 depicts the relationship between wave height and wind speed across all AIS data points in the test set. The Pearson correlation coefficient ($r = 0.658$) indicates a moderate linear association between these two environmental variables. The Pearson correlation coefficient between wave height and wind speed ($r = 0.658$, $p < 0.001$) indicates a statistically significant moderate positive linear association between these environmental variables.

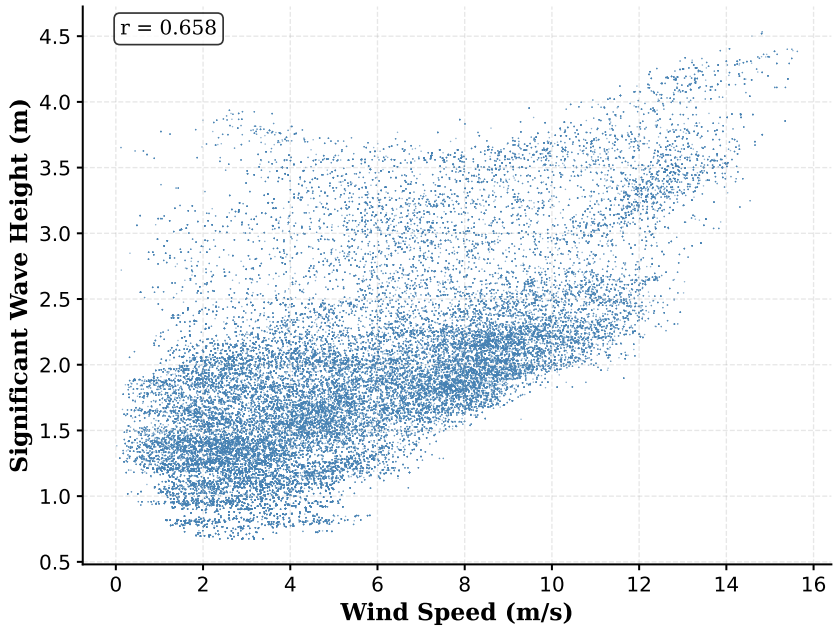

Figure 4.7: The relationship between Significant Wave Height (m) and Wind Speed (m/s) over the collection of individual AIS data points in the test set. The top left corner of the plot shows the Pearson correlation coefficient (r) for the two variables.

4.4.2 Performance Comparison

Model performance was evaluated separately for the groups representing higher and lower environmental severity by calculating ROC AUC, sensitivity, and specificity metrics. The differential values (high severity minus low severity) for the weather-enhanced model are presented in Table 4.13 for each anomaly type.

Table 4.13: Comparison of Metrics over Trajectories grouped by severity of Weather Conditions. Scores are computed for the test sets that correspond to the following test configuration parameters: $d \in \{0, 1\}$, $r = 0.1$, and $\rho = 0.05$

Anomaly	Difference Between Group (High-Low)		
	ROC AUC	Sensitivity	Specificity
Shift deviation	0.0637	-0.0545	0.1044
Abnormal heading	0.1099	0.2571	-0.0384
Abnormal speeding	0.0782	-0.0105	0.1545
Mean	0.0839	0.0640	0.0735

The results indicate that, on average, the weather-enhanced model achieved slightly higher ROC AUC, sensitivity, and specificity scores under more severe environmental conditions. Across the anomaly types evaluated, the differences in ROC AUC ranged from 0.0637 to 0.1099, with the largest increase observed for the abnormal heading anomaly. Sensitivity varied more substantially, with abnormal heading detection exhibiting a notable improvement (0.2571), while the shift deviation and abnormal speeding anomalies showed reduced sensitivity (0.0545 and -0.0105, respectively) in high-severity conditions. Specificity generally improved under higher severity, particularly for shift deviation (0.1044) and abnormal speeding (0.1545). The mean differences suggest a consistent, although moderate, enhancement in model discrimination capacity as envir-

onmental severity increases.

These findings suggest that the weather-enhanced model maintains, and in certain cases improves, its ability to distinguish between weather-induced and artificial anomalies under more severe meteorological conditions, primarily reflected in improved ROC AUC and specificity scores. However, sensitivity to certain anomaly types may decrease, indicating areas where further refinement of the model may be warranted for operational deployment in adverse environments.

4.5 Summary of Results

4.5.1 Impact of Gaussian Mixture Components

Systematic experiments varying the number of Gaussian Mixture Model (GMM) components (C) confirmed that contextual model performance is sensitive to this parameter, particularly since additional meteorological features increase input dimensionality. In conformance [18] we observed that, models with too few components oversimplified trajectory representations, however we also that high values for (C) occasionally led to numerical instability. The optimal number of components balanced these effects, supporting model convergence and effective discrimination between meteorologically influenced and artificially anomalous vessel trajectories.

4.5.2 Effect of Meteorological Data Integration

The comparative evaluation revealed that incorporating meteorological variables consistently improved model specificity across various experimental setups. For Setup A ($d \in \{0, 1\}$), specificity increased from 0.5411 (baseline) to 0.5783 (weather-enhanced), whereas for Setup B ($d \in \{0, 2\}$), specificity increased from 0.4522 (baseline) to 0.5096 (weather-enhanced). However, we note that this improvement was offset by moderate reductions in sensitivity and in the ROC AUC score. These findings suggest the weather-enhanced model is more effective at avoiding false positives, accurately; however, its ability to detect artificial anomalies is reduced.

4.5.3 Impact by Anomaly Type and Weather Severity

The detailed analysis by anomaly type (shift deviation, abnormal heading, and abnormal speeding) revealed that, on average, the weather-enhanced model demonstrated improved specificity under higher environmental severity. The most significant increase in discrimination ability (ROC AUC difference of 0.1099) was observed for abnormal heading anomalies in high-severity conditions, while other types showed more moderate improvements. Sensitivity improvements were highest for abnormal heading but decreased for shift deviation in adverse weather.

Finally, we observed that when comparing the baseline model to the weather-enhanced model, the ROC AUC score remained relatively stable, indicating that including meteorological data does not significantly affect the overall discriminative ability of the model. However, we do notice an overall improvement in specificity, indicating an improved capacity to predict normal trajectories.

Chapter 5

Discussion

This chapter discusses the experimental findings presented in Chapter 4, contextualizing them within the broader scope of maritime anomaly detection. The discussion is introduced with an analysis of the key findings related to the model's architecture and ability to distinguish between artificial anomalies and weather-induced behavior. It then addresses the inherent threats to the study's validity and concludes by synthesizing these points to evaluate the overall contribution of this research.

5.1 Key Findings and Implications for the Maritime Sector

This research builds upon the framework proposed by [18], presenting an approach that combines the strengths of clustering and sequence-based anomaly detection methods. The modeling architecture demonstrates considerable potential for addressing maritime trajectory anomaly detection, as evidenced by the comprehensive experimental evaluation conducted in [18], including extensive comparisons with alternative methodologies by [27]. The temporal dependency inherent in trajectory analysis necessitates the inclusion of temporally aware models [2]. Unlike clustering-based methods that often neglect the temporal aspect [25], the hybrid approach as presented by [18] captures both spatial clustering patterns and temporal dependencies, providing a more comprehensive representation of vessel behavior given its context [33].
In addition to spatial context, [55] demonstrated that vessel behavior depends on meteorological context. [24] demonstrated that the integration of AIS-based trajectory data with meteorological contextual information for deep learning approaches results in enhanced trajectory prediction capabilities. However, the potential for integration of contextual meteorological data with AIS-based trajectories remained unexplored, as noted in [36, 27, 33]. This thesis demonstrates the potential for maritime anomaly detection by combining meteorological information with AIS-based trajectory data.

Ability to Distinguish Between Weather-Induced Behavior and Anomalous Behavior
The core research objective of this thesis is to determine if meteorological data could help a model distinguish between synthetic anomalous behavioral patterns and legitimate, weather-induced deviations. The results presented in 4.2 show improved performance for the detection of normal behavior as opposed to the capacity to detect (artificial abnormal) behavior, measured by specificity.

These results under Section 4.2 suggest that including meteorological variables enhances the model's ability to identify normal (weather-influenced) trajectories, although it may slightly reduce the ability to detect artificially introduced anomalies. Practically, this means that the models will predict fewer false alarms at the cost of missing out on anomalies. This trade-off is analogous to the results reported in [40], where a decreased false alarm rate resulted in lower model accuracy.

The practical implications of this result depend on the system requirements in an operational setting [33].

Stabile Overall Discriminative Capabilities As we increased the dimensionality of the data from which vessel patterns are identified, we must determine if vessel behavioral patterns can be formed differently, as the ability to learn patterns correctly strongly depends on this parameter (C) [18]. In Section 4.1, we observed that with an increase of C, the overall discriminative ability of the model does not substantially improve with additional complexity. In addition, we observed that when C increased, model specificity increased, although at the expense of a decrease in sensitivity.

This trade-off indicates that as the model becomes more complex, it tends to classify normal trajectories more accurately. However, this comes at the expense of overlooking the trajectories in which artificial anomalous behavior is added. The stability of the ROC AUC scores indicates that the overall discriminative ability of the model remains stable when complexity is added.

This trade-off can be an indication of overfitting behavior of the model caused by architecture design choices. Due to the difference between training and inference objectives (see: Section 3.4.5), we hypothesize that the model learns behavioral patterns too well, causing synthetic anomalies to be overlooked for weather-induced behaviors (an increase for false negative predictions. See Figure: 3.9). This hypothesis relates to work by [24], where meteorological data was successfully utilized to enhance vessel trajectory prediction accuracy. Our work nuances their recommendation for utilizing meteorological data in maritime anomaly detection because of the risk of model overfitting weather behavioral patterns especially due to their substantial accuracy improvement (15.31%) over their baseline methods.

Type of Anomaly Influences Distinguishing Capabilities This research validates its implementation through the use of synthetically created anomalous trajectories. For that reason, we analyze the effect of the methodology used for creating artificial anomalies [18] by comparing the difference of the specificity score (weather-enhanced minus baseline) to assess the model's ability to distinguish between weather-induced behavior (see: Section 4.3).

For the discussion of the results, we exclusively refer to test configurations as reported in Table 4.6 ($d \in \{0, 1\}$, $r = 0.1$, and $\rho = 0.05$). We hypothesize this configuration involves anomalies with minor deviations from the normal trajectory, allowing us to report on the model's discriminative abilities in the optimal setting. However, we do include the results for the other test configurations, allowing for comparability with [18].

The results summarized in Table 4.6 indicate that the impact of weather integration on anomaly detection performance varies by anomaly type. We conducted independent significance tests per type of anomaly. Overall, these results suggest that the impact of meteorological integration on model specificity is significant for distinguishing between normal and anomalous vessel movements in the presence of shift and heading anomalies, but this effect does not generalize to all anomaly types, such as abnormal speeding.

Specifically, the inclusion of weather features led to increased specificity for abnormal heading and speeding anomalies when compared to the baseline configuration. However, we note decreased specificity for anomalies of type shift deviation. This finding corresponds to similar findings by [18], arguing that shift deviation anomalies being less distinguishable from genuine operational behavior. We assume that synthetic anomalies that do not represent actual anomalous behavior will result in better distinguishability from normal behavior, resulting in higher specificity. As opposed to shift deviation anomalies that especially relate to weather. When the model is shown trajectories in which synthetic shift deviation occurs - and severe weather conditions do not occur during that trajectory - The distinguishing capabilities are increased from an actual instance of the positive class, resulting in a higher score for sensitivity instead of specificity. A similar finding was presented in [1]: *"Graph (d) suggests a positive relationship of significant wave height on the prediction of a claim, with the effect tapering off for wave heights above three meters. The latter effect is likely related to evasive maneuvers made by the captain to ensure the safety of the ship and crew during very bad weather conditions, such as altering the heading of the vessel relative to*

the wave direction or reducing vessel speed." The same finding was also presented in [12] which further acknowledges our assumption.

Increased Performance For Severe Weather Conditions The findings presented in Section 4.4 show evidence for the influence of weather severity on the ability to distinguish between weather-induced anomalous behavior and artificial anomalies. The contextually-enhanced model demonstrated improved discriminative capabilities when comparing performance during periods of high and low severity weather conditions. The consistent increase in model specificity under these conditions is particularly relevant to our research objectives, as it signifies that the model successfully learned to associate certain movements (e.g., speed fluctuations, minor heading adjustments) with adverse weather, correctly classifying them as normal within that context. From a practical standpoint, the results of this finding are relevant as it indicates that during severe weather conditions, the ability to reduce false alarms (increased specificity) and overall discriminative performance is increased. As noted in [31, 33], a significant challenge of maritime anomaly detection is the high rate of false alarms due to the neglect of contextual information. In addition, false alarms cause cognitive overload for Vessel Traffic Monitoring staff [40]. Our results show the potential for meteorological data to be incorporated in maritime anomaly detection models to reduce false alarms.

5.1.1 Implications for Practitioners

Operational Resilience The integration of meteorological information into vessel traffic monitoring processes enhances Maritime Situational Awareness (MSA) by enabling a more robust decision-support system for operational staff [20]. Empirical evidence suggests that such integration may contribute to a reduction in false alarm rates, thereby improving the reliability of automated alert mechanisms [31, 36]. Moreover, the contextualization of detected anomalies by explicitly incorporating weather-related data supports more accurate risk assessment, as it allows operators to weigh the consequences of anomalous events based on prevailing environmental conditions [39]. Notably, the inherent interpretability of meteorological features further facilitates the development of explainable anomaly detection frameworks, providing a clearer rationale for system outputs and supporting post-incident analysis [56].

5.1.2 Implications for Researchers

Architecture Design This thesis provides a perspective for incorporating meteorologic data in a deep learning architecture. As noted in [24], challenges exist due to differing frequencies at which AIS data and meteorologic measurements are recorded. In this work, mean and standard deviations are included on the trajectory level showing improvements for reducing false alarm rate. For that reason, this work denotes the potential for exploring other frequencies.

5.2 Study Limitations and Threats to Validity

Within this research project, several limitations applied and possibly affect the validity and the extent to which this research is generalizable. The limitations relate to the geographic characteristics of the training and testing sets. The threads to validity relate to the evaluation framework

5.2.1 Study Limitations

Geographical Limitations The models were trained exclusively on the West Coast dataset provided by [18]. Due to implementation complexities that exceeded anticipated timelines, evaluation on the East Coast dataset was not completed. This limitation restricts the generalizability of the findings and may introduce geographical bias in the assessment of model performance.

Explainability Within maritime anomaly detection, there is a prevalent trend toward utilizing latent representations for trajectory prediction and anomaly detection through generative models [18, 25]. This preference stems from the high-dimensional nature of source data [33], which incorporates both temporal and spatial dimensions. Although deep learning models effectively handle dimensional complexity, they significantly complicate model interpretability, which is a critical operational concern in maritime applications where detection failures may result in loss of human life or substantial material damage [33].

Implementation Challenges Several significant implementation challenges emerged during the development phase. The multi-modal nature of the architecture introduced substantial complexity to the implementation process. The limited availability of detailed implementation specifications necessitated considerable engineering effort to reconstruct the system architecture and parameter configurations. Consequently, the baseline experiment results unperformed compared to those reported by [18], potentially due to implementation differences or suboptimal hyperparameter configurations.

5.2.2 Threats to Validity

External Validity Threats

External validity concerns the extent to which the findings of this study can be generalized beyond the specific experimental conditions under investigation. In this research, a primary limitation arises from the representativeness of weather conditions incorporated into the evaluation framework.

Statistical Testing This study examines the models' capacity to differentiate between normal and weather-induced vessel behaviors, with particular emphasis on scenarios involving severe weather. However, statistical testing was not performed to evaluate the models' performance across varying levels of weather severity. The importance of statistical significance testing is emphasized in literature on maritime anomaly detection [20, 28]. Statistical significance testing is vital to ascertain whether observed differences in model performance are genuinely meaningful or simply a product of random variation [28]. Conducting statistical tests would enhance external validity by providing quantitative evidence on whether observed differences in performance are statistically significant. Specifically, statistical significance testing could further elucidate the models' discriminative abilities in distinguishing weather-induced behaviors from anomalous patterns under differing meteorological conditions [20]. The absence of these tests limits the ability to draw robust generalizations regarding model performance in diverse operational environments.

Construct Validity

Validation with Artificial Anomalies The scarcity of labeled datasets represents a well-documented challenge in vessel anomaly detection research [33, 18, 51, 19]. In this work, we adopt the validation framework as presented in [18] and acknowledge the limited representativeness of synthetic anomalies to relation to actual anomalies. Our efforts to mitigate this limitation are as follows: we only reported on test configurations in which the ratio of anomalies (r) to the total number of samples is equal to 0.1 to ensure maximum representation of the positive class. Also, our key findings are solely based on anomalies with the smallest deviation from the original trajectory (d and ρ) allowing for optimal evaluation of our research questions.

Internal Validity Threats

Internal validity refers to the extent to which the observed effects can be confidently attributed to the experimental treatments, rather than to confounding variables. In this work the main internal validity threats relate to the data distribution of weather variables used for the testing dataset.

Test Set Distribution As this research was conducted using processed datasets as provided by [18], we were unable to ensure equal representation of weather conditions between datasets for training and testing purposes. In [18], trajectories are inferred from AIS data with a temporal dimension ranging from January to August for training and validation, and the test set is limited to the month of September.

Figure 5.1 shows the distributions for significant wave height and observed different distributions between the datasets. We mitigated this by randomizing trajectories for the train and validation sets, but were unable to do so for the dataset used for testing, since it contained anomalies. We assume this limits our research due to severe weather conditions not being represented equally in the test set, creating a substantial discrepancy, especially for the Deep Gaussian Mixture Model from the perspective of clustering trajectories on meteorological features.

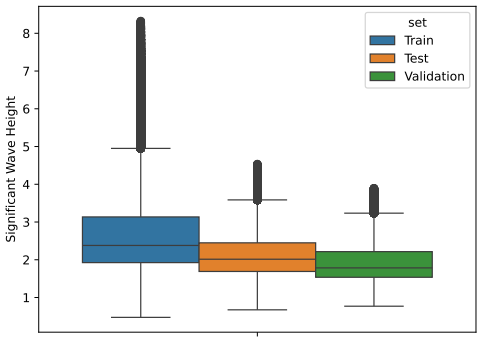

Figure 5.1: Distribution of Significant Wave Height Visualized per Dataset.

The test set partitioning methodology used in the experiment evaluation for RQ4 (see: Section 4.4 is transparently reported and grounded in prior literature [4, 12, 1, 24], supporting reproducibility and the specific objectives of RQ4. However, the reduction of environmental variability to a single aggregated measure and dichotomous grouping limits the granularity and ecological validity of the conclusions. Consideration of continuous or multivariate partitioning schemes, as well as potential confounders, would enhance both the interpretability and external validity of the findings.

Chapter 6

Conclusions

6.1 Answers to Research Questions

6.1.1 RQ1: Impact of Gaussian Mixture Components

Experimental results confirm that the overall discriminative capability, as measured by the ROC AUC score, remains stable as C increases, suggesting that added complexity does not yield substantial improvement in overall anomaly detection capabilities but rather refines the classification of normal behavior. These findings emphasize the necessity of careful model selection in operational contexts to optimize operational trade-offs between false-positive and false-negative alarm rates.

6.1.2 RQ2: Impact of Meteorological Variables

The integration of meteorological variables consistently improves model specificity across experimental setups. For both Setup A and B, which span differing spatial deviations and anomaly types, the weather-enhanced model yields a notable reduction in false positive rates, indicating increased ability to correctly classify weather-influenced trajectories as normal. However, this improvement in avoiding false positives corresponds to moderate reductions in both sensitivity and ROC AUC, indicating a slight decrease in the model's capacity to identify artificial anomalies. Thus, the principal benefit of adding meteorological data is the enhanced resilience to weather-induced "anomalous" maneuvers, making the model less prone to mistakenly flag legitimate weather-induced behaviors as anomalous.

6.1.3 RQ3: Impact of Meteorological Data Integration

Performance analysis by anomaly type reveals that weather data integration significantly differs per type of anomaly. The most substantial increases in specificity (false positive rate reduction) occur for abnormal heading and abnormal speeding anomalies. This suggests that these anomaly classes are better distinguished from environmental effects by the model when meteorological context is included. In contrast, shift deviation anomalies register a decrease in specificity with meteorological integration, possibly due to representing the genuine corrective vessel behavior better in heavy weather than abnormal heading and abnormal speeding anomalies. This finding was found to be statistically significant and shows consistency with the literature on vessel behavior during extreme weather conditions.

6.1.4 RQ4: Model Performance Under Varying Weather Conditions and Environmental Scenarios

The analysis of model performance under varying weather severity demonstrates that weather-aware models maintain, and in certain cases improve, overall discriminative ability when exposed to more challenging environmental conditions. Specifically, increased ROC AUC scores in high-severity weather are an indicator of more accurate recognition of legitimate, context-driven maneuvers as normal. The practical implication is that the integration of meteorological data enhances anomaly detection during periods of high environmental stress. This finding supports the view that contextual information can be leveraged to reduce false alarms.

6.2 Research Contributions

This thesis makes several novel contributions to the field of maritime anomaly detection through the integration of meteorological data in the context of deep learning-based trajectory analysis. First, it demonstrates that external environmental factors—such as wind, wave height, and other weather variables—substantially influence normal vessel movements and, if unaccounted for, may confound the detection of artificial anomalies. By proposing and empirically validating a weather-enhanced anomaly detection model, this research advances state-of-the-art methodology and provides robust evidence supporting the inclusion of meteorological information for improved understanding of anomalous weather-induced behaviors. Second, this work introduces a rigorous evaluation framework that employs synthetic anomaly injection and a comprehensive suite of statistical analyses, enabling the systematic benchmarking of model performance under various operational and environmental conditions. The proposed approach, including the adaptation of the Gaussian Mixture Model component to incorporate meteorological features, represents a methodological innovation with direct relevance to practitioners deploying operational maritime surveillance and safety systems. Third, the study provides insights into the impact of model configuration choices (e.g., number of GMM components) and puts the conditions under which weather integration is most beneficial, particularly in scenarios with elevated environmental variability.

6.3 Recommendations for Future Work

6.3.1 Generalizability and Data Diversity

This thesis is limited by the representativeness of meteorological extremes in the test set and distributional discrepancy between training and test splits. Future studies should aim for geographically and temporally diverse datasets, encompassing different traffic densities, vessel types, and a broader spectrum of environmental conditions. Further validation efforts are needed for operational assessment, a potential area to explore could be domain expert consultation.

6.3.2 Architecture Optimization

In this study, the clustering component was integrated within the overall model architecture to enable the transformer module to directly leverage the clustering characteristics of vessel trajectories. While this integrated approach confers benefits in terms of contextual representation and end-to-end learning, it also entails increased computational complexity and potential scalability constraints.

Future research should systematically explore the optimization of hybrid modeling strategies, with particular emphasis on decoupled architectures. A decoupled approach, separating clustering and sequence modeling components, offers several advantages. Firstly, such architectures may enhance the flexibility of context integration, enabling modular inclusion or exclusion of environmental features as operational scenarios require. This modularity can support broader Maritime

Situational Awareness (MSA) objectives, particularly by facilitating model sharing and adaptation between governmental or interagency partners. Secondly, decoupled models inherently improve explainability, as the contributions of clustering and sequential dynamics to anomaly detection can be more transparently quantified, potentially reducing the operational burden on vessel monitoring personnel.

Furthermore, future work should focus on optimizing architectures for enhanced scalability, addressing both geographical coverage and the diversity of operational contexts. While mapping continuous and circular variables (e.g., heading, speed, meteorological variables) to discrete feature spaces is a common practice in the domain, standardization in discretization methodologies remains limited. Further investigation is warranted to assess whether the adoption of standardized discretization frameworks can improve comparability across studies and support the development of globally scalable maritime anomaly detection systems.

6.3.3 Explainable Anomaly Detection

Meteorological data presents a valuable complementary data source for enhancing explainable maritime anomaly detection models. As noted by [56], the integration of domain knowledge and contextual information significantly improves both detection accuracy and explanation fidelity. Weather conditions such as wind speed, wave height, visibility, and storm patterns directly influence vessel behavior and can provide crucial context for distinguishing between normal operational responses to environmental conditions and genuine anomalies. The inherent interpretability of meteorological variables aligns well with their recommendations for explainability in [56], as maritime domain experts can readily understand how weather factors contribute to trajectory deviations or operational anomalies. Furthermore, incorporating weather data addresses one of the key challenges identified in XAD literature: the consistency between Oracle-Definition (what maritime experts consider anomalous) and Detection-Definition (what the model identifies), especially since weather-induced behavioral changes have well-established patterns in maritime operations [43, 46, 24, 33, 7].

6.3.4 Challenges and Directions in Anomaly Definition and Feature Engineering

This research primarily focused on assessing the influence of meteorological variables on identifying vessel behaviors that, while potentially labeled as anomalous by conventional models, are in fact legitimate responses to prevailing weather conditions. The primary objective was to distinguish between artificial anomalies and those behaviors induced by environmental factors, thereby reducing the occurrence of false positive detections.

A critical limitation within the field of maritime anomaly detection remains the absence of comprehensive, labeled datasets with validated annotations of anomalous events. The current lack of consensus regarding the operational definition of anomalies impedes both the development and benchmarking of advanced detection methodologies. Progress in this domain would be substantially advanced through collaborative efforts to establish standardized taxonomies of maritime anomalies, including rigorous validation of what constitutes anomalous versus contextually appropriate behavior, particularly under severe weather conditions.

Furthermore, future research should prioritize the development of precise definitions and formal representations of the kinematics underlying anomalous vessel actions. This includes characterizing trajectory deviations, abnormal speed profiles, and atypical course alterations in relation to contemporaneous meteorological phenomena. By systematically relating detailed kinematic patterns to specific environmental triggers, new studies can provide a clearer distinction between operationally justified and genuinely suspicious maritime behaviors. Such advancements in feature selection and engineering are essential for enhancing model robustness, interpretability, and operational utility in real-world maritime monitoring scenarios.

Bibliography

[1] Roar Adland, Haiying Jia, Tønnes Lode, and Jørgen Skontorp. The value of meteorological data in marine risk assessment. *Reliability Engineering & System Safety*, 209:107480, 2021. 5, 6, 10, 11, 13, 14, 17, 31, 44, 48, 51

[2] Julien Audibert, Pietro Michiardi, Frédéric Guyard, Sébastien Marti, and Maria A Zuluaga. Usad: Unsupervised anomaly detection on multivariate time series. In *Proceedings of the 26th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 3395–3404, 2020. 10, 21, 47

[3] Xiangen Bai, Zhexin Xie, Xiaofeng Xu, and Yingjie Xiao. An adaptive threshold fast dbscan algorithm with preserved trajectory feature points for vessel trajectory clustering. *Ocean Engineering*, 280:114930, 2023. 9

[4] Peter Brandt, Ziaul Haque Munim, Meriam Chaal, and Hooi-Siang Kang. Maritime accident risk prediction integrating weather data using machine learning. *Transportation Research Part D: Transport and Environment*, 136:104388, 2024. 1, 2, 4, 5, 6, 7, 10, 11, 13, 14, 17, 51

[5] Varun Chandola, Arindam Banerjee, and Vipin Kumar. Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):1–58, 2009. 7, 8, 9, 29

[6] Davide Chicco, Niklas Tötsch, and Giuseppe Jurman. The matthews correlation coefficient (mcc) is more reliable than balanced accuracy, bookmaker informedness, and markedness in two-class confusion matrix evaluation. *BioData mining*, 14:1–22, 2021. 27

[7] Andrej Dobrkovic, Maria-Eugenia Iacob, and Jos Van Hillegersberg. Using machine learning for unsupervised maritime waypoint discovery from streaming ais data. In *Proceedings of the 15th international conference on knowledge technologies and data-driven business*, pages 1–8, 2015. 6, 13, 17, 54

[8] EMSA. Automated Behaviour Monitoring — emsa.europa.eu. `https://www.emsa.europa.eu/combined-maritime-data-menu/abm.html`, 2015. [Accessed 23-05-2025]. v, 8, 27

[9] Martin Ester, Hans-Peter Kriegel, Jörg Sander, Xiaowei Xu, et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In *kdd*, volume 96, 1996. 9

[10] European Maritime Safety Agency. Annual Overview of Marine Casualties and Incidents 2024. Annual Overview Ares(2024)8229157, European Maritime Safety Agency, Lisbon, Portugal, june 2024. 1

[11] John A Hanley et al. Receiver operating characteristic (roc) methodology: the state of the art. *Crit Rev Diagn Imaging*, 29(3):307–335, 1989. 29

[12] Christiaan Heij and Sabine Knapp. Effects of wind strength and wave height on ship incident risk: regional trends and seasonality. *Transportation Research Part D: Transport and Environment*, 37:29–39, 2015. 1, 5, 6, 11, 13, 14, 17, 29, 49, 51

[13] Hans Hersbach, Bill Bell, Paul Berrisford, Gionata Biavati, András Horányi, J Muñoz Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Iryna Rozum, et al. Era5 hourly data on single levels from 1979 to present. *Copernicus climate change service (c3s) climate data store (cds)*, 10(10.24381), 2018. 14

[14] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 24

[15] Kustwacht. Vessel Traffic Monitoring (vtmon). https://kustwacht.nl/hulpverlening/vtm/, 2021. [Accessed 05-05-2025]. 13

[16] Jesse M Lane and Michael Pretes. Maritime dependency and economic prosperity: Why access to oceanic trade matters. *Marine Policy*, 121:104180, 2020. 1, 4

[17] Huanhuan Li, Jingxian Liu, Kefeng Wu, Zaili Yang, Ryan Wen Liu, and Naixue Xiong. Spatio-temporal vessel trajectory clustering based on data mapping and density. *IEEE Access*, 6:58939–58954, 2018. 9

[18] Hui Li, Wengen Li, Shuyu Wang, Hanchen Yang, Jihong Guan, and Yichao Zhang. Stad: Ship trajectory anomaly detection in ocean with dynamic pattern clustering. *Ocean Engineering*, 313:119530, 2024. v, 1, 2, 5, 10, 11, 13, 15, 17, 18, 19, 20, 21, 22, 23, 24, 25, 27, 29, 32, 33, 36, 38, 46, 47, 48, 49, 50, 51

[19] Maohan Liang, Jianlong Su, Ryan Wen Liu, and Jasmine Siu Lee Lam. Aisclean: Ais data-driven vessel trajectory reconstruction under uncertain conditions. *Ocean Engineering*, 306:117987, 2024. 5, 50

[20] Maohan Liang, Lingxuan Weng, Ruobin Gao, Yan Li, and Liang Du. Unsupervised maritime anomaly detection for intelligent situational awareness using ais data. *Knowledge-Based Systems*, 284:111313, 2024. 1, 4, 7, 10, 11, 12, 13, 17, 18, 24, 49, 50

[21] Qi Liu, Rudy Klucik, Chao Chen, Glenn Grant, David Gallaher, Qin Lv, and Li Shang. Unsupervised detection of contextual anomaly in remotely sensed data. *Remote Sensing of Environment*, 202:75–87, 2017. 11, 17

[22] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017. 24

[23] Meifeng Luo and Sung-Ho Shin. Half-century research developments in maritime accidents: Future directions. *Accident Analysis & Prevention*, 123:448–460, 2019. 1, 4

[24] Saeed Mehri, Ali Asghar Alesheikh, and Anahid Basiri. A context-aware approach for vessels' trajectory prediction. *Ocean Engineering*, 282:114916, 2023. 1, 5, 6, 11, 13, 14, 22, 44, 47, 48, 49, 51, 54

[25] Brian Murray and Lokukaluge Prasad Perera. A dual linear autoencoder approach for vessel trajectory prediction using historical ais data. *Ocean engineering*, 209:107478, 2020. 1, 4, 5, 7, 10, 11, 47, 50

[26] Duong Nguyen, Rodolphe Vadaine, Guillaume Hajduch, René Garello, and Ronan Fablet. A multi-task deep learning architecture for maritime surveillance using ais data streams. In *2018 IEEE 5th International Conference on Data Science and Advanced Analytics (DSAA)*, pages 331–340. IEEE, 2018. 11

[27] Duong Nguyen, Rodolphe Vadaine, Guillaume Hajduch, René Garello, and Ronan Fablet. Geotracknet—a maritime anomaly detector using probabilistic neural network representation of ais tracks and a contrario detection. *IEEE Transactions on Intelligent Transportation Systems*, 23(6):5655–5667, 2021. 1, 8, 9, 10, 13, 27, 29, 47

[28] John Paparrizos, Paul Boniol, Themis Palpanas, Ruey S Tsay, Aaron Elmore, and Michael J Franklin. Volume under the surface: a new accuracy evaluation measure for time-series anomaly detection. *Proceedings of the VLDB Endowment*, 15(11):2774–2787, 2022. 21, 27, 29, 50

[29] Matilda QR Pembury Smith and Graeme D Ruxton. Effective use of the mcnemar test. *Behavioral Ecology and Sociobiology*, 74:1–9, 2020. 29, 30, 43

[30] PyTorch Team. Cosineannealinglr. `https://docs.pytorch.org/docs/stable/generated/torch.optim.lr_scheduler.CosineAnnealingLR.html`, 2025. Accessed: 2025-07-01. 59

[31] Aungon Nag Radon, Ke Wang, Uwe Glässer, Hans Wehn, and Andrew Westwell-Roper. Contextual verification for false alarm reduction in maritime anomaly detection. In *2015 IEEE International Conference on Big Data (Big Data)*, pages 1123–1133. IEEE, 2015. 1, 11, 30, 49

[32] Douglas Reynolds. Gaussian mixture models. In *Encyclopedia of biometrics*, pages 827–832. Springer, 2015. 21

[33] Claudio V. Ribeiro, Aline Paes, and Daniel de Oliveira. Ais-based maritime anomaly traffic detection: A review. *Expert Systems with Applications*, 231:120561, 2023. 1, 4, 5, 6, 7, 8, 9, 10, 11, 12, 13, 14, 17, 24, 27, 29, 47, 48, 49, 50, 54

[34] Rijkswaterstaat. Monitoring, 2024. 13

[35] Maria Riveiro, Giuliana Pallotta, and Michele Vespe. Maritime anomaly detection: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8(5):e1266, 2018. 7, 9, 17

[36] Hao Rong, AP Teixeira, and C Guedes Soares. A framework for ship abnormal behaviour detection and classification using ais data. *Reliability Engineering & System Safety*, 247:110105, 2024. 5, 7, 9, 10, 11, 12, 14, 24, 29, 47, 49

[37] Sebastian Ruder. An overview of gradient descent optimization algorithms. *arXiv preprint arXiv:1609.04747*, 2016. 24

[38] Marcus D Ruopp, Neil J Perkins, Brian W Whitcomb, and Enrique F Schisterman. Youden index and optimal cut-point estimated from observations affected by a lower limit of detection. *Biometrical Journal: Journal of Mathematical Methods in Biosciences*, 50(3):419–430, 2008. 27

[39] Rohan Sinha, Amine Elhafsi, Christopher Agia, Matthew Foutter, Edward Schmerling, and Marco Pavone. Real-time anomaly detection and reactive planning with large language models. *arXiv preprint arXiv:2407.08735*, 2024. 25, 49

[40] Fernando Terroso-Saenz, Mercedes Valdes-Vela, and Antonio F Skarmeta-Gomez. A complex event processing approach to detect abnormal behaviours in the marine environment. *Information Systems Frontiers*, 18:765–780, 2016. 11, 47, 49

[41] Michael Tschannen, Olivier Bachem, and Mario Lucic. Recent advances in autoencoder-based representation learning. *arXiv preprint arXiv:1812.05069*, 2018. 21

[42] UNCTAD. Review of maritime transport. `https://unctad.org/system/files/official-document/rmt2024overview_en.pdf`, 2024. [Accessed 12-05-2025]. 1, 4

[43] Solange van der Werff, Mark van Koningsveld, and Fedor Baart. Vessel behaviour under varying environmental conditions in coastal areas. In *35th PIANC World Congress*, Cape Town, South Africa, April 2024. 29 April – 23 May 2024. 1, 5, 6, 11, 17, 18, 29, 54

[44] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017. 17, 20

[45] Michel Verleysen and Damien François. The curse of dimensionality in data mining and time series prediction. In *International work-conference on artificial neural networks*, pages 758–770. Springer, 2005. 9, 21

[46] Roberto Vettor and C Guedes Soares. Development of a ship weather routing system. *Ocean Engineering*, 123:1–14, 2016. 6, 54

[47] Cinzia Viroli and Geoffrey J McLachlan. Deep gaussian mixture models. *Statistics and Computing*, 29:43–51, 2019. 21, 22, 23

[48] Colin Wei, Sham Kakade, and Tengyu Ma. The implicit and explicit regularization effects of dropout. In *International conference on machine learning*, pages 10181–10192. PMLR, 2020. 34

[49] Rüdiger Wirth and Jochen Hipp. Crisp-dm: Towards a standard process model for data mining. In *Proceedings of the 4th international conference on the practical applications of knowledge discovery and data mining*, volume 1, pages 29–39. Manchester, 2000. v, 12

[50] Konrad Wolsing, Linus Roepert, Jan Bauer, and Klaus Wehrle. Anomaly detection in maritime ais tracks: A review of recent approaches. *Journal of Marine Science and Engineering*, 10(1):112, 2022. 11

[51] Dawen Xia, Yunsong Li, Yuce Ao, Xiaoduo Wei, Yan Chen, Yang Hu, Yantao Li, and Huaqing Li. Parallel recurrent neural network with transformer for anomalous trajectory detection. *Applied Intelligence*, 55(6):519, 2025. 10, 20, 50

[52] Lei Xie, Tao Guo, Jiliang Chang, Chengpeng Wan, Xinyuan Hu, Yang Yang, and Changkui Ou. A novel model for ship trajectory anomaly detection based on gaussian mixture variational autoencoder. *IEEE Transactions on Vehicular Technology*, 72(11):13826–13835, 2023. 10

[53] William J Youden. Index for rating diagnostic tests. *Cancer*, 3(1):32–35, 1950. 27

[54] Chuang Zhang, Songtao Liu, Muzhuang Guo, and Yuanchang Liu. A novel ship trajectory clustering analysis and anomaly detection method based on ais data. *Ocean Engineering*, 288:116082, 2023. 5, 11, 17, 18

[55] Yang Zhou, Winnie Daamen, Tiedo Vellinga, and Serge P Hoogendoorn. Impacts of wind and current on ship behavior in ports and waterways: A quantitative analysis based on ais data. *Ocean Engineering*, 213:107774, 2020. v, 5, 6, 10, 11, 47

[56] Bohan Zhuang, Jing Liu, Zizheng Pan, Haoyu He, Yuetian Weng, and Chunhua Shen. A survey on efficient training of transformers. *arXiv preprint arXiv:2302.01107*, 2023. 10, 24, 49, 54

Appendix

Table 1: Model Hyperparameters

Hyperparameter	**Value**	**Description**
	Model Architecture	
Embedding dimension	32	Dimension for trajectory window embedding (e_t) (see: Section 3.4.2)
Transformer heads	8	Number of attention heads in transformer encoder model (see: Section 3.4.2)
Transformer layers	4	Number of transformer encoder layers (see: Section 3.4.2)
Max sequence length	10	Maximum trajectory (e_t) window size
Latent dimension (AE)	32	Autoencoder compression dimensionality (see: Section 3.4.2)
GMM hidden dimension	$\{32, 48\}$	Hidden layer size in GMM network. (See: Section 3.4.2) Please note that for the experiments conducted, this parameter was scaled according to the specification in Section 4.1.1.
GMM components (C)	20	Number of Gaussian mixture components (see: Section 3.4.2)
Dropout	0.1	Dropout probability
GMM epsilon (p_{GMM})	1×10^{-7}	Added to the GMM covariance matrix to prevent numerical instability
Loss epsilon	1	Added to the penalty term formulation of the covariance matrix (see: Section 3.4.5)
	Training Configuration	
Optimizer	AdamW	Optimizer with weight decay
Learning rate	1×10^{-5}	Peak learning rate
Weight decay	0.1	L2 regularization strength
Epochs	$\{100, 250\}$	100 for experiments conducted for RQ1 (see: 4.1) and 250 for the experiments conducted for RQ2, RQ3 and RQ4 (see: 4.1
Scheduler type	OneCycleLR	Base configuration adapted from [30]
Anneal strategy	Cosine	Learning rate decay pattern
	Loss Function Weights (see: Section 3.4.5)	
λ_1 (GMM energy)	1.0	Energy loss weight
λ_2 (AE reconstruction)	1.0	Autoencoder loss weight
λ_3 (GMM penalty)	0.005	Regularization penalty weight