Enhancing Situation Awareness of Maritime Surveillance Operators using Deep Learning based Abnormal Maritime Behaviour Detection

Olesen, Kristoffer Vinther

Publication date:
2023

Document Version
Publisher's PDF, also known as Version of record

Enhancing Situation Awareness of Maritime Surveillance Operators using Deep Learning based Abnormal Maritime Behaviour Detection

Kristoffer Vinther Olesen

Kongens Lyngby 2023

Technical University of Denmark
Department of Applied Mathematics and Computer Science
Richard Petersens Plads, building 324,
2800 Kongens Lyngby, Denmark
Phone +45 4525 3031
compute@compute.dtu.dk
www.compute.dtu.dk

Summary (English)

During recent years, we have seen a growing importance of maritime security to ensure the safety of maritime traffic, territorial protection, and the protection of key infrastructure assets. Maritime Surveillance Command and Control systems offer automated tools for enhancing the situation awareness of surveillance operators to improve their decision-making and detect abnormal illicit maritime behavior. Most automated tools in use currently scale poorly with large amounts of data and data sources which raises issues regarding the generalization across space and time and makes detection more prone to errors that may have fatal consequences. The goal of the thesis is to assess the applicability of deep learning methodologies to enhance situation awareness of surveillance operators.

In the first part of the thesis, we present an overview of current methods for the detection of maritime abnormalities and discuss how they address some of the issues found in practical application. We then introduce deep learning architectures for the analysis of maritime trajectories. These architectures are based on Recurrent Neural Networks that model trajectories of variable length sequentially. Abnormal trajectories may be detected based on predictive errors in a Sequence-2-Sequence architecture or reconstructive errors in sequential Variational AutoEncoders. This results in models that can analyze maritime trajectories, detect abnormal trajectories, and be trained in a scalable way using large unlabelled datasets.

In the second part of the thesis, we evaluate how deep learning architectures may enhance situation awareness of surveillance operators. On the basis of manually annotated abnormal trajectories related to a recent collision accident,

we make a quantitative comparison of the automated detection of abnormal trajectories. Qualitative comparison of flagged trajectories indicates that deep neural networks of different architectures flag different types of abnormal behavior. Therefore, we suggest ensembles of different deep learning architectures in which each member is designed specifically with the detection of a certain type of abnormal behavior in mind. We investigate how to extract the learned normalcy models through interpretable latent variables, but find the encoded information lacking in describing local behavior and insufficient to be used for man-in-the-loop style detections. We propose a two-step clustering approach to describe local behavior and find that this is superior in describing behavioral differences along the same maritime route and for discovering abnormal behavior outside the main maritime routes.

Summary (Danish)

I de senere år er den maritime sikkerhed blevet et vigtigere og vigtigere emne for at sikre søtrafikken, beskyttelse af territoriale grænser og vigtig infrastruktur. Maritime Surveillance Command and Control systemer tilbyder automatiserede værktøjer til at øge overvågningsoperatørernes situationsbevidsthed for at forbedre deres beslutningstagen og opdage anormal ulovlig maritim adfærd. De fleste automatiserede værktøjer i brug i dag skalerer dårligt med store mængder data og datakilder, hvilket rejser spørgsmål vedrørende generalisering på tværs af rum og tid, og gør detektionen mere tilbøjelig til fejl, der kan have fatale konsekvenser. Målet med denne afhandling er at vurdere anvendeligheden af deep learning-metoder for at øge overvågningsoperatørernes situationsbevidsthed.

I den første del af afhandlingen præsenterer vi et overblik over aktuelle metoder til detektion af maritime abnormaliteter og diskuterer, hvordan de løser nogle af de problemstillinger, der findes i praktisk anvendelse. Vi introducerer derefter deep learning-arkitekturer til analyse af maritime trajektorier. Disse arkitekturer er baseret på Recurrent Neurale Netværk, der sekventielt modellerer trajektorier med variabel længde. Anormale trajektorier kan detekteres baseret på prædiktionsfejl i en Sequence-2-Sequence-arkitektur eller rekonstruktive fejl i sekventielle Variational AutoEncoders. Dette resulterer i modeller, der kan analysere maritime trajektorier, detektere anormale baner og trænes på en skalerbar måde ved hjælp af store uannoteret datasæt.

I anden del af afhandlingen evaluerer vi, hvordan deep learning-arkitekturer kan øge overvågningsoperatørernes situationsbevidsthed. På baggrund af manuelt annoterede anormale trajektorier relateret til en kollisionsulykke sket for nyligt, foretager vi en kvantitativ sammenligning af den automatiske detektering

af unormale trajektorier. Kvalitativ sammenligning af detekteret trajektorier indikerer, at dybe neurale netværk af forskellige arkitekturer detekterer forskellige typer af anormal maritim adfærd. Derfor foreslår vi ensembler af forskellige deep learning-arkitekturer, hvor hvert medlem er designet specifikt til detektion af en bestemt type anormal adfærd. Vi undersøger, hvordan man kan ekstrahere de lærte normalitetsmodeller gennem fortolkbare latente variabler, men finder, at den kodede information er mangelfuld i beskrivelsen af lokal adfærd og er utilstrækkelig til at blive brugt til detektioner i en man-in-the-loop stil. Vi foreslår en to-trins cluster-tilgang til at beskrive lokal adfærd og finder, at denne er overlegen til at beskrive adfærdsforskelle langs den samme maritime rute og til at opdage anormal adfærd uden for de veldefinerede maritime ruter.

Preface

This thesis was prepared at the Section for Statistics and Data Analysis at DTU Compute, Department of Applied Mathematics and Computer Science, Technical University of Denmark. It constitutes a partial fulfillment of the requirements for acquiring a Ph.D. at the Technical University of Denmark.

This PhD project was financed by the Danish Defense Acquisition and Logistics Organization, grant no.: 4600005159, and by DTU Compute, and was supervised by Line Katrine Harder Clemmensen, Anders Nymark Christensen (DTU Compute), and Sune Hørlück (Terma A/S). The PhD project was carried out at DTU during the period January 2020 - January 2023.

The thesis is divided into two main parts. The first part consists of an introduction to maritime abnormality detection and deep learning, while the second part contains research papers (2 submitted for first review and 2 preprints) that investigate the application of deep learning and clustering models to the problem of maritime abnormality detection. The first two papers are mainly concerned with the application and comparison of techniques and deep learning methods to the problem of maritime abnormality detection. The final two papers question the interpretability of deep learning methods, and we propose a new clustering approach for maritime trajectories.

Lyngby, 14-January-2023

Kristoffer Vinther Olesen

Acknowledgements

This thesis is the culmination of an exciting three-year journey, and I consider myself lucky to have been given the opportunity to conduct research on such an important topic as maritime security. If anything, events during my PhD have highlighted the necessity of effective tool to ensure maritime safety and security. I am extremely grateful to DTU, Terma A/S, and the Danish Defense Acquisition and Logistics Organization for funding and providing me with this amazing opportunity.

First of all, I would like to thank my supervisors Associate Professor Line Katrine Harder Clemmensen and Associate Professor Anders Nymark Christensen for their fundamental role in everything I have accomplished during my PhD. You have been exemplary in your support and your inspiration and enthusiasm have guided me through a challenging but also enjoyable journey. Thank you also to my industrial co-supervisor Data Scientist Sune Hørlück for our discussions and your extensive work in capturing, cleaning, storing, and preparing the data I have used in my work. Your domain knowledge and practical approach has often helped keep my work grounded and have ensured practical and operational applications were always accounted for. The work presented in this dissertation is the result of countless hours of our joint work.

A special thanks goes to Asger Lund Christensen for managing the greater project between three large institutions and helping me stay on schedule. Thank you to Felix Gravila for interesting discussions on everything deep learning related to my and his own work. Our discussion often helped stimulate my intellectual energy and curiosity. Also, thanks to all the other great Terma people I have met and that helped me during my work. Thank you to Katrine Feld and

Jesper Holm of the Danish Defense Acquisition and Logistics Organization for their incredible work in organizing feedback sessions and workshops with operational units. A big thank you to everyone at MOC, NMOC, SOE, and other branches of the Danish Navy for your interest in the project and invaluable feedback on my work.

In Winter and Spring of 2022 I had the pleasure of visiting Professor Robert Jenssen and the Machine Learning Group at the University of Tromsø, Norway. I am grateful to Robert, Ahcene Boubekki, Michael Kampffmeyer, and everyone at UiT for their hospitality and ensuring that my stay was both professionally and personally rewarding. Chapter 9 of this dissertation is a direct result of our collaborative work. Thank you to Visual Intelligence for hosting my stay and thank you to Otto Mønsteds, Thomas B. Thriges, and Knud Højgaards Foundations for making my stay possible. Leaving the office at night under the northern lights is something I will always remember.

I also express gratitude to my friends and colleagues at the Section for Statistics and Data Analysis. In particular, my office mates in Room 324-210 for making our long office days much more fun, lunches, coffee breaks, and for academic as well as nonacademic discussions. Thank you for your friendship and support, even during lockdown.

Last but not least, I am grateful to my friends and family in Denmark and USA for your unconditional support. A special emotional thank you to my wife Linnea, who showered me with all her love and support throughout my PhD, and who were understanding and kind enough to bear with me when I was busy and frustrated. Without you I would never have made it and I am forever grateful.

Contents

Chapter 1

Introduction

The aim of this thesis is to progress the research within deep learning for maritime abnormality detection. This thesis develop, discuss, and evaluate methods from an operational perspective and provides an evaluation framework for practitioners, who are looking to develop automated models for enhancing situation awareness of maritime surveillance operators

The primary target audience of this thesis is practitioners who are looking to develop automated models for enhancing situation awareness of surveillance operators, or operators trying to evaluate the potential and limitations of deep learning models for automated abnormality detection. Basic knowledge of machine learning and deep learning is recommended, but not required, to understand this thesis.

This chapter provides an overview of the research presented in this thesis. First, the motivation and background of the study are presented. The research objectives and main research contributions are outlined next. Finally, an outline of the remainder of the thesis is presented.

1.1 Motivation and Background

Historically, human civilization and society were often established in close proximity to water. Easy access to water is a requirement for sustained growth and forms the basis for farming, sanitation, and production. Similarly, water transport is more efficient, convenient, and cheaper than land travel, so securing waterways has always been important for the spread of goods and ideas. To this day, the oceans remain one of the most important ways of connecting and supporting a growing globalized world. Maritime shipping is the most efficient and cost-effective form of long-distance transportation and is responsible for 80% of the world's trade [IMO, 2021]. Recent geopolitical issues in Europe have shown the vulnerability of our society to sabotage of critical infrastructure and the necessity of free trade and transportation. The restriction of free trade and transportation of food, feed materials, fuel, energy, etc. have propelled into a food- and energy-crisis due to rising commodity costs [World Food Program, 2022]. Most recently, the sabotage of the Nord Stream gas pipelines in the Baltic Sea [Ljungqvist] has raised the issue of territorial protection and protection of key infrastructure assets.

These events highlight the need for maritime security and safety to protect the global supply chain and key societal infrastructure. Maritime Surveillance Command and Control play a crucial role in ensuring safe navigation, general safety of marine traffic, and maritime border security. Surveillance operators monitor large areas of the ocean and predict emerging critical situations from several threats such as collisions, smuggling, piracy, overfishing, maritime pollution, territorial incursions, etc. Current operational systems rely strongly on human experts. Data is collected from a wide range of sensors, like radar, sonar, and Automatic Identification System (AIS) as well intelligence reports and weather data. As the number of data sources increases, it becomes increasingly difficult for operators to process the amount of information due to a number of factors, such as cognitive overload, time pressure, fatigue, and uncertainty, in addition to the complex and heterogeneous nature of the data. In order to provide support for the operators, methods and systems capable of abnormality detection is one of the most important tasks within the domain of maritime surveillance and is a very active research area Pallotta et al. [2013a], Nguyen et al. [2021], Singh et al. [2021].

Human operators rely on their level of situation awareness [Endsley, 1995], developed from experience, to detect critical situations and initiate the appropriate response. Situation awareness is the perception of your surrounding environment and events and how these affect your decision making now and in the future. However, Endsley [2017] highlight potential issues when human decision making is based on autonomous systems.

- The so-called Automation Conundrum describes the degradation of human situation awareness as automation increases and the observation that human operators are less likely to be able to take over manual control when needed.
- Automated support of decision-making has been found to be problematic due to a decision bias effect, where operators tend to follow the decision of the automated system.
- Human-autonomy interaction tends to follow a serial model in which operators use decision support systems as input along with other information to make a final decision. This has been found to be less reliable than if operators and automated systems reached a decision separately in parallel, which is then combined for a final decision.

As a consequence of these issues, Endsley [2017] recommends automation systems to support situation awareness rather than decision making.

Surveillance operators likely use their situation awareness to mentally construct a model of the normal maritime picture and base their decision of normality on comparing observed behavior with the established normal behavior model. Automatic construction of the normalcy models may support operators in obtaining a sufficient level of situation awareness. Currently, the most common way to construct the normality model is to use clustering on individual AIS updates Pallotta et al. [2013a], Liu et al. [2015b] or using trajectory similarity measures Zhao and Shi [2019a], Yang et al. [2022b]. However, real-life maritime laws and regulations are complex. Ship types, such as cargo and tankers, mainly follow well-defined shipping lanes with near-constant speeds, whereas other ship types, such as fishing vessels and sailing ships, have less constrained and more complex behavior. The exact route of these trajectories is less important than the local behaviour, i.e. speed and course changes, due to the additional freedom of movement. The mixture of densely populated shipping lanes and more sparse regions with increased behavioural freedom complicates clustering.

During the past decade, research on methods for the detection of maritime abnormalities has been accelerated by easy access to large amounts of historical Automatic Identification System (AIS) data. Every day, AIS provides hundreds of millions of messages on a global scale MarineTraffic [2016], which contain the identifier of the ships, their coordinates from the Global Positioning System (GPS), speed, course, etc. In many areas, this data is freely available and may be collected into large amounts of historical maritime trajectories for maritime surveillance.

The potential limitations of clustering in constructing sufficient normalcy models

and the large amount of AIS data available raise the question whether realistic normalcy models could be learned by deep learning techniques. Deep learning applies very large neural networks, that were initially designed to resemble human brain functions, and scales very well with large amounts of training data. Deep learning methods have been used for representation learning for clustering tasks later down the pipeline Protopapadakis et al. [2017], Yao et al. [2017], trajectory prediction Forti et al. [2020], Capobianco et al. [2021b], Spadon et al. [2022], and automated detection of maritime abnormalities [Nguyen et al., 2021, Singh et al., 2021]. Thus, the potential of AIS-based deep learning methods to support surveillance operators is clear. Although automated decision support may not be warranted in all situations, it represents a clear way to test the suitability of the estimated normalcy model. Additionally, automated detection of maritime abnormalities may be used to identify situations which require additional development of systems for enhancing situation awareness.

1.2 Research Objectives and Contributions

This thesis aims to provide an initial study on the feasibility and evaluation of deep learning techniques to construct maritime normalcy models and for automated detection of maritime abnormalities. We study deep learning architectures from an operational point of view and evaluate them based on how they tackle known requirements for the practical application of abnormality detection. This will be discussed in more detail in section 1.3. The hypothesis is that deep learning techniques applied to the analysis of maritime trajectories can be used to predict future actions, intention, or to detect abnormalities that indicate dangerous, suspicious, or malicious behavior. In order to evaluate this hypothesis, the PhD project will cover the following topics:

A) Research, develop, and evaluate deep learning frameworks for describing normal maritime behavior using historical AIS data to support maritime situation awareness of surveillance operators.

B) Development of deep learning methods for the automated detection of maritime abnormalities.

C) Study the generalization of deep learning methods for the automated detection of maritime abnormalities over time and across geographical areas.

D) Methods to describe abnormal local behavior that is not restricted to major shipping lanes.

In line with the main topics of the thesis, the research outcomes evaluate deep learning methods based on historical AIS data to improve maritime security and safety. Focusing on situation awareness, methods are discussed both from the perspective of automated detection of maritime abnormalities and the level of interpretation that models may provide related to the detection of maritime abnormalities. The following research contributions are considered to be provided by the research:

1. Methodology to evaluate deep learning methods for enhancing situation awareness, including the publication of an annotated data set of maritime abnormalities.
2. Ensemble methods for automated detection of maritime abnormalities that detect different types of abnormal behavior and outperform the current state-of-the-art.
3. Analysis of the interpretability of state-of-the-art deep learning models and illumination of the limitations of the current state-of-the-art.
4. A multi-step clustering approach to disentangle positional and kinematic information resulting in a better description of behavioral patterns in a large ROI.

1.3 Outline of the Thesis

This thesis is separated into two major parts. In Part I, the methodology and context of the thesis are presented and consists of chapters 2-3 and Part II presents the research outcomes first as a summary, Chapter 4, and conclusion, Chapter 5, for those looking for a brief overview and chapters 6-9 corresponds to the paper contributions prepared during the PhD.

Chapter 2 introduces the concepts of situation awareness in relation to the detection of maritime abnormalities. We then present a detailed review of systems and models for automated detection of maritime abnormalities. We argue that rule-based systems are incapable of sufficiently describing the complexity of maritime behavior, even though they make up the majority of systems in use at the moment. Based on previous reviews of the field, we identify five main issues that data-driven methods for automated maritime abnormality detection must address. We then discuss the most common methods for expressing the normalcy picture from the point of how the five main issues are addressed.

Chapter 3 presents the deep learning models investigated in this thesis. The basic concepts of AutoEncoders and variational inference are presented, and we introduce recurrent neural networks (RNNs) for the modeling of sequential data. These two concepts are first combined into the Recurrent Variational AutoEncoder in which an entire trajectory is encoded into a single latent variable z. We then present the Variational Recurrent Neural Network (VRNN) which introduces a variational AutoEncoder into every time update in an RNN. The result is a model in which the dynamic information is modelled using the recurrent hidden states, h_t, and the random effects of the environment is modelled by a sequence of stochastic latent variables, z_t. Lastly, we discuss how to detect abnormalities using deep neural networks.

Chapter 4 summarizes the research presented in the papers contained in this thesis. We discuss how the results relate to the research objectives and contributions presented above and relate the results to the theory of situation awareness and how the methods may find use in practice.

Chapter 5 summarizes the main contributions of this thesis, discusses open questions, and directions for future work.

Chapter 6 contains the paper Olesen et al. [2022c]. This paper presents a review of deep learning methods for analysis of maritime trajectories and relates them to the five main issues identified by previous reviews. The paper provides a comparison of four deep learning architectures for the unsupervised detection of maritime abnormalities and evaluates their performance on a data set with annotated abnormalities related to a collision accident. Based on the review and comparison of deep learning techniques, the paper presents a set of guidelines for future research on deep learning models for the detection of maritime abnormalities.

Chapter 7 contains the paper Olesen et al. [2022d] . This paper follows up on some of the ideas presented in Chapter 6. First, different pre-processing techniques are evaluated for the purpose of abnormality detection. An alternative temporal training strategy is explored to create models generalizable to different resampling periods. Ensembles of different model architectures, objective functions, and resampling periods are proposed and evaluated. Finally, transfer learning is explored to transfer learnt normalcy models to different regions of interest (ROI).

Chapter 8 contains the paper Olesen et al. [2021]. This paper investigate the physical interpretability of a state-of-the-art model for automatic detection of maritime abnormalities. The training objective is modified with a static consistency loss and an ElasticNet loss to induce information in the stochastic latent variables. The latent variable are clustered to investigate the physical information encoded in the latent layer.

Chapter 9 contains the paper Olesen et al. [2022a]. Based on the findings regarding the information encoded in the latent space of deep AutoEncoders in Paper III, this paper suggests a two-step clustering method to disentangle positional and kinematic behavior. Disentanglement of these features allows for a more detailed normalcy model in terms of local kinematic behavior. Feedback from surveillance operators highlight local kinematic behavior as indicative of abnormal behaviour. Trajectories are first clustered on the basis of positional similarity. Each positional cluster is then further clustered on the basis of kinematic similarity.

Part I

Methodology and Context

Chapter 2

Maritime Abnormality Detection

In this chapter, we present a description of maritime abnormalities and methods for detecting them. In the first section, the concept of situation awareness and the relation to maritime anomaly detection is presented. We then discuss maritime abnormalities and their detection using rule-based vs. data-driven methods. Finally, we discuss the most common ways of constructing the normal maritime picture and how they address the main issues identified with automated methods for maritime abnormality detection.

2.1 Situation Awareness

The term situation awareness was first coined during World War 1 [Endsley, 1995] where it was recognized as an important tool for military pilots. To this day, it still remains an important concept of military operations, but has also found use in several civil domains such as traffic modeling, medical decision making, operation of heavy machinery, etc. Endsley [1995] provides a general definition of situation awareness:

"Situation awareness is the perception of the elements in the environment

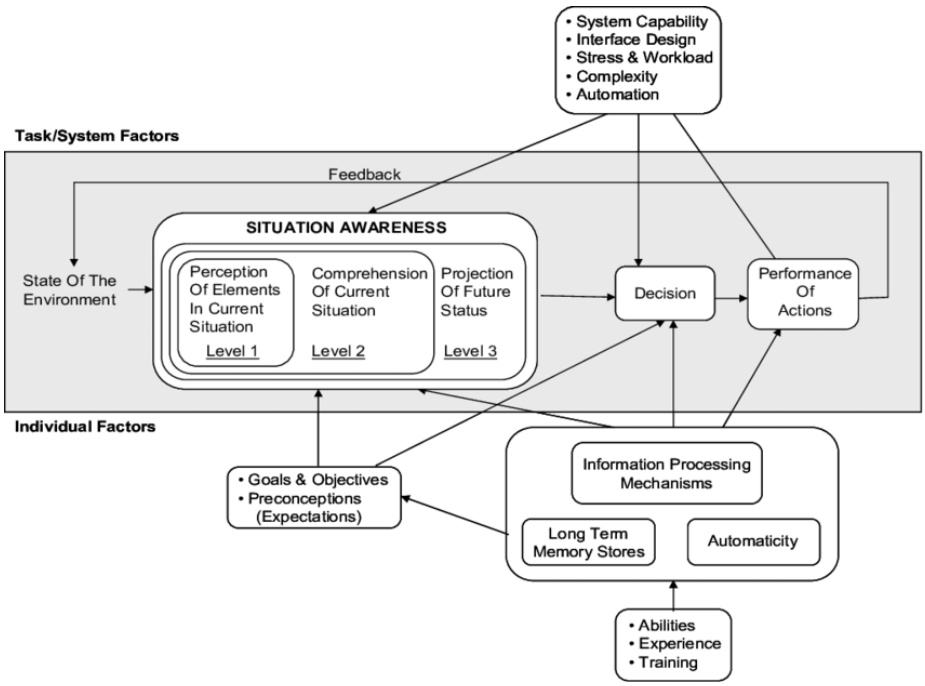


Figure 2.1 – Model of situation awareness in dynamic decision making. Adapted from Endsley [1995] by Salerno et al. [2022]

within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future"

Situation awareness may be interpreted as the perception and understanding of your nearby surroundings and prediction of events in the near future. This is key in dynamical decision making, where accurate predictions of the environment's response to your actions are key to fulfilling your goals and objectives. A diagram of dynamical decision making and the role of situation awareness is shown in Figure 2.1. The perception and analysis of the environment forms the basis for decision and action to achieve the objectives of the operator. As such, only when full situation awareness has been achieved can we make successful decisions and implement the required actions. Thus, situation awareness serves as the basis for military Command and Control operations and has been equated with the observe and orient phases of the observe-orient-decide-act (OODA) loop [Grant and Kooter, 2005].

Endsley [1995] decomposed situation awareness into three levels shown in Figure 2.1.

1. Perception of the elements in the environment.
2. Comprehension of the current situation.
3. Projection of the future status.

These levels increase in complexity, and each level serves as a requirement for the following level. The first level of situation awareness is associated with the perception of the environment, the second level the analysis of the environment and the third level is the prediction of the future environment. The primary task of surveillance operators monitoring maritime safety and security is to detect unexpected deviation from normality, i.e. abnormalities. Since these abnormalities are discovered in real time, the problem of maritime abnormality detection can be considered as a contribution to level 2 situation awareness. In relation to maritime abnormalities, level 3 situation awareness consists of predicting the impact on maritime safety and security if the abnormality is ignored and allowed to continue.

Several technologies and products have been develop for level one automation support within the domain of maritime surveillance. Data on surroundings is collected from a wide range of sensors like radar, AIS, sonar, weather, tactical data links, open source intelligence, satellite imagery, and communication systems. The data is gathered, fused, and presented in a concise and easily accessible, typical graphically, manner. Recently, computer vision techniques have been utilized to detect and classify objects, thereby contributing to level 1 situation awareness, Freitas et al. [2019], Chang et al. [2019], Mazur et al. [2017]. However, while much attention has been focused on automation support for level 1 situation awareness, automated systems for maritime abnormality detection to support level 2 situation awareness remain a challenge. Human surveillance operators monitor the maritime safety and security of large maritime regions, including charts of all vessel movements and taking into account a wide range of sensory input mentioned above. Despite visual aids, human operators can easily overlook abnormalities due to cognitive overload, time pressure, fatigue, and uncertainty due to the complex and heterogeneous nature of the data. This makes automated systems for maritime abnormality detection a key ingredient to support level 2 situation awareness.

2.2 Maritime Abnormalities

The situation awareness of maritime abnormality detection by experienced human operators relies heavily on domain knowledge and familiarity with vessels,

shipping routes, rules and regulations, the weather, ect. As such, a precise definition of maritime abnormalities is very difficult to make, as this would depend on deterministic factors such as location, time period, ship types, etc. as well as random factors such as weather, local traffic picture, and noisy sensory input. Additionally, maritime abnormalities can range from abnormal behavior due to weather or local traffic to severe accidents or illegal activities that impact maritime safety and security.

Designing a rule system for maritime abnormality detection capable of encompassing all maritime knowledge would be immensely difficult. First, the inclusion of new data sources for level 1 situation awareness requires designing new rules, which have to be consistent with the predetermined rule base. Secondly, studies are required to determine the applicability of rule sets in both time and geographical location. This requires a system to constantly evaluate the applied rule base to identify rules producing false positives, as well as new potential rules to include. Additionally, relocation of the system to another geographical location may require the design of completely new rules. Both of these are difficult and computationally heavy if the number of data sources increases and may make a rule system infeasible for general use. Additionally, even if such comprehensible rule systems could be designed, it would have difficulty dealing with the uncertainties of the real world.

Alternatively, data-driven approaches derive a model to represent the normal picture and assess anomalies compared to the normal picture. However, compared to human operators, data-driven approaches may struggle to understand the context of maritime situations. As such, the best solution seems to be designing automated data driven systems for detection of maritime abnormalities and to leave the interpretation of the situation and decision regarding the appropriate operational response largely up to the human operator.

In the work of data-driven automated abnormality detection there are generally three ways of defining abnormalities for training and evaluating models; Simulated abnormalities, self annotated abnormalities, and unsupervised anomalies. Self-annotated or simulated data sets may be trained purposefully for abnormality detection of well-defined abnormal behavior using supervised learning. However, it is important that the models are trained using meaningful abnormalities of operational interest and while extreme values or observations with added noise are outliers compared to the training set they might not define abnormal trajectories of operational interest to surveillance operators.

In unsupervised models, a normalcy model is learned on the training set. Then it is assumed that abnormalities are poorly modeled by the normalcy model and can be detected as outliers using this model. One drawback of unsupervised methods is that they flag statistical outliers as abnormal. This means

that they can potentially learn to model abnormalities of operational interest if they occur frequently causing false negatives. Thus, evaluation and comparison of unsupervised methods is an important issue for them to find practical application. In chapter 6 we discuss evaluation of deep unsupervised methods in more detail and provide a comparison of models previously suggested in the literature.

2.3 Maritime Abnormality Detection

The previous research on maritime abnormality detection has been summarised in several recent reviews; [Riveiro et al., 2018, Yan and Wang, 2019, Sidibé and Shu, 2017, Zhang et al., 2020a]. They all identify some of the same trends and future challenges for maritime anomaly detection presented below. As mentioned in the previous section, the main approaches to maritime anomaly detection can be separated into rule based and data-driven approaches. Rule-based approaches look for predefined patterns in the data source while data driven approaches derive a model for representing the normal picture and evaluate anomalies compared to the derived normalcy picture. In the next section, we shall discuss some of these models in more detail and their relation to the challenges in the following.

Evaluation:
Most works approach the problem of anomaly detection as an unsupervised problem. As a result, it is very difficult to compare the performance of various methods due to a lack of established, publicly available, annotated data sets for performance evaluation. However, annotated data sets are difficult to obtain since most surveillance operations are conducted by the military or law enforcement. Therefore, information related to the detection of abnormal behavior is often classified for security purposes. Additionally, many works limit their data to either small geographical areas or a few different ship types. As such, the training data may not be a true representation of the expected real-life data, and the generalization of any trained model may be questionable.

Incorporation of External Features:
Most works in the literature are centered on historical AIS data, and most approaches use only kinematic information from AIS messages to build a normalcy model. However, static factors such as ship type, and ship size as well as contextual factors such as weather, traffic density or regional regulations also dictate the behaviour of maritime vessels. As such, integrating and combining multiple sources of data may help improve situation awareness and reduce the risk of false alarms. Additionally, Riveiro et al. [2008] remark that there is an

observed lack of studies on feature extraction and finding relevant features in high-dimensional maritime datasets. However, due to evaluation issues, it is very difficult to evaluate the impact of additional external features.

Zhang et al. [2020a] and Yan and Wang [2019] also raise an issue regarding the quality of the AIS data in particular. Several parts of the AIS system require manual data entry, which may be prone to manual errors, and not all ships are equipped with or properly operate the AIS system. Additionally, basing automated abnormality detection models solely on AIS data makes the detection prone to illicit interference with the AIS signal such as on/off switching or jamming [Hovgaard, 2022]. Therefore, combining different sources of tracking such as radar or satellite imagery could help to refine the maritime picture.

Online Detection:
All reviews identify a lack of methods for online detection in real time. Many works are based on an assumption that the abnormality detection is performed after the entire trajectory of the vessel has been observed. This is a critical limitation in any real-life application, as it will delay detection until after the abnormal event has ended. Thus, in order to enhance situation awareness for maritime safety and security, the research should focus on methods capable of performing real time or near real time detection.

Scaleability:
As mentioned many previous works contain only a limited data sets. This is closely related to the scalability issues of most suggested methods. Typical approaches such as Kernel Density Estimators (KDE) and clustering methods are very sensitive to hyper-parameters and require time-consuming tuning of these hyper-parameters. Furthermore, the time complexity of the clustering methods increases quadratically with the size of the training data, while KDE increases with the product of the sizes of the training and evaluation sets. Osekowska et al. [2017] finds significant levels of seasonality in the maritime traffic patterns which might require frequent re-training of automated abnormality detection models. If no data reduction strategy is used, models for automated abnormality detection may become infeasible due to the size of the training sets.

Interpretation of alarms:
Riveiro et al. [2008] discuss the end-user aspects related to maritime abnormality detection systems and find that little work has focused on specifying user requirements for such systems. Riveiro et al. [2008] argue that in order to build system trust and support user understanding of the system's inner workings and limitations, detection systems needs to be transparent and interpretable for operators. In particular, Riveiro et al. [2008] highlight issues such as the insertion of expert knowledge into the anomaly detection process, the interpretability of

the anomaly detection process, providing explanations for detections to operators, finding efficient human–machine collaborations, how to initialize parameters related to the detection process and how operators can update and improve normalcy models. According to Riveiro et al. [2008] development of abnormality detection systems has focused primarily on technological challenges. In order to improve abnormality detection performance, Riveiro et al. [2008] suggests greater involvement of end users in system design and a focus on efficient human-machine collaboration for abnormality detection.

2.3.1 Rule Based Detection Methods

As discussed previously, a comprehensive set of rules for abnormality detection would be practically impossible to design due to scaling issues as the number of data sources increases. However, the results are very easy to interpret and analyze, since rule-based systems may output which rules were broken and clearly show operators the reason for an alert in real time. This makes the majority of automated systems for maritime abnormality detection in use today apply rule based detection on a small scale limiting the ROI and abnormalities detected [Thomopoulos et al., 2019, Neves et al., 2019].

The OCULUS sea forensics toolbox [Thomopoulos et al., 2019] uses a combination of rules and fuzzy inference suggested in Rizogiannis and Thomopoulos [2019] to perform detection of several different forms of anomalies; Gap in reporting, speed change, fake MMSI, rendezvous incidents, and collision notification. To detect reporting gaps, speed changes, and MMSI spoofing, the system mines all AIS messages in an area every t seconds and sends an alert if it detects any of the above things based on user-defined thresholds. To perform detection of rendezvous incidents, and collision notification they utilize a fuzzy inference system. Positions reported from AIS messages are used to calculate the distance to closest point of approach, the time to closest point of approach and variation of compass degree. This information is used together with the current time of day and geographical position as input to the fuzzy inference system. They suggest a filtering mechanism to filter out anchored ships and ships too far away from each other before the fuzzy inference system. The system is designed to achieve a high degree of interpretability by translating abstract numbers into easy-to-understand linguistic rules, and almost all parameters are adjustable by operators. However, the system fails to account for contextual and external information. The design of fuzzification functions and rules is not discussed, and as the number of data sources increases, the fuzzification mechanism and inference rules become increasingly difficult to design.

The MARISA project [Neves et al., 2019] uses similar rules for online detection of

signal loss detection, abnormal change of direction, or abnormal change of speed based on AIS messages. They also conduct offline analysis of AIS messages to verify navigational status, perform rendezvous detection, and verify the reported position using the previously reported position of the vessel.

Another popular approach to express rules is Markov Logic Networks. Lauro Snidaro et al. [2012] suggests the use of Markov logic to incorporate contextual information into detection. Markov Logic Networks consist of a set of rules explaining the interaction between parameters known as the knowledge base. Each rule in the knowledge base is assigned a weight, and the probability that an event x occurs is proportional to the exponential of the sum of the weights multiplied by the number of times each rule is upheld during the event.

$$P(X = x) \propto exp\left(\sum_{i=1}^{|L|} \omega_i n_i(x)\right) \tag{2.1}$$

where $|L|$ denotes the cardinality of the knowledge base, ω_i the weight of the i-th rule, and $n_i(x)$ the number of times the i-th rule is upheld during the event. The knowledge base can consist of both of observable context and prior determinable contextual information. Observable context may be knowledge about heading, route, or intelligence reports. Prior determinable contextual information is related to knowledge about zoning of waters and shipping lanes ect. Snidaro et al. [2015] extends the method to handle complex events made up of multiple simple events through the combination of rules. The knowledge base for rendezvous detection used by Snidaro et al. [2015] is shown in Table 2.1. The complex event *rendezvous* depends on the level of suspicion we have of the participating vessels and the proximity between them. Rule 17 defines a *rendezvous* abnormality as two vessels in the same area during overlapping time intervals. Rule 16 applies further weight to this *rendezvous* abnormality if the vessels are flagged as suspicious by operators. Rule 18 specifies that *rendezvous* is not possible if either vessel has not stopped. Similarly, rule 19 states that *rendezvous* has not happened when the two vessels are in the same area in non-overlapping time intervals. The level of suspicion and proximity we assign to vessels has their own rules based on simple observational or contextual knowledge. Snidaro et al. [2015] argues that in chains of rules, the weights acts as an uncertainty measure. A strength of Markov logic is the ability to be applied even when data sources are missing and that output probabilities are associated with understandable events, allowing operators to quickly assess the level of risk. Another key feature is the possibility to query possibilities for arbitrary events which allows operator to build on-the-fly queries required by the certain situations. Like for fuzzy inference the inclusion of many data sources makes the knowledge base increasingly difficult to design and a potential inconsistent knowledge base will cause global inconsistencies and affect the inference of the system.

#	Rule	Weight
1	overlaps(v,y) ⇔ overlaps(y,v)	ω
2	meets(v,y) ⇔ meets(y,v)	ω
3	proximity(v,y) ⇔ proximity(y,v)	ω
4	rendezvous(v,y) ⇔ rendezvous(y,v)	ω
5	stopped(v) ∧ (isIn(v,openSea) ∨ isIn(v,intWaters)) ⇒ suspicious(v)	$^4/_5\ \omega$
6	stopped(v) ∧ (isIn(v,harbour) ∨ isIn(v,nearCoast)) ⇒ ¬ suspicious(v)	$^2/_5\ \omega$
7	¬ AIS(v) ⇒ alarm(v)	ω
8	¬ insideCorridor(v) ⇒ suspicious(v)	$^4/_5\ \omega$
9	humint(v,smuggling) ⇒ suspicious(v)	$^3/_5\ \omega$
10	humint(v,clear) ⇒ ¬ suspicious(v)	$^1/_5\ \omega$
11	suspicious(v) ⇒ alarm(v)	ω
12	¬ suspicious(v) ⇒ ¬ alarm(v)	$^1/_5\ \omega$
13	isIn(v,z) ⇒ (z ≠ zp) ∧ ¬ isIn(v,zp)	ω
14	isIn(v,z) ∧ isIn(y,zp) ∧ (z ≠ zp) ⇒ ¬ proximity(v,y)	ω
15	¬ proximity(v,y) ⇒ ¬ rendezvous(v,y)	ω
16	suspicious(v) ∧ suspicious(y) ∧ (overlaps(v,y) ∨ meets(v,y)) ∧ proximity(v,y) ⇒ rendezvous(v,y)	ω
17	(overlaps(v,y) ∨ meets(v,y)) ∧ proximity(v,y) ⇒ rendezvous(v,y)	$^1/_5\ \omega$
18	¬ stopped(v) ∨ ¬ stopped(y) ⇒ ¬ rendezvous(v,y)	$^3/_5\ \omega$
19	before(v,y) ∧ proximity(v,y) ⇒ ¬ rendezvous(v,y)	$^4/_5\ \omega$

Table 2.1 – Knowledge base suggested by Snidaro et al. [2015] for incorporating contextual information and uncertainty using Markov Logic Networks for rendezvous detection.

2.3.2 Data-driven Detection Methods

Data-driven approaches generally try to construct a model to represent the normal picture and evaluate anomalies compared to the normal picture in terms of the trained model. Compressing the normal maritime picture may increase interpretability, as operators can easier relate to an idea of average behavior compared to a large ensemble. The most common way to express the normal maritime picture is by clustering [Pallotta et al., 2013b, Zhao and Shi, 2019b]. Other popular approaches include stochastic processes [D'Afflisio et al., 2018, Forti et al., 2018, Kowalska and Peel, 2012], graph-based approaches [Venskus et al., 2019, Osekowska et al., 2017, Varlamis et al., 2019], non-parametric methods [Laxhammar and Falkman, 2015, Smith et al., 2014a, Anneken et al., 2015], or predictive/reconstructive models [Nguyen et al., 2021, Singh et al., 2021]. Some skip this step and train a classifier directly on the available data using supervised learning [Singh and Heymann, 2020a, Sfyridis et al., 2013].

2.3.2.1 Supervised Learning

Supervised models applied directly on raw data points include decision trees [Bombara et al., 2016], neural networks [Singh and Heymann, 2020a, Liu et al., 2022a], and Support Vector Machines (SVM) [Handayani et al., 2013, Sfyridis et al., 2013].

Bombara et al. [2016] suggests the use of decision trees trained using coordinate trajectories to classify normal and abnormal behavior. The training data is simulated to represent two abnormal cases, i.e., smuggling and pirate/terrorist activity. The simulated trajectories have clear differences between normal and abnormal behavior, making detection trivial for a human operator. Thus, the contributions of the system is debatable and the performance when exposed to real life trajectories and abnormalities remains an open question.

Handayani et al. [2013] and Sfyridis et al. [2013] both train SVMs on data extracted from historical AIS messages such as speed, location, course, heading and timestamp. Handayani et al. [2013] use a self-labeled data set and obtain a 91.63% accuracy using the raw AIS data. Sfyridis et al. [2013] focus on migrant carrying vessels in the Mediterranean. Since only a few vessels have been labeled as migrant carrying by national coast guards upon inspection, they use a One-Class SVM. With this algorithm, they identify another 9 vessels which show similar behavior to the vessels previously found to be smuggling refugees.

Several works have applied shallow or deep neural networks to directly classify trajectories as normal or abnormal using sequences of location, speed, and course. Wang [2020] train a feed-forward neural network to predicts the probability of abnormality directly using simulated labels obtained by adding noise. Singh and Heymann [2020a] suggests a model for detection of AIS on/off switching. Incoming AIS messages are re-sampled to every two seconds and missing data is considered as abnormal cases of on/off switching and trajectories with continuously missing data are labeled as abnormal. Singh and Heymann [2020b] finds neural networks outperform traditional machine learning methods such as; SVMs, K-Nearest neighbors, Decision Trees, etc. on detection of AIS on/off switching. Liu et al. [2022a] and Hu et al. [2022] self-annotate the dataset based on extreme position, speed, or course values decided using the COLREGS international collision avoidance rules for the area. Zhang et al. [2021b] cluster the data using density-based clustering and propose a Long-short Term Memory (LSTM) model to classify abnormal trajectories identified as outliers in the clustering.

Supervised learning using self-annotated or simulated data sets may be trained purposefully for abnormality detection of well-defined abnormal behavior, and evaluation of their performance is trivial. Additionally, depending on the choice of model (eg. deep sequential neural networks) the may also allow for incorporation of additional features and online detection of abnormalities. However, as discussed previously large annotated data sets for training purposes are difficult to procure. Although extreme values or observations with added noise are outliers compared to the training set, they may not define abnormal trajectories of operational interest to surveillance operators. In addition, the process of manual annotation or the annotation using clustering suffers from scaleability

issues, negating the potential modeling power of deep neural networks applied to very large data sets.

2.3.2.2 Stochastic Processes

Several papers treat normal maritime navigation as a stochastic process. Millefiori et al. [2016] and Vivone et al. [2017] argue that vessel dynamics in open seas is well described by an Ornstein-Uhlenbeck stochastic process. Ornstein-Uhlenbeck stochastic processes closely resemble constant-velocity processes; however, in Ornstein-Uhlenbeck processes, the velocity is allowed to drift around a long-term mean. Ornstein-Uhlenbeck processes have been used for trajectory forecasting [Pallotta et al., 2014, Uney et al., 2019], route deviations [Forti et al., 2018, 2019], and detection of AIS on/off switching [D'Afflisio et al., 2018, Braca et al., 2018]. For route deviations, the switching between normal and abnormal behavior is modeled as an unknown Bernoulli velocity set that is either empty under normal behavior or a singleton under abnormal behavior. Using the trajectory extracted from AIS messages or recorded using radar, the probability of this set being a singleton can be derived and used as the anomaly score.

The parameters of the Ornstein-Uhlenbeck process can be estimated on the fly for each trajectory, or estimated using historical data. Estimating the parameters from using historical data may provide valuable contextual knowledge. Pallotta et al. [2014], Uney et al. [2019] make a position prediction model based on vessel kinematics that is described by an Ornstein-Uhlenbeck process. Pallotta et al. [2014] estimate the long-term mean velocity, revert rate, and noise from historical AIS tracks. Density based clustering [Pallotta et al., 2013b] is used to extract routes from historical AIS data and the parameters of the Ornstein-Uhlenbeck process is estimated using the trajectories belonging to the same route as the current vessel. Having defined the Ornstein-Uhlenbeck process, long-term predictions of the position up to seven hours ahead is made. The prediction model is evaluated on a few test cases of different ship types and the variance of the prediction error is found to grow linearly with the prediction horizon. The same principal idea is applied in Uney et al. [2019]. Historical AIS data is classified into several classes based on the start and ending positions. For each class the number of changes in the route is determined by a GMM, in which the number of components equals the number changes along the route. This also splits the class trajectories into sections in which the vessel kinematics is modeled by an Ornstein-Uhlenbeck process. The posterior probability density of the future vessel position is approximated by performing the marginalization over the kinematic parameters of all classes using the Monte Carlo approximation. The algorithm is tested on three trajectories initiating from a position where several different routes converge, and it correctly identifies the split in

routes observed in the historical data.

D'Afflisio et al. [2018], Braca et al. [2018] suggest a model to detect intentional AIS on/off switching based on Ornstein-Uhlenbeck processes. D'Afflisio et al. [2018] derive a test statistic to evaluate whether the vessel kinematics between two AIS contacts, which might be widely separated both spatially and temporally, can be described as an Ornstein-Uhlenbeck process. The process is evaluated using a single case from real life and is found to both flag abnormal behavior when observations are present as well when observations of the abnormality are not present. Braca et al. [2018] extends this approach to include situations in which secondary detections of the vessel are made between the two points of interest. These secondary detections could stem from radar measurements or reported sightings from human observations of the vessel. The added value of the extra detections is not investigated in a real world setting, but their effect on the test statistic is tested on simulated trajectories.

Kowalska and Peel [2012], Smith et al. [2012], and Smith et al. [2014b] suggest Gaussian Processes to model vessel kinematics. Kowalska and Peel [2012] models the velocity of ships as function of their physical position. Due to limitations in the Gaussian Process they only consider velocities where both components are positive. A model of normal vessel behaviour is created from historical AIS data and then tested on simulated anomalies by moving an observation to another position as well as some test cases from the English Channel in which movements where reenacted to show sign of smuggling and terrorism. The model correctly classifies the test cases as abnormal, and on the simulated abnormality, they report a global average accuracy of 80%. However, in regions where normal and abnormal behavior is present, the accuracy is very poor. They suspect this is due how the abnormality is simulated. When an observation is moved along the shipping lane, the velocity may still be normal but is labeled abnormal.

Smith et al. [2012, 2014b] also suggest using Gaussian Process in combination with extreme value theory to perform sequential anomaly detection. Smith applies a sequentially updating Gaussian Process in order to predict the distance of the next track update from the origin based on the sequence of previous distances. Extreme value theory estimates the distribution of the maximum value observed in a sequence of stochastic variables and can be used to define a threshold for abnormal observations. If future observations are within the threshold for the predicted mean determined by extreme value theory it is considered normal and is included in the next model update. Smith reports an AUC-ROC of 0.8032 using simulated anomalies which could indicate the model struggles with false positives or there are some anomalies that can not be predicted. The models are also evaluated on real life test cases. The model correctly identifies a single abnormal trajectory but also flags all trajectories of sailing vessels as abnormal since these trajectories do not exhibit common movement characteristics.

Models based on stochastic process make underlying assumptions regarding normal vessel dynamics. These assumptions must be reflected by the underlying stochastic process or in the choice of kernel function for Gaussian Processes. Stochastic processes may be applied to large data sets, where the parameters are estimated sequentially for each trajectory. The parameters of the processes may also be estimated from historical data; however, this often requires some form of data clustering, which scales poorly with the size of training data. As seen in Smith et al. [2014b] the underlying assumptions about vessel dynamics may not hold for all types of vessels, which could cause issues when applied in a real-life setting. Additionally, the models have generally not been evaluated on a large noisy data set representative of real-life data.

2.3.2.3 Clustering

The most common way of expressing the normal maritime picture is using clustering. Several different clustering algorithms have been applied for clustering of trajectories. Methods such as K-means, Klaas et al. and K-medoids, Zhen et al. [2017] have been suggested in collaboration with similarity measures. However, density-based clustering techniques have long been the predominant approach to data mining within maritime trajectory analysis. Most clustering methods are based on the Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [Ester et al., 1996]. An important feature of DBSCAN is that it directly identifies outliers whose local density is lower than a predefined threshold. This is used in Sun et al. [2016] to identify abnormal AIS updates based on the reported location. Pallotta et al. [2013b] suggests the widely used Traffic Route Extraction and Anomaly Detection (TREAD) framework. TREAD first clusters observations to form waypoint objects in the ROI. These waypoint objects are classified as stationary points, entry points, or exit points within the ROI boundary box. A route between waypoints is then formed whenever a certain number of transitions between them have been observed.

The obtain normalcy model may then be used for anomaly detection. In Pallotta et al. [2013b] the observed trajectory of a vessel is described by a sequence of state vectors that contain information on position and velocity extracted from AIS messages.

$$V = \{v_1, v_2, ..., v_T\} \tag{2.2}$$

$$v_t = [x_t, y_t, \dot{x}_t, \dot{y}_t] \tag{2.3}$$

The trajectory of the vessel is associated with a time series of regions denoted as the corresponding temporal state sequence $S = \{s_1, s_2, ..., s_T\}$. These regions are identified by circles of radius, d, centered in the observed positions $[x_t, y_t]$,

and used to describe the local route behavior. It is clear that the radius, d, of the associated states s_t is an important hyperparameter. If d is too small, the behavior of the local route would be based on a reduced number of neighbors, leading to poor generalization. On the other hand, if d is too large, the behaviour would be biased by the mixing of different behaviors. The extracted routes and their sequence of local behaviour can be used for anomaly detection purposes. Pallotta et al. [2013b] suggest a simple probability threshold detection. The maximum probability of the observed vessel track and associated temporal state sequence conditioned on the k-th possible route is used as a test criterion.

$$\underset{k}{\operatorname{argmax}}\, P(V, S|R^k)P(R^k) \underset{H_1}{\overset{H_0}{\gtrless}} T_h \tag{2.4}$$

where V denotes the observed trajectory of the vessel of interest, S the corresponding temporal state sequence, and R^k the k-th possible route. If the test criteria is greater than a user-defined threshold, the behavior falls under the established normality model. However, if the value is smaller, the behavior is flagged as abnormal. Pallotta et al. [2013b] suggest the probability of observing positional and velocity components $(x_t, y_t, \dot{x}_t, \dot{y}_t)$ within the state s_t, i.e. $P(V, S|R^k)$ is estimated using non-parametric methods such as the Kernel Density Estimator. The anomaly detection is not evaluated, but the usability is shown through a real-life example of a double u-turn in the shipping lane.

Several extensions of the TREAD framework have been suggested. Pallotta and Jousselme [2015] associate several features to the routes extracted using TREAD. Each route is associate with an average route computed as a sequence of waypoints $(x_{WP_t}, y_{WP_t}, \dot{x}_{WP_t}, \dot{y}_{WP_t})$, where x_{WP_t} and y_{WP_t}, are related to the coordinates of an ideal vessel moving along the "average route" and passing by that specific waypoint. The velocity components $\dot{x}_{WP_t}, \dot{y}_{WP_t}$ are derived by combining the median speed over ground and the course over ground of the waypoint. Pallotta and Jousselme [2015] proceed to perform anomaly detection in a hierarchical manner. First, the positional distance between tracks and the discovered average routes are computed. Tracks are then either flag as being off-course abnormalities or assigned to a single route. The kinematic distance between the course over ground and speed over ground features and the kinematic values of the average route is then calculated and large distances flagged as abnormal. Finally, the transition probability between route waypoints is modeled to detect abnormal transitions. The proposed hierarchical detection is evaluated using simulated abnormalities and has a precision of 87.3% for abnormal trajectories. In Jousselme and Pallotta [2015] different distances or relation measures between tracks and routes are studies in terms of their uncertainty representation capabilities. A similar approach is suggested in Arguedas et al. [2015], Varlamis et al. [2019]. To build a maritime traffic network, each average route is divided into segments according to the detected break points or turns as

described in Arguedas et al. [2014]. Abnormalities are then detected using the distance between the reported vessel position and the declared route Arguedas et al. [2015], transition probabilities, or behavioral statistics of the subsegment between two waypoints Varlamis et al. [2019]. Singh et al. [2021] suggest the generation of a graph network based on waypoint discovery using DBSCAN. Each trajectory update is assigned to an edge in the derived network, and this assignment is provided to a LSTM model as a contextual input.

The TREAD method and its derivatives are all based on a clustering of individual AIS updates not accounting for the temporal relation between them until the abnormality detection phase. The DBSCAN algorithm has also been used for the clustering of complete trajectories by computing trajectory similarities Yang et al. [2022b], Wang et al. [2021], Zhao and Shi [2019a]. First, the trajectories are simplified using the Douglas-Peucker (DP) algorithm Douglas and Peucker [2011]. The similarities are then computed using the Hausdorff distance or Dynamic Time Warping (DTW) before being clustered by DBSCAN. Wang et al. [2021] considers a hierarchical search over the hyperparameters of DBSCAN, which allows for groups with different densities, and helps to find clusters in sparsely populated geographical regions.

Recently, deep neural networks have been used for feature extraction to cluster trajectories [Yao et al., 2017, Murray and Perera, 2021, Protopapadakis et al., 2017]. Yao et al. [2017] and Murray and Perera [2021] use auto-encoders based on a recurrent neural network (RNN) to compress trajectories to a fixed dimension. Encodings in the latent space are clustered using k-medoids and hierarchical DBSCAN respectively and clusters are found to corresponds to the major shipping lanes in the ROI. Murray and Perera [2021] further use the discovered clusters to train neural networks for predicting the future position.

Protopapadakis et al. [2017] use stacked AutoEncoders in combination with density-based clustering (OPTICS [Ankerst et al., 1999]) to detect abnormalities in tracks from Over-The-Horizon radar measurements. At each time step, all tracked ships are projected into the new latent space using AutoEncoders and clustered using the OPTICS algorithm. Abnormalities are then detected as outliers flagged by the OPTICS algorithm. The proposed feature extraction is evaluated against raw data points or Principal Component Analysis using the Calinski–Harabasz Index, Davies–Bouldin Index and Silhouette values of the calculated clusters. AutoEncoder feature extraction is found to give more coherent and densely packed clusters. This causes outliers to be more well determined, and thus makes it easier to find abnormalities. Additionally, more outliers can be found when using AutoEncoders with around 2-3 outliers detected at each time step.

The clustering methods mentioned above only consider the positional input,

yielding clusters that mostly correspond to the primary shipping lanes. As discussed, Pallotta and Jousselme [2015] consider the speed and course separately in the abnormality detection stage, however, this information may also be incorporated into the clustering. Zhen et al. [2017] introduce the difference of the average course in their similarity measure and Liu et al. [2015b] extends the DBSCAN clustering models to consider not only the geographical distance of the coordinates, but also the difference in speed and course. This allows them to distinguish between shipping lanes in opposite directions and find speed differences within the main shipping lanes. However, the work is limited to small geographical areas and a limited number of ship types.

Previously, we argued that models should encourage human-machine interaction and that the interpretation of the situation and the decision regarding the appropriate operational response should be largely left to the human operator. This approach is studied in Radon et al. [2015]. Firstly, extraction of the normal movement patterns is done using a clustering of trajectory segments using the OPTICS algorithm. Any outliers flagged in this clustering is considered a potential abnormality. For all potential abnormalities Radon et al. [2015] then consider the direction of the wind, the speed of the wind, the speed of wind gusts, and the height of the waves at the location and time that the abnormal behavior was detected. A rule-based approach to explain abnormal behavior using the contextual features is then suggested. The proposed approach is verified using recorded AIS data from the entry of Vancouver bay. The data is split into 6 groups based on the total duration and each trajectory segment is labeled by a subject matter expert. The proposed contextual verification is found to significantly reduce the number of false alarms and increase precision in all groups.

Clustering approaches may be used to detect abnormalities in different ways. Density based clustering methods may directly flag trajectories found to be outliers [Radon et al., 2015, Wang et al., 2021] or clustering can be used to extract a sequence of state vectors representing the average cluster behaviour [Pallotta et al., 2013b, Liu et al., 2015b]. As discussed before, it is very difficult to compare these methods as they mostly only report a few qualitative examples or performance on simulated abnormalities. However, an important factor in the comparison is the issue of online detection. Methods based on trajectory similarities are often restricted to offline detection, whereas point-based methods allow for real-time detection. Pallotta et al. [2013b] use a sliding window that captures only the most recent points of the partially observed track to perform online detection. Clustering methods may provide more interpretability for operators as clusters of trajectories a generally easy to visualize and comprehend. However, as the time complexity of density-based clustering algorithm is squared w.r.t. the size of the training set, they are not applicable for large-scale use. Similarly, the inclusion of additional data sources in the clustering may

negatively affect the clustering performance due to sparsely populated clusters. However, as illustrated in Radon et al. [2015] a human contextual verification step can reduce the number of false alarms.

In Chapter 9 we propose a two-step clustering procedure that can disentangle positional and kinematic features. This method is intended to increase the user interpretability through a proposed man-in-the-loop detection scheme.

2.3.2.4 Graph-Based Approaches

As discussed previously, the waypoints discovered using density-based clustering may be extended to form a graph representation of the normal maritime picture Arguedas et al. [2015], Varlamis et al. [2019], Singh et al. [2021].

Another common way to represent the normal picture is by using a fixed grid representation. Lei [2016] suggests a grid-based clustering algorithm. The ROI is partitioned into grid cells of equal size and the number of trajectory segments passing through each cell is calculated. A cell is identified as a frequent region if the number of trajectory segments passing through exceeds a predefined threshold. After extracting the frequent regions, each raw trajectory is transformed into a sequence of frequent regions. Based on these sequences the transition probability and average kinematics behaviour for each transition is computed. Abnormality detection is then based on three metrics; spatial outlying score, sequential outlying score, and behavioral outlying score. The spatial outlying score is calculated using the probability of a vessel passing through the current frequent region, the sequential outlying score is calculated using the transition probabilities, and the behavioral outlying score by comparing the kinematic behaviour of the vessel to the typical kinematic behaviour using Gaussian Mixture Models. These three scores are combined into a single anomaly score that is used for detection.

Other grid-based clustering approaches have been suggested in Venskus et al. [2015, 2017, 2019], and Osekowska et al. [2014], Osekowska and Carlsson. Osekowska *et al.* applies a fixed geographical grid over the ROI while Venskus *et al.* learns the grid by training a self-organizing map. Both methods take inspiration from nature to perform outlier detection; Venskus *et al.* from ant pheromone navigation and Osekowska *et al.* from potential fields, respectively. Both methods assign a value to each node in the grid which increases whenever a vessel transitions to the node, but then slowly evaporates or decays over time. Venskus et al. [2015] flags abnormalities whenever an update is assigned to a node in which the pheromone level is less than a defined threshold, while Osekowska et al. [2014] computes a continuous potential as a weighted average of

the charge of all nodes.

Graphs have also been used to represent the relationship between different trajectories. Hu et al. [2022] suggest a graph variational AutoEncoder in which each node corresponds to a ship in the ROI and the adjacency matrix is calculated using trajectory similarity measures. A similar approach is suggested in Liu et al. [2022d] for the problem of trajectory prediction. Adjacency matrices are calculated based on the geographical distance, time to closest point of approach, and vessel size differences. A graph convolution network is then trained to predict the future position based on these adjacency matrices. Instead of modelling individual trajectories. Eljabu et al. [2021] detects abnormalities in traffic patterns on longer time scales of weeks or months. They propose a graph neural network where semantic stopping points in the ROI, e.g. ports, form nodes in a graph. The number of transitions between these nodes over a long period of time is then embedded using graph convolutions. The embeddings are assumed to be normally distributed, and outliers in this distribution are declared abnormal transition patterns.

Like with most other suggested models, evaluation of the abnormality detection performance of graph-based approaches is difficult and is often only evaluated using qualitative examples or simulated abnormalities. Venskus et al. [2017] evaluate the proposed method using expert-labeled information; however, the model is only tested on data from small harbor areas. Thus, it may not generalize well to larger areas with less restricted traffic.

Graph-based preprocessing may provide a similar level of intuitive detection as clustering methods due to the semantic interpretation of nodes in most graph approches. However, this intuition may begin to break down as additional features are added, as inVenskus et al. [2017, 2019]. As nodes become high dimensional, they will start to overlap in dense regions of the input space and only differ slightly for a few input features. Contrarily, nodes in sparse regions will have large differences in multiple input features and contain less graniulity compared to dense regions. Thus, we may see a trade-off between the interpretability of the graph and the number of features. Similarly, construction of the graph from data must be scaleable to large amounts of training data. Expensive similarity measures or clustering methods might not be useful for large amounts of training data.

Detection of abnormalities in most graph-based methods is based on transition probabilities Venskus et al. [2019], Varlamis et al. [2019] or behavioural statistics of the transition Lei [2016], Varlamis et al. [2019]. Both methods require a full trajectory sub-segment between nodes, which may delay detection in sparse areas where nodes are far between. Thus, to be useful in large ROIs, detection based on transition statistics should be avoided.

2.3.2.5 Non-Parametric Methods

Anneken et al. [2015] compare the abnormality detection performance of Kernel Density Estimators to Gaussian Mixture Models on an self-annotated data set of AIS messages from tankers and cargo vessels. Both methods are trained using only the points marked as normal and abnormality detection is based on the assumption that the log-likelihood for abnormal points will be lower. Anneken et al. [2015] conclude that both methods are prone to false positives and false negatives and argue that looking at the entire trajectory instead of single points could improve performance.

Laxhammar and Falkman [2011], Smith et al. [2014a] both suggest using K-nearest neighbors with the Hausdorff distance to detect abnormalities using conformal prediction. Laxhammar and Falkman [2011] focus on making an online classifier by considering a directed Hausdorff distance to calculate the similarity between incomplete trajectories. For each trajectory, the k-th highest similarity scores are saved and used for conformal prediction of new unseen trajectories. Smith et al. [2014a] compares non-conformity measures based on K-nearest neighbors and Kernel Density functions with a Gaussian Kernel. They find that K-nearest neighbours performs better when minimizing the number of false positives is important, but Kernel Density Estimators achieved an overall better performance. The methods were evaluated using a combination of real trajectories from vessel types not included in the training data and simulated abnormalities. Smith et al. extends proposed approach by conditioning the detection according to the type of ship. This is found to improve detection, finding more abnormalities and fewer false positives.

Laxhammar and Falkman [2012] propose to swap the K-nearest neighbours non-conformity measure to one based on Local Outlier Factor (LOF, Breunig et al. [2000]). LOF computes the local neighbourhood density around each data point and computes an abnormality score comparing the local neighbourhood density of each data point to that of its nearest neighbours. Laxhammar and Falkman [2012] suggests a sliding window to accommodate for possible incomplete trajectories and allow for online detection. For each full trajectory in the training data, this is done by finding the subsegment with the highest similarity with the incomplete trajectory under investigation.

A major drawback with conformal abnormality detection detection [Laxhammar and Falkman, 2012, Smith et al., 2014a] and Kernel Density Estimation [Smith et al., 2014a, Anneken et al., 2015] is that they require storage of and comparison with the complete training data. Therefore, it is computationally inefficient and scales very poorly to large training sets. Laxhammar and Falkman [2015] attempts to combat this issue using an inductive approach to calculate non-

conformity scores. The training set is split into the proper training set and a calibration set. The nonconformity score of the calibration set is then precalculated using the proper training set. Given a test sample the non-conformity score is calculated relative to the training set, and detection is based on comparison with the non-conformity score of the calibration set. With the proposed approach, one does not need to recalculate the non-conformity score of each previous data point; however, calculation of the non-conformity score for the test trajectory still may be infeasible for large training sets.

2.3.2.6 Predictive/Reconstructive Models

Predictive and reconstructive models are unsupervised approaches that detects abnormalities based on the predictive or reconstructive error. We previously mentioned SVMs applied for direct prediction of abnormality labels; however, they may also be applied for prediction of future positions for abnormality detection. Kim et al. [2019] use Support Vector Regression to learn the historical shipping lanes. The future position is predicted based on the learned shipping lanes, current position, heading, and speed. If the future position, heading, or speed deviates from the predicted values by a fixed threshold, the trajectory is classified as abnormal.

Deep neural networks have been especially prevalent as predictive and reconstructive models. In Chapter 6 we provide a more in depth review of deep learning methodologies used for analysis of maritime trajectories. In this section, we shall briefly discuss how deep learning methods tackle the five issues raised previously.

For trajectory prediction/reconstruction, the architecture most applied has been RNNs either as one-step predictions [Liu et al., 2021, Sørensen et al., 2022], direct multistep predictions [Chondrodima et al., 2022, Mandalis et al., 2022, Spadon et al., 2022], iterative one-step prediction [Forti et al., 2020, Capobianco et al., 2021b, Dijt and Mettes, 2020] in a sequence-2-sequence fashion or in a reconstructive AutoEncoder architecture [Nguyen et al., 2021, Hu et al., 2022, Murray and Perera, 2021]. Several different combinations of recurrent architectures have been suggested in the literature with different recurrence cells, unidirectional vs. bidirectional, context lengths, etc. Additionally, a few extensions to the basic RNN structure have been proposed. Spadon et al. [2022] suggest a hybrid solution where each recurrence is preceded by a 1-dimensional convolution and Capobianco et al. [2021b] suggests the use of an attention mechanism to allow the decoder to more easily focus on specific time updates of the input during prediction. In order to account for external information, some works include additional features extracted from AIS or external data sources. Liu

et al. [2021, 2022c] suggest an encoder/decoder structure in which the encoded latent vector from multiple ships is averaged and used to initialise the decoder. This is supposed to make the decoder able to take into account not just the previous trajectory of the current ship being predicted but also all other vessels in the vicinity. Dijt and Mettes [2020] include a sequence of radar images centered on the modeled vessel to provide context regarding the local environment, including the shoreline and nearby vessels.

For the purpose of abnormality detection, it is necessary to study the effect of different architectures including additional input features on the type of behavior flagged as abnormal. Features such as final destination, external weather data, nearby vessel positions, etc. all seem intuitive to include, but it is extremely difficult to evaluate their value without a publicly available baseline validation set. Even if these features are discovered to be useful for abnormality detection, it may not be trivial to include them in a real-life operational setting. Even comparison of trajectory prediction models is difficult due to a lack of a standardized available dataset. Spadon et al. [2022] compare the performance of deep learning models with traditional machine learning models for regression and find deep learning models significantly outperform other models. Spadon et al. [2022] similarly compare different architectures of RNNs. They find that their proposed CNN-RNN hybrid stabilizes the performance across data complexity and improves feature extraction for multiple vessels of different types. However, in general, they find little difference in terms of the Root Mean Squared Error (RMSE) between deep learning models, and the variance over multiple trainings is larger than the differences between models.

Most applications of deep learning methods allow for real-time detection of abnormalities. RNNs process trajectories sequentially and predictive/reconstructive errors of individual updates may be used to flag abnormalities. Most work utilize a global threshold for detection [Zhao and Shi, 2019b, Singh et al., 2021], however, the use of a global threshold for abnormality detection may cause a bias in sparse regions in the input space. Since models will naturally perform better in regions with more training data, the use of global thresholds will cause an increase in detection in regions with lower modeling performance. The A-Contrario detection method [Nguyen et al., 2021] overcomes this issue by only considering the reconstruction/predictive errors of the observation in the local vicinity when deciding the threshold. However, the time complexity of A-Contrario detection is quadratic with respect to the trajectory length, thus, there may be a need for more efficient real-time detection methods.

Predictive/reconstructive deep learning methods can be trained on large corpora of data using mini-batches and stochastic optimization algorithms. This allows unsupervised deep learning models for abnormality detection to scale very well with large amounts of training data. Similarly, neural networks allow for rel-

atively easy inclusion of explanatory features. However, these features needs to be widely available in real time to be useful in an operational setting and recording of the features must be scaleable to large data amounts as well. Dijt and Mettes [2020] include a sequence of radar images centered on the modeled vessel. This approach may work for large ships with on-board radar capabilities centered on their own position but applications to off-site surveillance centers when radar images are off center is less clear. Capobianco et al. [2021b] include the final destination as an optional categorical variable in the decoder of a Seq-2-Seq model predicting the future position. This distinction may not be scaleable to global or large ROIs with many different shipping lanes, harbors, and traffic that does not follow the well-defined shipping lanes. First, the discovery of exit points requires a clustering of AIS messages similar to the procedure suggested in TREAD [Pallotta et al., 2013b]. Secondly, in order to be used in real-time it requires a mapping from the final destination reported using AIS to the discovered exit points. Similarly, Zhao and Shi [2019b], Singh et al. [2021] both suggest preprocessing trajectories using DBSCAN to make the LSTM easier to train. Zhao and Shi [2019b] trains different predictive models for each cluster discovered using DBSCAN [Zhao and Shi, 2019a] and Singh et al. [2021] suggest the generation of a graph network based on waypoint discovery using DBSCAN. Each trajectory update is then assigned to an edge in this network, and this assignment is provided to the LSTM model as input. As we discussed previously, clustering does not scale very well to large amounts of training data, thus training of the proposed models might prove computationally expensive in practice.

To a large extent, the literature has focused on the design of increasingly deeper and more complex models and interpretation of detections has been largely ignored by previous works. As discussed previously, clustering of the latent space of deep recurrent AutoEncoders found that the latent space encodes information related to the route chosen through the ROI. This information might be useful for the detection of abnormalities that evolve on a global scale, that is, across the entire ROI. However, information related to behavior on a global scale does not provide useful contextual information related to the detection of local abnormalities that depends on more localized behavioral differences. In Chapter 8, we investigate the interpretability of sequential reconstruction based deep neural networks and if information related to the detection of abnormalities can be induced in the latent variables.

Chapter 3

Deep Neural Network

In this chapter, we present the deep neural networks used in this thesis for learning the normalcy model using historical AIS data. These networks combine ideas from two classes of models, the variational AutoEncoder (VAE) and recurrent neural networks (RNNs). VAEs are used to model complex high dimensional data by introducing latent variables and use neural networks to parameterize conditional distributions for the latent variables and the observed data. RNNs are used to model temporal dependencies in data (usually time series) through their internal memory units.

3.1 Variational AutoEncoder

An AutoEncoder (AE) is a neural network trained to copy the provided input, $\mathbf{x}$, to the output, $\hat{\mathbf{x}}$. This is done by first learning a compressed description, $\mathbf{z}$, which is the used to make a reconstruction as faithful as possible to the input, $\hat{\mathbf{x}} \approx \mathbf{x}$. The unobserved variable $\mathbf{z}$ is referred to as the latent variable and the layer is commonly called the bottleneck layer. The network consists of two parts; an encoder function, $\mathbf{z} = f(\mathbf{x})$, that learns the latent variable given an input, and a decoder function, $\hat{\mathbf{x}} = g(\mathbf{z})$, that generates a reconstruction of the original data given a latent representation.

Historically, AEs have been used for dimensionality reduction and feature extraction Protopapadakis et al. [2017]. However, the Variational AutoEncoder (VAE) [Kingma and Welling, 2013, Rezende et al., 2014] extends the encoder and decoder functions from deterministic functions to probability distributions $p_{\text{encoder}}(\mathbf{z}|\mathbf{x})$ and $p_{\text{decoder}}(\hat{\mathbf{x}}|\mathbf{z})$, making VAEs excellent for generative modeling..

The joint distribution of the input and latent variables defines the generative model as

$$p(\mathbf{x}, \mathbf{z}) = p_{\text{decoder}}(\mathbf{x}|\mathbf{z})p(\mathbf{z}) \tag{3.1}$$

The prior over the latent variable, $p(\mathbf{z})$, is often chosen to be a simple Gaussian distribution. The decoder $p_{\text{decoder}}(\mathbf{x}|\mathbf{z})$ is typically a Gaussian distribution (for continuous data) or a Bernoulli distribution (for binary data) whose parameters are computed by passing the latent state $\mathbf{z}$ through a deep neural network. The weights and biases in the deep neural network defines the parameters, θ, over which we wish to optimize. Since we are interested in learning a model that explains the observed data well, we aim at maximizing the probability assigned to $\mathbf{x}$. Therefore the optimal parameter θ^* is given by

$$\theta^* := \underset{\theta}{argmax}\ p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) d\mathbf{z} \tag{3.2}$$

Computation of $p_\theta(\mathbf{x})$ using (3.2) becomes intractable since the marginalization over $\mathbf{z}$ is prohibitively expensive due to the non-linearities in the decoder. Similarly, the true posterior distribution $p_{\text{encoder}}(\mathbf{z}|\mathbf{x})$ is also intractable, since this requires computation of $p_\theta(\mathbf{x})$. Variational Inference (VI) overcomes the intractability of the true posterior $p_{\text{encoder}}(\mathbf{z}|\mathbf{x})$ by introducing an approximate posterior

$$q_{\text{encoder}}(\mathbf{z}|\mathbf{x}) \approx p_{\text{encoder}}(\mathbf{z}|\mathbf{x}) \tag{3.3}$$

The approximate posterior is often chosen to be a Gaussian distribution with diagonal variance.

$$q_{\text{encoder}}(\mathbf{z}|\mathbf{x}) = N(\boldsymbol{\mu}, \text{diag}(\boldsymbol{\sigma}^2)) \tag{3.4}$$

where the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\sigma}$ are modeled using a neural network referred to as the encoder network.

The introduction of the approximate posterior allows to express the likelihood of $p_\theta(\mathbf{x})$ as an expectation over the approximate posterior $q_{\text{encoder}}(\mathbf{z}|\mathbf{x})$.

$$p_\theta(\mathbf{x}) = \int_{\mathbf{z}} p(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \int_{\mathbf{z}} \frac{q_{\text{encoder}}(\mathbf{z}|\mathbf{x})}{q_{\text{encoder}}(\mathbf{z}|\mathbf{x})} p(\mathbf{x}, \mathbf{z}) d\mathbf{z} = \mathbb{E}_{q_{\text{encoder}}(\mathbf{z}|\mathbf{x})}\left[\frac{p(\mathbf{x}, \mathbf{z})}{q_{\text{encoder}}(\mathbf{z}|\mathbf{x})}\right] \tag{3.5}$$

Taking the logarithm defines the Evidence Lower Bound (ELBO) loss function.

$$\mathcal{L}(\mathbf{x}) := \mathbb{E}_{q_{\text{encoder}}(\mathbf{z}|\mathbf{x})}\left[log(p_{decoder}(\mathbf{x}|\mathbf{z})) - log(q_{encoder}(\mathbf{z}|\mathbf{x})) + log(p(\mathbf{z}))\right] \quad (3.6)$$

$$= \overbrace{\mathbb{E}_{q_\phi(\mathbf{z}|\mathbf{x})}\left[\log p_\theta(\mathbf{x}|\mathbf{z})\right]}^{\text{Reconstruction Error}} - \overbrace{\mathcal{D}_{\text{KL}}\left(q_\phi(\mathbf{z}|\mathbf{x}) \mid p(\mathbf{z})\right)}^{\text{Regularization}} \quad (3.7)$$

Thus, learning advances by maximising the ELBO (3.6). Optimizing the ELBO results in a trade-off between two terms (3.7). The first term measures the quality of the reconstruction, while the second term enforces the posterior $q_{\text{encoder}}(\mathbf{z}|\mathbf{x})$ to match the prior $p(\mathbf{z})$. Regularization is done using the Kullback-Leibler (KL) divergence. The KL divergence is a measure of the similarity between two probability distributions. Thus, a stronger regularization (weighting the KL term higher) drives the approximate posterior closer to the prior, making it more difficult to produce good reconstructions. Maximization of (3.6) can be done by standard backpropagation Rumelhart et al. [1986] using the reparametrization trick [Kingma and Welling, 2013, Rezende et al., 2014] to sample and backpropagate through the latent variable $\mathbf{z}$.

3.2 Recurrent Neural Net

Recurrent neural networks (RNN) are an extension of deep neural networks designed to model sequential data of variable length. RNNs assume that the joint distribution over the sequence $\mathbf{x_{1:L}}$ follows the factorization.

$$p_\phi(\mathbf{x}_{1:L}) = \prod_{t=1}^{L} p_\phi(\mathbf{x}_t|\mathbf{x}_{t-1}, \ldots, \mathbf{x}_1) \quad (3.8)$$

We assume that, at time t, all the relevant information from the past $\mathbf{x}_{1:t-1}$ can be encoded into a single variable $\mathbf{h}_{t-1}$. The latent variable $\mathbf{h}_t$ evolves over time, and at each time step incorporates the information from the previous elements of the sequence, using the update equation $\mathbf{h}_t = f_\phi(\mathbf{h}_{t-1}, \mathbf{x}_{t-1})$. With this definition, the factorization (3.8) becomes

$$p_\phi(\mathbf{x}_{1:L}) = \prod_{t=1}^{L} p_\phi(\mathbf{x}_t|\mathbf{h}_{t-1}) \quad (3.9)$$

The function f_ϕ used to update the latent variable is a differentiable non-linear function that has to be powerful enough to capture long-term dependencies in

the sequence. Common choices for f_ϕ are memory cell units such as Long-Short Term Memory (LSTM Hochreiter and Schmidhuber [1997]) or Gated Recurrent Unit (GRU Chung et al. [2014]).

RNNs normally process sequences in the forward direction, i.e. from the past to the future. However, sometimes future events may influence the past. In translation tasks the order of the words may change between the languages or in the maritime domain a ship may reduce the speed to prepare for a future turn. Bidirectional RNNs combine two RNNs to provide an architecture where sequences are processed in both the forward and backward temporal directions. At each time step, the latent variable is obtained by concatenating the forward and backward variables $\mathbf{h}_t = \left[\overrightarrow{\mathbf{h}_t}, \overleftarrow{\mathbf{h}_t}\right]$.

RNNs may be trained for different purposes. If RNNs are used to predict the next word given the previous, it may be sufficient for the latent variable to only store information a few words back. However, we may also use the latent variable in an AE framework where we reconstruct an entire sequence from a single variable encoding the entire input sequence.

3.3 Sequence-2-sequence models

RNN Sequence-2-Sequence (Seq-2-Seq) models have been proposed in order to train an RNN to map an input sequence to an output sequence of different lengths. They are widely used in natural language processing for translation [Sutskever Google et al., 2014] and text generation [Graves, 2013]. Similarly to the AE it employs 2 RNNs in an encoder-decoder architecture. The encoder reads the input sequence $\{\mathbf{x}_1, \ldots, \mathbf{x}_l\}$ and obtains the fixed-dimensional latent variable $\mathbf{h}_l$. The final latent variable $\mathbf{h}_l$ is passed to the decoder which recursively predicts the output sequence $\{\mathbf{y}_1, \ldots, \mathbf{y}_m\}$.

$$\hat{\mathbf{y}}_t = \mathbf{W}_{u\hat{x}}\mathbf{u}_t + \mathbf{b}_{u\hat{x}} \tag{3.10}$$

$$\mathbf{u}_t = g_\theta(\hat{\mathbf{y}}_{t-1}, \mathbf{u}_{t-1}, \mathbf{h}_l) \tag{3.11}$$

where $\mathbf{u}_{t-1}$ is the hidden state of the decoder RNN modelled by the function g_θ. The length l of the input sequence and the length m of the output sequence are referred to as the context length and the prediction length, respectively. Training is carried out by minimizing the prediction error.

$$\mathcal{L}(\mathbf{x}, \mathbf{y}) = \frac{1}{m}\sum_{t=1}^{m}\|\hat{\mathbf{y}}_t - \mathbf{y}_t\|^2 \tag{3.12}$$

A simple Seq-2-Seq network as described above is shown in Figure 3.1.

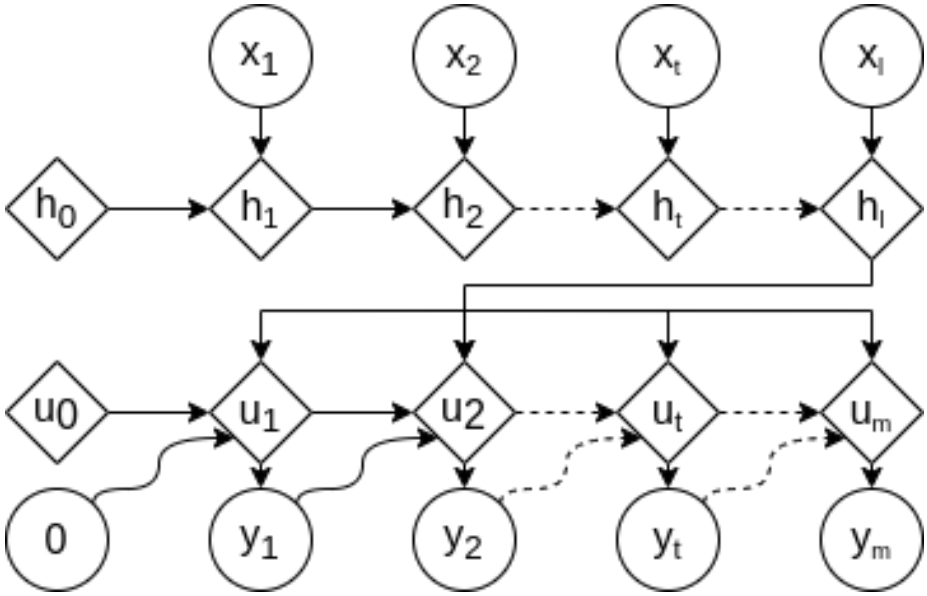


Figure 3.1 – A basic Seq-2-Seq network. Diamonds denote deterministic variables and circles random variables.

A clear limitation of the encoder-decoder strategy occurs when the size of the hidden state is too small to properly summarize an entire sequence. To alleviate this problem, the attention mechanism can be added as an intermediate layer between the encoder and decoder. The attention mechanism allows each step in the decoder to focus on the hidden state at intermediate time steps instead of just the last. This is done by taking a weighted average of the encoder hidden states. Consider the sequence of encoder hidden states $\{\mathbf{h}_1, \ldots, \mathbf{h}_l\}$ and the hidden state of the decoder $\mathbf{u}_j$, then each context vector $\mathbf{z}_j$ is found by

$$\mathbf{z}_j = \sum_{t=1}^{l} \alpha_{jt} \mathbf{h}_t \tag{3.13}$$

The attention weights α_{jt} are calculated using the softmax operator as

$$\alpha_{jt} = \frac{\exp e_{jt}}{\sum_{t=1}^{l} \exp e_{jt}} \tag{3.14}$$

with

$$e_{jt} = \mathbf{v}_a^T tanh(\mathbf{W}_h \mathbf{h}_t + \mathbf{W}_u \mathbf{u}_{j-1}) \tag{3.15}$$

The weights $\mathbf{v}_a$, $\mathbf{W}_h$, and $\mathbf{W}_u$ are trainable network parameters used to score the spatio-temporal relationship between inputs around time t and the output

at time j. A graphical representation of a Seq-2-Seq with bidirectional encoder and attention is shown in Figure 3.2.

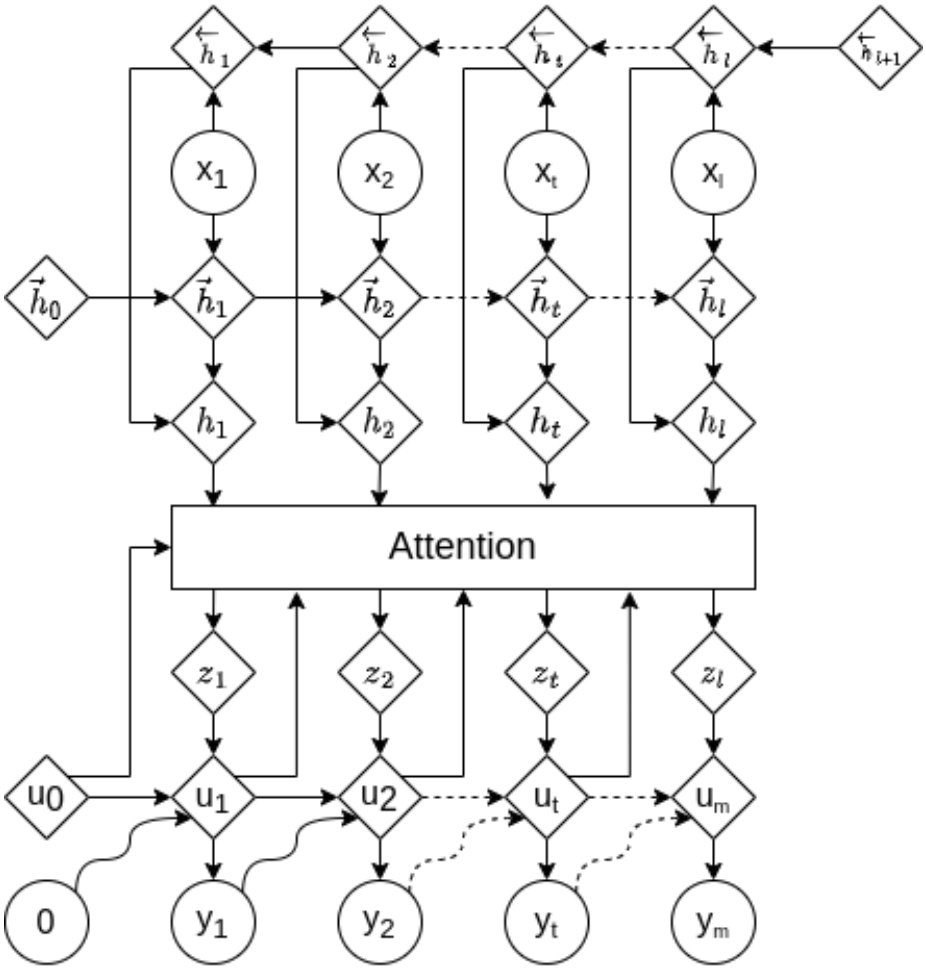


Figure 3.2 – A bidirectional Seq-2-Seq network with attention. Diamonds denote deterministic variables and circles random variables.

3.4 Recurrent Variational AutoEncoder

In the previous section, we saw how the hidden state of the RNN can be used to make sequential predictions of future positions. In a Recurrent Variational

AutoEncoder (RVAE, Srivastava et al. [2015]) the final latent hidden state is instead used to make a reconstruction of the entire input sequence. This creates an AE where both the encoder (3.4) and decoder (3.1) are approximated by recurrent neural networks. The final hidden states of encoder network is passed through a fully connected layer to give the parameters of the Gaussian distribution (3.4)

$$\mu_z = \mathbf{W}_\mu \mathbf{h}_L + \mathbf{b}_\mu \tag{3.16}$$
$$\sigma_z = \mathbf{W}_\sigma \mathbf{h}_L + \mathbf{b}_\sigma \tag{3.17}$$
$$\tag{3.18}$$

We then sample a latent variable z from (3.4) and calculate the initial hidden state of the decoder

$$\mathbf{u}_0 = tanh(\mathbf{W}_{zu}\mathbf{z} + \mathbf{b}_{zu}) \tag{3.19}$$

Reconstruction of the trajectory is done sequentially using (3.10) and (3.11), however, the decoder is not updated using the hidden state of the encoder. The RVAE is with bidirectional encoder is sketched in Figure 3.3

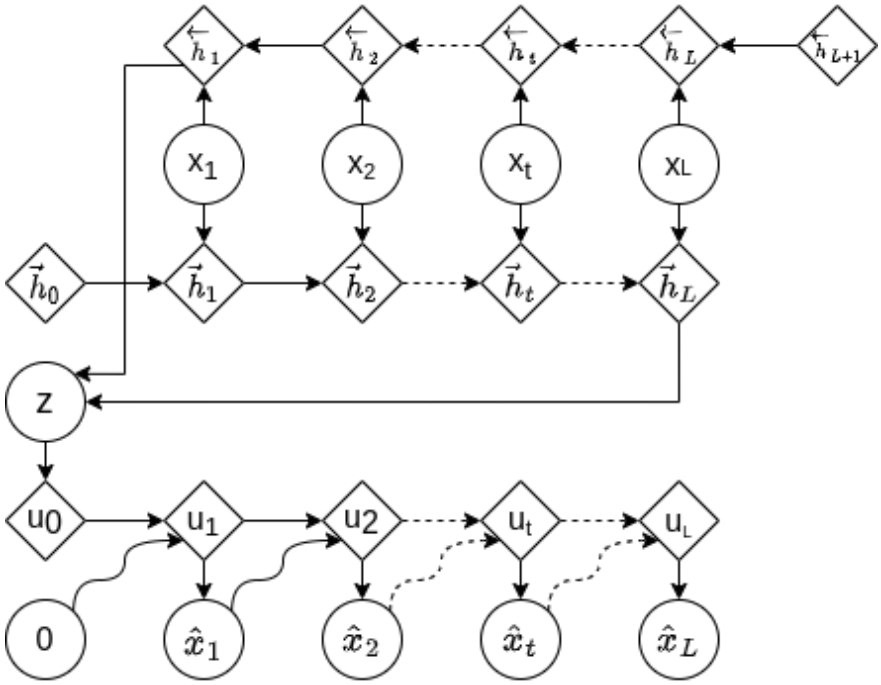


Figure 3.3 – A RVAE with bidirectional encoder. Diamonds denote deterministic variables and circles random variables.

3.5 Variational Recurrent Neural Network

The Variational Recurrent Neural Network (VRNN, Chung et al. [2015]) can be considered a RNN which at each time step consists of a VAE conditioned on the recurrence model. As previously, we model the sequence using (3.9). However, we then introduce the latent stochastic variables $\mathbf{z_t}$. The prior distribution of the latent stochastic variable at time t, $\mathbf{z}_t$, is given by a Gaussian with parameters $\boldsymbol{\mu}_{0,t}$ and $\boldsymbol{\sigma}^2_{0,t}$, obtained by a neural network ϕ^{prior} taking as input the recurrent hidden states $\mathbf{h}_{t-1}$.

$$\begin{aligned} &\mathbf{z}_t \sim N(\boldsymbol{\mu}_{0,t}, \mathrm{diag}(\boldsymbol{\sigma}^2_{0,t})), \\ &\text{where } [\boldsymbol{\mu}_{0,t}, \boldsymbol{\sigma}^2_{0,t}] = \phi^{prior}(\mathbf{h}_{t-1}) \end{aligned} \tag{3.20}$$

Similarly the generating distribution and the approximate posterior $q(\mathbf{z}_t|\mathbf{x}_t)$ is also modelled using neural networks, ϕ^{dec} and ϕ^{enc}, taking $\mathbf{h}_{t-1}$ as an input. The generating distribution also depends on the latent variable $\mathbf{z_t}$, which first passes through a feature extractor $\phi^{\mathbf{z}}$.

$$\begin{aligned} &\mathbf{x}_t|\mathbf{z}_t, \mathbf{h}_t \sim P(\boldsymbol{p}_{x,t}), \\ &\text{where } P(\boldsymbol{p}_{x,t}) = \phi^{dec}(\phi^{\mathbf{z}}(\mathbf{z}_t), \mathbf{h}_{t-1}) \end{aligned} \tag{3.21}$$

As mentioned, the distribution $P(\boldsymbol{p}_{x,t})$ is often chosen as a Gaussian or Bernoulli distribution depending on the nature of the input $\mathbf{x}$. The approximate posterior depends on the input $\mathbf{x_t}$ that passes through another feature extractor $\phi^{\mathbf{x}}$.

$$\begin{aligned} &\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_t \sim N(\boldsymbol{\mu}_{z,t}, \mathrm{diag}(\boldsymbol{\sigma}^2_{z,t})), \\ &\text{where } [\boldsymbol{\mu}_{z,t}, \boldsymbol{\sigma}^2_{z,t}] = \phi^{enc}(\phi^{\mathbf{x}}(\mathbf{x}_t), \mathbf{h}_{t-1}) \end{aligned} \tag{3.22}$$

At each time step the recurrence $\mathbf{h_t}$ is updated according to

$$\mathbf{h}_t = f_\theta(\phi^{\mathbf{x}}(\mathbf{x}_t), \phi^{\mathbf{z}}(\mathbf{z}_t), \mathbf{h}_{t-1}) \tag{3.23}$$

The VRNN network is pictured in Figure 3.4. The probability of observing the sequence $\mathbf{x}_{1:T}$ is obtained by integrating out $\mathbf{z}_t$ from (3.21)

$$log(p(\mathbf{x}_{1:T})) = \sum_{t=1}^{T} log(p(\mathbf{x}_t|\mathbf{h}_{t-1})) \tag{3.24}$$

$$= \sum_{t=1}^{T} log(\mathbb{E}_{p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_{t-1})}\left[p(\mathbf{x}_t, \mathbf{z}_t|\mathbf{h}_{t-1})\right]) \tag{3.25}$$

$$= \sum_{t=1}^{T} log(\mathbb{E}_{p(\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_{t-1})}\left[p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{h}_{t-1})p(\mathbf{z_t}|\mathbf{h}_{t-1})\right]) \tag{3.26}$$

This integral is intractable but can be approximated using variational inference as explained in Section 3.1. Thus, learning is done by maximizing the timestep-wise ELBO.

$$\mathcal{L}(\mathbf{x_{1:T}}) = \sum_{t=1}^{T} \mathbb{E}_{q(\mathbf{z}_t|\mathbf{x}_t,\mathbf{h}_{t-1})} p(\mathbf{x}_t|\mathbf{z}_t,\mathbf{h}_{t-1}) - KL\left[q(\mathbf{z}_t|\mathbf{x}_t,\mathbf{h}_{t-1})|p(\mathbf{z}_t,\mathbf{h}_{t-1})\right] \tag{3.27}$$

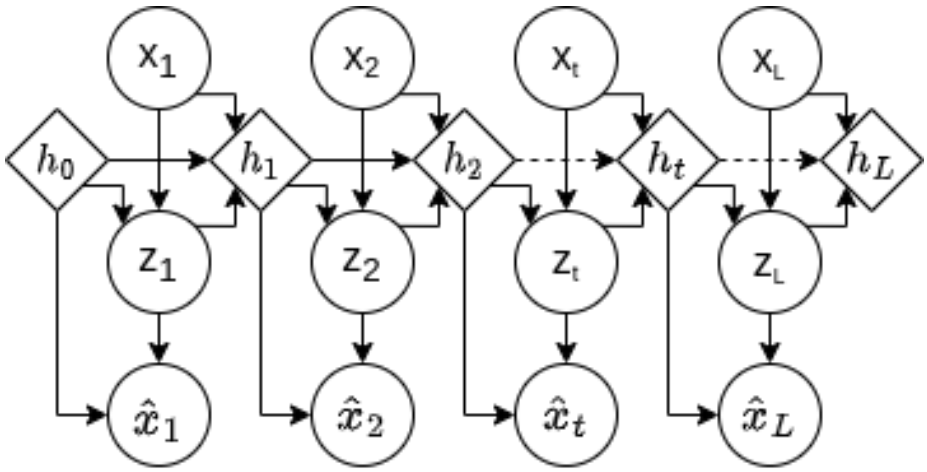


Figure 3.4 – The VRNN network is a VAE which at each time step is conditioned on a recurrence model. Diamonds denote deterministic variables and circles random variables.

3.6 A-Contrario detection

In this section we present a description of A-Contraio detection suggested in Nguyen et al. [2021]. In this section, the method will be derived using reconstructive probabilities as suggested in Nguyen et al. [2021]. However, the method is equally applicable using reconstruction or predictive errors from point predictions as in (3.10), by changing from detection of probabilities lower than the training set to errors higher than the training set. .

Having learned the distribution $p(\mathbf{x}_{1:T})$ (3.24), one may apply a global threshold to mark all trajectories with low probability. These trajectories may be regarded as abnormal since the model have not learnt to accurately model the dynamics. However, in local regions with a limited number of trajectories, the model may not have learned to reconstruct trajectories to the same degree as in densely

populated regions. Thus, applying a global threshold might cause the detection of trajectories in a sparsely populated region, regardless of whether they are abnormal or not. This issue is alleviated using A-Contrario detection.

A-Contrario detection divides the ROI into geographical cells C_i. Instead of considering the global reconstruction, we limit ourselves to a small geographical cell around the current target C_i, and look at the reconstruction within this cell. Let $l^{C_i}_{\mathbf{x'}_t}$ denote the log-probabilities $log(p(\mathbf{x'}_t|\mathbf{h}_{t-1}))$ of AIS messages within C_i. P^{C_i} then denotes the distribution of $l^{C_i}_{\mathbf{x'}_t}$ modelled by Kernel Density Estimation.

$$l^{C_i}_{\mathbf{x'}_t} \sim P^{C_i} \tag{3.28}$$

An AIS-message is the considered to be abnormal if the log probability, $L_{\mathbf{x}_t}$, is less than the $1/p$-quantile of P^{C_i}. In other words, an AIS message is abnormal if the probability of observing worse log-probabilities within the same cell is less than p. When using reconstruction errors, the error must be greater than (1-1/p)-quantile to be considered abnormal.

$$\mathbf{x}_t \text{ is abnormal} \leftrightarrow P^{C_i}(l^{C_i}_{\mathbf{x'}_t} < L_{\mathbf{x}_t}) < p \tag{3.29}$$

In Nguyen et al. [2021] the value p is set to 0.1. Therefore, the probability that any individual randomly sampled AIS message is abnormal is 10%.

Assuming that the event "$\mathbf{x}_t$ is abnormal" is independent for updates in a trajectory $\mathbf{x}_{1:T}$, the probability that at least k of n AIS messages are abnormal follows the tail of a binomial distribution.

$$B(n,k,p) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \tag{3.30}$$

A-Contrario detection detects whether a trajectory $\mathbf{x}_{1:T}$ contains any abnormal sub-segments. A subsegment is abnormal if the probability of the observed binomial tail is lower than some threshold. If an abnormal subsegment exists, the entire trajectory is denoted as abnormal.

$$\mathbf{x}_{1:T} \text{is abnormal} \Leftrightarrow \exists (n,k), N_s \cdot B(n,k,p) < \epsilon \tag{3.31}$$

The scaling factor N_s accounts for the number of different subsegments that can be created from the trajectory $\mathbf{x}_{1:T}$ of length T. Its value can be calculated using $N_s = \frac{T(T+1)}{2}$. The quantity $N_s \cdot B(n,k,p)$ denotes the outlier score of the

trajectory $\mathbf{x}_{1:T}$ that we can use to gauge the degree of abnormality. We define the outlier factor of the trajectory $\mathbf{x}_{1:T}$ as the inverse of this quantity.

$$OF(\mathbf{x}_{1:T}) \equiv \frac{1}{N_s \cdot B(n,k,p)} \tag{3.32}$$

Part II

Research Outcomes

Chapter 4

Summary and Discussion of Research

In this chapter, the included publications presented in the following chapters are summarized. The research follows two main sections. In Section 4.1 papers I and II are presented. These papers study deep learning models for the automated detection of maritime abnormalities in real time. Papers III and IV are presented in Section 4.2. These papers study the interpretability of a state-of-the-art model evaluated in the first section and ways to extract the learned normalcy model in order to provide detection explanation or to allow for more man-in-the-loop style detection. Methods for more abnormality detection of more localized kinematic behaviour are developed.

4.1 Reconstruction based Abnormality Detection

In this section, deep learning models for automated detection of maritime abnormalities are investigated. A major issue with using unsupervised deep learning methods for extracting the normalcy picture is evaluation of the learned normalcy model, i.e. can the learned representation of maritime traffic be used for detection of abnormal behaviour. One possible solution is to evaluate the abnormality detection performance of trained models on a data set representative

of real-life maritime traffic.

For this purpose, we use a dataset with annotated abnormal traffic related to a collision accident in the waters northwest of the Danish island of Bornholm on December 13^{th} 2021 originally prepared for Paper IV. In total, the data set consists of 521 trajectories extracted from AIS data, where 25 trajectories were found to be abnormal. Trajectories consist of latitude and longitude coordinates, as well as speed and course over ground reported using the AIS system. In addition to the colliding vessels, the abnormal trajectories involve commercial traffic, which had to deviate from the planned course, Search-and-Rescue and law enforcement vessels responding to the accident, and any other vessel participating in the search of two missing sailors. For training, trajectories from the period of June 1^{st} to November 30^{th} from the same region was extracted.

The datasets cover a rectangular ROI over the Bornholm island bounded by (54.5^o N, 13^o E) to (56^o N, 16^o E). The data was limited to moving updates, that is, AIS messages with a stationary flag were discarded, and 8 different types of ship covering a number of different priorities and complex behaviors. These ship types are; Cargo, Tanker, Fishing, Passenger, Sailing, Pleasure, Military, High Speed Vessel. For the test set the ship types Search-and-Rescue and Law Enforcement where added. The speed was limited to 20 m/s, and updates with higher speeds were discarded. Unless otherwise mentioned tracks shorter than 10 minutes were discarded and tracks exceeding 12 hours were split into smaller tracks, each between 10 minutes and 12 hours. Tracks are resampled every 120 seconds using linear interpolation. In order to ensure updates for interpolation, tracks were split into two contiguous tracks if the time interval between two successive AIS messages exceeded 15 min,

Paper I - A Review of Current Deep Learning Techniques for Maritime Abnormality Detection and Directions for Future Progress

Paper I presents a review of deep learning methodologies for analyzing maritime trajectories from a perspective of abnormality detection. The review finds several different applications of deep learning models for the detection of maritime abnormalities and other problems requiring the analysis of maritime trajectories. Other problems within the field of maritime trajectory analysis can in many circumstances be interpreted as detection models for a special case of maritime abnormalities. For instance, classification of the ship type may be used to detect vessels trying to hide their true identity and prediction of the vessel state may be used to detect illegal fishing activities in exclusion zones. In fact, many abnormality detection methods are based on prediction of the future trajectory.

We find that previous literature has mainly focused on the design of more complex models for both abnormality detection and other problems within maritime deep learning. However, it remains an open question how these models compare with each other quantitatively and qualitatively. Likewise, it remains an open question what impact the choice of architecture, hyperparameter values, and input features have on abnormality detection.

We then evaluate four methodologies suggested in the literature for abnormality detection using the test set described above. We find that the models generally achieve similar performance as measure by the ROC-AUC, however, by inspecting the false positives we find that they flag different forms of behavior.

We find that predictive Seq-2-Seq models generally learned a linear projection of future states causing false positive for trajectories with varying course and turns at high speeds. Course variations at low speeds were found not to induce alarms due to the relatively smaller prediction errors. This stresses a potential issue with the prediction of abnormal behavior at low speeds. We find that the GeoTrackNet model is susceptible to flag trajectories with frequent course changes, and RAE better detects traffic that follows the shipping lanes with small deviations or is located near the shipping lanes. For this reason, we argue that an ensemble of the two AutoEncoder based models is a better abnormality detector than any of the studied models individually. Additionally, the paper investigate the outliers scores of different ship types and find large differences between ship types indicating the ship types should be included in the abnormality detection.

Based on the review and evaluation of models the paper presents the following guidelines for future research:

- Models should be trained and evaluated on data sets representative of the expected traffic in real-time applications.
- Abnormality detection models should focus on unsupervised methods.
- Preprocessing steps should be kept to a minimum unless a change in the abnormality detection performance can be verified.
- We propose a study of the impact of additional inputs such as kinematic AIS information, weather data, vessel-2-vessel interaction, etc. on the behavior flagged as abnormal.
- We suggest that future model designs include ideas of how prediction interpretability may be achieved.
- The ship type should be included in the abnormality detection.

- The literature should work towards ensembles of models that detect different types of abnormal behavior.

In terms of the research objectives presented in Section 1.2 Paper I contributes primarily to objectives A and B. Deep learning frameworks are discussed in detail and evaluated on the basis of the performance for automated abnormality detection. The paper shows by using objective measures that deep learning methods can be used for automated detection of maritime abnormalities associated with collision accidents. Although both colliding vessels are detected as abnormal, it is generally due to events after the collision occurred. As such an impending collision was not detected and models which better model vessel interactions should be investigated.

As discussed previously, automated abnormality detection models may not be a desired end goal Endsley [2017]. It is likely that automated abnormality detection models will be implemented as a decision support tool instead of a tool to enhance level 2 situation awareness of human surveillance operators. As such special care should be taken as to how automated models are utilized in a practical operational setting if at all. A very conservative model with only few alarms might be implemented in a parallel setting with human operators and contribute with detection of only the most extreme cases concerning operational security. However, previous cases of this sort are generally unknown to the public, making it very difficult to objectively measure the performance of these extreme cases. Automated tools could potentially also be utilized in limited capacity for training purposes or for application in new ROIs where little is known about the normal maritime picture and human operators may lack level 2 situation awareness.

Paper II - Detecting Abnormal Maritime Trajectories using Ensembles and Transfer Learning

Paper II builds upon the findings of Paper I and addresses some of the questions raised for future work. Initially, the different inputs and objective functions are investigated based on their practicality for abnormality detection. Overall, the behavior flagged as abnormal is very similar between the three learning objectives considered. The model trained for point predictions achieves the overall best abnormality detection performance. Models trained for predictions of the full generative distribution are generally more susceptible to noise. This sensitivity can be reduced by implementing a discretization of the inputs. However, this makes the models tied to the ROI and makes application to new regions or implementation on moving platforms impossible.

Models are then trained on trajectories with irregular sampling. The training trajectories are subsampled by randomly selecting a number of updates from each trajectory, each mini-batch. The number of sampled points depends on the length of the entire trajectory. Therefore, the previous limitation of a maximum trajectory length of 12 hours is removed. We found that training using random irregular sampled trajectories did not negatively affect the abnormality detection of models trained for point predictions. Training with random irregular samples made the models generalizable to different resample periods during evaluation. As such, the same model was evaluated with a resample period of 120 seconds as described above, but also using a resample period of 600 seconds. Depending on the type of behavior of interest, different sampling periods may be useful, and a model trained on random irregular samples can be used to evaluate all sampling periods.

We then evaluate ensembles of models with different model architecture, objective function, and resample periods and find that an ensemble consisting of GeoTrackNet evaluated using resample periods of 120 and 600 seconds and a RAE model evaluated using resample a period of 120 seconds achieves the best performance as measured by ROC-AUC. The ensemble members were qualitatively found to flag different types of abnormal trajectories, which explain the performance boost.

Lastly, the generalization to different ROIs is studied and it is investigated whether transfer learning can be applied to reduce the training time for new ROI. Models are trained using trajectories captured around the Danish island of Anholt within the bounding box (56^o N, 10.7^o E) to (57.5^o N, 13^o E) using the same preprocessing strategy. It is found that models trained for point prediction does not require any form of fine-tuning on the target domain to achieve similar level of abnormality detection performance.

Paper II address the generalization of deep learning methods and contributes towards objective C and B. Ensembles are found to improve abnormality detection when made up of members that flag different types of behavior as abnormal. In order to improve automatic detection further additional ensemble members that flag other abnormal forms of behaviour could be introduced. For this purpose, additional cases with different forms of abnormal behavior should be annotated for evaluation purposes.

Construction of such ensembles might not only be beneficial for automated models, but also useful for enhancing level 2 situation awareness. With different deep learning models, the trajectories can be evaluated against different normalcy models. However, this requires that it is possible to extract the normalcy model in a form that is interpretable for human operators, which might prove difficult for every ensemble member.

4.2 Clustering based Abnormality Detection

An issue of particular interest to surveillance operators is the general lack of interpretability of model detections. This is especially true for the deep learning architectures explored in papers I and II. In this section, we study how to extract normalcy models from deep learning models and we develop a method for constructing normalcy models focusing on the local kinematic behavior. In particular, we wish to construct normalcy models with disentangled and physically interpretable information. That is, we wish that dynamic information, e.g. local kinematic behavior, is disentangled from static information, e.g. the general location or route choice.

In this section, we shall briefly use the previously mentioned annotated test set for a quantitative comparison of clustering and deep learning-based automated detection of maritime abnormalities in Paper IV. However, the primary discussion of results is based on qualitatively comparisons. The datasets used are described in each respective paper.

Paper III - Towards latent representation interpretability for maritime anomaly detections

Paper III study how to extract the learned normalcy model from the state-of-the-art GeoTrackNet model studied in papers I and II. The study focuses primarily on the GeoTrackNet model, as the RAE model has been investigated from this perspective in Murray and Perera [2021]. Murray and Perera [2021] found that clusters in the latent space of the RAE model correspond to the global behavior through the ROI, i.e. the route chosen through the ROI. As such, clustering in the latent space of the RAE model can be used to extract information about the shipping lanes and major routes through the ROI. This study was initially conducted before the works of Paper I and II and some of the takeaways from these two papers could have been used to change the scope of the paper or further strengthen the analysis. We shall comment on this where ever necessary in the discussion.

To extract information related to the learned normalcy model, we studied the activation of the stochastic latent space in the GeoTrackNet model. The latent space is found to be nonzero only in the most extreme cases of vessel maneuvering at high speeds with a highly varying course. Thus, the information extracted from the latent space may not be useful for explaining flagged trajectories or to detect abnormalities in a man-in-the-loop style. GeoTrackNet is based on the VRNN described in Section 3.5 and introduces the latent variables z_t described by the inference model Eq. (3.22) to account for the random effects from exter-

nal sources. However, the trajectories are then preprocessed using discretization so that only the most extreme cases of random effects are reflected in the data. From this aspect, it is not surprising that the stochastic latent space is mostly inactive, and the majority of information is modeled using the recurrent model. However, using the continuous inputs investigated in Paper II the random effects will not be suppressed to a similar degree, and a larger amount of information in the stochastic latent space would be expected.

In order to induce more information encoded into the stochastic latent space the learning objective is modified. To allow for a more flexible latent space, the Kullback-Leibler regularization term in Eq. 3.27 is relaxed by slowly annealing the weight, β, of the term from 0 to 1 over the first 10 training epochs. To improve the disentanglement of physically interpretable information in the stochastic latent space, an ElasticNet loss is added to the weights of the encoder ϕ^{enc}, and prior ϕ^{prior}, networks. Additionally, it is preferred if dynamic information related to the reconstruction of the current update is modeled using the recurrence model and that information in stochastic latent space is related to an overall or local description of the current vessel state across several time updates, independent of current location, speed, and course. For this purpose, a static consistency loss is added to drive the latent encodings from the same trajectory at different times towards one another. To avoid the trivial zero encoding, we also add a loss to drive away the latent encodings of other vessels.

In a disentangled latent space, we would expect to see a connection between updates in a trajectory, and we would expect to see clusters of trajectory segments with similar behavior, i.e. clusters denoting different behaviors such as steaming at constant speed, acceleration, turn to starboard, etc. However, clustering in this latent space, we found that latent encodings generally corresponds to segments of shipping lanes in the ROI. We also observe a significant overlap between the segments of the shipping lanes assigned to different clusters. We suspect this to be an artifact of the static consistency loss forcing apart latent encoding of trajectories sharing similar behavior. Due to a large overlap, we do not identify any specific type of vessel behavior associated with each cluster as we would have hoped from a fully disentangled latent space. This indicates that the stochastic latent space alone may not be sufficient for the detection of maritime trajectories with abnormal behavior in a man-in-the-loop style. Sequences of latent vectors jump between different areas of the latent space as the trajectory moves through the ROI making it difficult to detect abnormal behavior based on latent space transitions.

Compared to the clusters reported in Murray and Perera [2021], the clusters obtained using the static consistency loss only make up parts of each possible route through the ROI. Therefore, two trajectories that share parts of the same route will be clustered together for the common sections but clustered differently

when they do not follow the same routes. Thus, we have obtained a more localized description of the behavior through the ROI, but not local enough to identify common families of behavior as desired. An interesting open question is whether limiting the static consistency loss in time could induce a further localized description in the latent variables. I.e., instead of driving the latent variable of the full trajectory towards each other, the static consistency loss is computed using a sliding window centered on the current position.

Paper III address the interpretability of the state-of-the-art GeoTrackNet model in an attempt to extract the learned normalcy model to support for level 2 situation awareness. Paper III contributes to research objective A from a perspective of extracting the learned normalcy model. Paper III also attempts to address research objective D. However, it is evident that a more localized representation of behavior is needed in order to make man-in-the-loop style detections. The Seq-2-Seq models suggested in Paper I could be interpreted as a localized model as the latent encodings only depend on the most recent part of the trajectory and not the full history. As such, Seq-2-Seq models may be preferable for clustering of local behavior. However, as discussed previously, an AutoEncoder architecture may be better suited for automated detection. These two approaches could perhaps be combined in a hierarchical recurrent neural network architecture, which has been suggested for video summarization [Zhao et al., 2019] and intersession memory [Sordoni et al., 2015, Quadrana et al., 2017].

Another interesting possibility is the combination of disentangled local and global stochastic latent variables. This has been a wide studied problem within the video analysis [Zhu et al., 2020, Bai et al., 2021]. Both papers apply the same static consistency loss as suggested in Paper III to ensure that the global latent variable remains constant subject to random shuffling of the input frames. However, training of the local latent variable is less trivial for spatio-temporal time series. Augmentations, as suggested in Bai et al. [2021], may not be applicable, since permutations break the physical dependence between updates, resulting trajectories with no physical meaning, and auxiliary tasks, as suggested in Zhu et al. [2020], are not trivial, as time-dependent classification problems, such as location of the maximum optical flow or prediction of the volume, are not readily apparent using the information contained in the AIS data.

Paper IV - A two-step clustering method for maritime behaviour identification

Paper IV investigates how clustering can be used to derive a normalcy model that describes local kinematic behavior that allows for man-in-the-loop style abnormality detection. As such, Paper IV primarily contributes to research

objective D.

The paper designs two trajectory similarity measure meant to capture different aspects of trajectory similarities. The first similarity measure calculates the position similarity using the Haversine distance. It has a linear-time complexity w.r.t. trajectory length, meaning that it can be applied relatively quickly across a large dataset. The second similarity measure calculates the kinematic similarity between trajectories. This similarity measure is the sum of speed- and course-based similarities measured using Dynamic Time Warping. This distance measure is quadratic w.r.t. to trajectory length so before calculating the kinematic similarity, the trajectories are compressed using a 2-stage Douglass-Peucker algorithm. Additionally, the trajectory similarity is calculated only for a subsample of the training set as described in the next section.

Paper IV proposes to use these two similarity measure in a two-step clustering approach. In the first stage all trajectories are clustered using their positional similarity. This results in clusters of trajectories with the same origin in the ROI and then divided into different maritime routes originating at this location. This clustering serves the purpose of filtering the training set into clustering in which we expect to find similar kinematic behavior. Then, within each of these positional clusters, another round of clustering is performed using kinematic similarity. The final result is clusters whose members start in the same general area and follow the same kinematic behavior. This means that not only can the major routes be detected, they can also be separated into different speed profiles. We can also detect similar types of behavior that are not restricted to the major shipping lanes. For example, pilot boats are generally found to be different from the rest of the traffic in the ROI and mostly similar to one another even though they do not follow well-defined routes.

We may also use the kinematic similarity measure to identify abnormal traffic that does not resemble any other traffic in the ROI. This means that the similarity measure can also be used for automated detection of abnormalities. In this application, the positional clustering is conducted as before, but the second step clustering is exchanged with an abnormality detection step using Local Outlier Factor. This may also be used to evaluate new unseen trajectories using inductive clustering. A K-Nearest-Neighbor classifier is trained to classify new trajectories into one of the positional clusters. The kinematic similarity with the cluster members is then computed and the new trajectory is evaluated for normality using Local Outlier Factor. Automated detection of abnormalities was evaluated on the annotated collision dataset and found to perform significantly worse than the predictive/reconstructive methods investigated in Section 4.1. However, the more intuitive interpretability an explainability makes the clustering approach more applicable for practical use.

Paper IV designs a clustering model that can disentangle positional and kinematic behavior resulting in a normalcy model that may be used to evaluate local kinematic behavior. This normalcy model may be used for automated detection, as outlined above, but we see it better utilized in a man-in-the-loop style in cooperation with human operators. Human operators may define the characteristic behavior of each clusters and evaluate trajectories against these characteristics. For example, if a vessel claims to be a sailing ship, it does not make much sense if it suddenly sails at a speed of 10 m/s along the shipping lanes or if a fishing vessel is found to have the same behavior as pilot boats, one might suspect some kind of illicit behavior, as fishing boats should not sail close to large commercial traffic. Additionally, operators may flag certain trajectories in the training set as abnormal and can flag new test trajectories if any of these known abnormal trajectories are found to have the highest similarity.

It is natural to seek a combination of the clustering approach of Paper IV with the reconstruction based deep learning methods discussed in Section 4.1. Such a model could be used for automated detection of abnormalities as discussed in Paper I-II and man-in-the-loop style detection using clustering. Boubekki et al. [2021] suggest an AutoEncoder architecture for combined reconstruction and clustering (AECM) using isotropic Gaussian Mixture Models (GMM). Two problems arise when adapting the AECM model to the two-step clustering, and preliminary experiments addressing both of these issues are presented here.

The first is related to number of clusters which is unknown. The AECM model provides a Dirichlet prior over the mixing coefficients of the GMM, which potentially may eliminate unused components. In practice, however, we often see clusters split equally into multiple components if a symmetric prior is used. This is illustrated in Figure 4.1a which shows the AECM with 20 clusters applied to the MNIST dataset. In order to eliminate unused components, we suggest updating the Dirichlet prior after each epoch using the posterior distribution of the previous epoch. Since the Dirichlet prior is the conjugate prior of the categorical distribution, this posterior is of the form

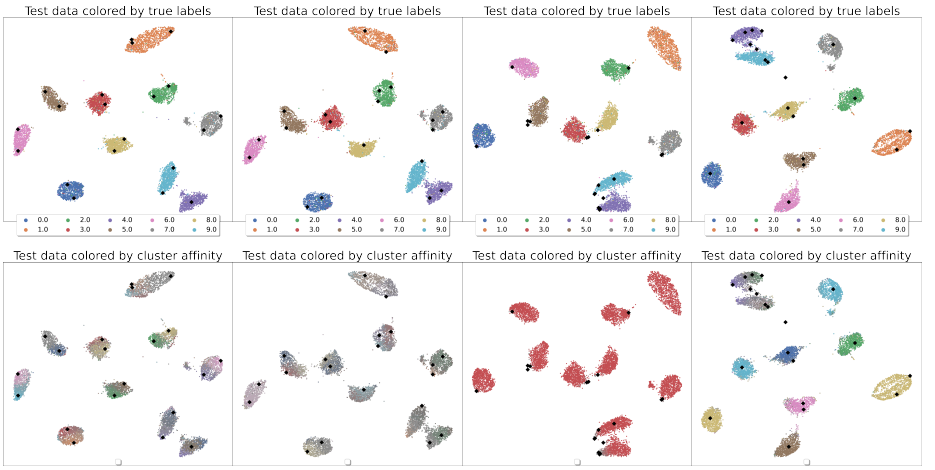
$$Dir(\alpha_0 + N_1, \dots, \alpha_0 + N_K), \tag{4.1}$$

where α_0 denote the initial symmetric prior and N_k is the number of observation assigned to the k'th component; $N_k = \sum_{i=1}^{N} \gamma_{i,k}$. Using this update the rule the Dirichlet prior loss, E_4, of the AECM becomes

$$\sum_{k=1}^{K} (1 - \alpha_0 - N_k \cdot C)\ log\ \tilde{\gamma}_k\ , \tag{4.2}$$

where $\tilde{\gamma}_k = {}^1/_N \sum_{i=1}^{N} \gamma_{i,k}$ and C is hyperparameter scaling the weights of the prior updates.

Preliminary findings on tuning these hyperparameters using the MNIST dataset are illustrated in figure 4.1. For a component k to die out, we must require that $(1 - \alpha_0 - N_k \cdot C) \approx 0$ since $log\ \tilde{\gamma}_k \to -\infty$ as $N_k \to 0$. Therefore, α_0 serves as an upper bound for the sparsity that we can achieve in the mixing components. High values of α_0 result in a solution in which all components are used equally, and low values of α_0 result in a solution that diverges to a single component. However, tuning the alpha values we can slowly eliminate unnecessary components and find the correct number of clusters without knowing how many clusters to initialize.



(a) No prior update. (b) High α_0. (c) Low α_0. (d) Tuned α_0.

Figure 4.1 – Cluster association of the MNIST testset using the AECM with 20 clusters. Models are trained with varying levels of prior updating and initial prior, α_0.

The second issue with adapting the AECM for the two-step clustering is related to the kinematic clustering being conditioned on the positional clusters.

$$P(\text{kin}|\text{pos}) = \frac{P(\text{kin}, \text{pos})}{P(\text{pos})} \tag{4.3}$$

The distribution $P(\text{pos})$ is modeled by a GMM using only positional inputs. However, joint distribution P(kin,pos) is not as simple. This distribution is given by

$$P(\text{kin}, \text{pos}) = \prod_{k=1}^{K} \pi_k N(kin|\theta_{\text{kin},k}) N(pos|\theta_{\text{pos},k}), \tag{4.4}$$

where π_k denotes the mixing coefficients and θ_k the parameters of the Gaussian distribution that describe positional and kinematic features, respectively. If the covariance matrices are diagonal, these two Gaussian distributions can be

combined to a single Gaussian distribution of the concatenated feature vectors and may be modeled using a GMM.

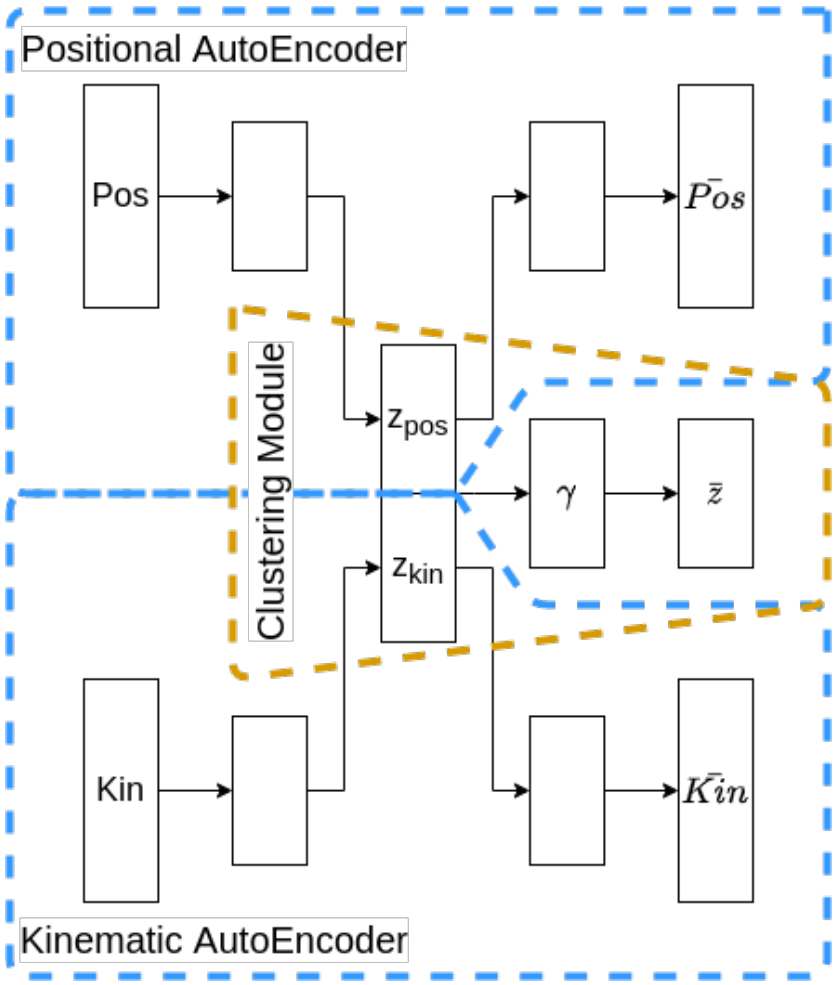


Figure 4.2 – The AECM architecture adapted for the clustering of the joint distribution $P(\text{kin}, \text{pos})$.

We have obtained latent encodings using two RAE's from Paper I trained on only the positional and kinematic features, respectively and trained two GMM's each with 50 components and spherical covariance matrices using these latent encodings. Evaluating the distribution in Eq. (4.3) for the testset from December 13^{th}, we find an AUC of 0.64 which is far below the both the reconstructive based outlier detection from Paper I and the clustering of Paper IV. However, considering that we have not put any effort into fine-tuning the approach, this

may serve as an affirmation that the AECM approach could be a promising solution. The joint distribution, $P(\text{kin}, \text{pos})$, requires a small change in the architecture of the AECM model, as the latent encodings of the positional and kinematic parts must be independent but concatenated for the clustering module. Thus, we need a two-branched architecture with two input and output branches for the positional and kinematic features, respectively, and the latent encodings of each branch are concatenated for the clustering module. A sketch of this updated network is shown in Figure 4.2.

4.3 General Discussion

In this section, we discuss the research contributions, how they advance the research objective presented in Section 1.2 and the limitations of the research presented in this thesis.

A) Research, develop, and evaluate deep learning frameworks for describing normal maritime behavior using historical AIS data to support maritime situation awareness of surveillance operators.
This thesis found that deep learning models may be used to describe normal maritime behaviour and that they natively address many of the issues of other data-driven methods for constructing the normal maritime picture. However, application of these models to support maritime situation awareness of surveillance operators is not straight forward.

Contribution 1 advances this objective from the perspective of automated abnormality detection as a way to evaluate the learned normalcy model. It is important that the dataset used for evaluation is representative of the true real-life traffic situation and that the abnormalities reflect actual abnormal behavior of operational interest. For this purpose, we have released an annotated dataset for public use. However, data sets with more types of abnormal behavior need to be made public to perform a comprehensive evaluation of the detection models.

Accurate evaluation of detection models is very important for operational use in order to have accurate estimates of the false positives and negatives. False negatives may have large societal, environmental, or fatal consequences and reduction of them is the most important issue in operational use. Normally, reducing the number of false negatives requires acceptance of false positives. However, detection models with an abundance of false positives may suffer from "alarm blindness", where operators slowly disregard all alarms by the system, since in their experience they are mostly incorrect. For this reason operators may tend to favor man-in-the-loop style detection systems.

Contribution 3 addresses objective A from this perspective. Extraction of the learned normalcy model from deep neural networks in order to facilitate man-in-the-loop style detection requires large amounts of interpretability. Contribution 3 clearly shows this as a major limitation of current application of deep neural networks. In future work, ideas for extracting and presenting the learned normalcy should be included already in the design phase.

This thesis only considers a methodology for evaluating deep learning methods for enhancing situation awareness and does not consider how this may be presented or visualized for operators. Clustering facilitates a very intuitive detection scheme and seem like a natural choice for presenting the normal maritime picture. Thus, the combination of deep learning and clustering should be a focus of future work.

A natural limitation of using historical AIS data is that many smaller vessels are not outfitted with or switch off their AIS transponders or spoof their AIS signal. Many abnormalities of operational interest might be carried out by so-called dark vessels that have turned off their AIS transponder [MI News Network, 2022, Kroodsma et al.]. To overcome these limitations, new data sources must be used such as radar or satellite imagery, e.g., synthetic aperture radar (SAR).

B) Development of deep learning methods for the automated detection of maritime abnormalities.
This thesis found that ensemble models may be utilized for a more robust detection of different types of abnormal behaviour. As such, this objective is addressed by contribution 2, which provides an ensemble model that outperforms the state-of-the-art. The models considered in this study are limited to analysing trajectories individually and may benefit from the addition of other members detecting other kinds of abnormal behavior. For example, models that facilitate vessel-2-vessel interaction may detect collisions and rendezvous situations before they occur.

Similarly, other inputs may be considered to reduce the amount of false positives or to detect abnormalities that depend on certain external conditions. However, it is important that the addition of these does not have negative effects on the scaleability of the models and the impact of the added data sources must be documented using the datasets with real-life abnormalities.

C) Study the generalization of deep learning methods for automated detection of maritime abnormalities over time and across geographical area.
As discussed previously, ship behavior change over time and differs between geographical areas. Deep learning models facilitate easy adjustments of normalcy models to new areas or seasons through fine-tuning of existing models.

Contribution 1 address this objective by evaluating the abnormality detection performance of models trained on different geographical areas and subject to different degrees of fine-tuning. This thesis found models trained for point predictions require very little degree of fine-tuning if any in order to achieve similar performance of abnormality detection. Although very little fine-tuning is required, a good amount of historical data is required to construct the reference of predictive/reconstructive errors required for abnormality detection. Therefore the abnormality detection systems investigated in this thesis will not function in regions with no historical data.

D) Methods to describe abnormal local behavior not constrained to major shipping lanes.
This thesis aims to construct normal models that disentangle positional and kinematic information such that kinematic behavior may be described independently from the location it happens. Contributions 4 addresses this objective by providing a two-step clustering approach where the second stage clusters are able to differentiate between kinematic behavior. It is natural to seek a combination of this clustering approach with reconstruction based detection using deep neural networks. As such, contribution 1 also addresses this objective to some extent.

The limitations of the data discussed above apply equally to the proposed clustering approach. Additionally, while deep learning models have been studied using data recorded over longer periods, clustering is based only on a single month of data. Thus, the generalization of the normal model in time remains an open question. Similarly, each normal model is restricted to the ROI in which it is trained and has to be retrained for every new ROI.

Chapter 5

Conclusion

The main objective of this study was to assess the feasibility of deep learning techniques for enhancing the situation awareness of maritime surveillance operators. Surveillance operators likely use level 2 situation awareness to mentally construct a model of the normal maritime picture and base their decision of normality on comparing observed behavior with the established normal behavior model. Therefore, this thesis aims to provide level 2 situation awareness using automated tools to build a model of the normal maritime picture.

Deep neural networks are used to construct maritime normalcy models which are evaluated from an operational perspective. Networks are evaluated based on the performance of automated abnormality detection and the interpretability of learned normalcy models to facilitate man-in-the-loop style detection. The networks are trained using historical AIS data which are freely available in large quantities in many areas of the world.

This thesis shows that deep neural networks address many of the issues with data-driven methods for constructing the normal maritime picture and that deep neural networks can learn a representation of the normal maritime picture using historical AIS data. However, most of the previous works have focused on the creation of more complex models, and little work has focused on the evaluation of the learned normalcy model. Annotating abnormal behavior related to a collision accident, it was found deep normalcy models may be used for automated

detection of maritime abnormalities. Using different architectures of deep neural networks in an ensemble, it is possible to detect different types of abnormal behavior, improving the state-of-the-art.

Using a random resampling strategy during training it is possible to train models which generalize to different resampled periods. Thus, the same model may be used to evaluate abnormalities that evolve on different time-scales. Further, deep representations of the normal maritime picture may generalize to different ROI reducing the need for training area specific models. However, it is still necessary to obtain some historical data from the new ROI in order to create a reference of historical errors for comparison with errors of new traffic.

Although deep neural networks are found to be effective for automated detection of maritime abnormalities, clustering of stochastic latent variables is ineffective in extracting and visualizing the learned normalcy model. Latent variables were found to encode information related to the current position in the main shipping lanes of the ROI. Information on such a large time scale might not be useful for man-in-the-loop style detections, which require knowledge regarding the kinematic evolution on a much smaller local time scale.

There is a general lack of normal models that focus on local kinematic behavior which is often only incorporated during the abnormality detection step. Using clustering in a two-step process, it is possible to disentangle positional and kinematic features and achieve normal models which separate trajectories based on the local kinematic behavior. Clustering provides clear advantages over deep neural networks for visualization and interpretation purposes, however, scales very poorly with large training data. Thus, a combination of the two-step clustering result with deep neural networks remains an open but interesting problem.

Overall, this thesis has discussed methods to support level 2 situation awareness of maritime surveillance operators. While deep neural network has clear benefits, the application in an operational setting is still an issue due to lack of interpretability and that previously suggested models generally focus on automated detection over man-in-the-loop style detection. Therefore, the following extensions to the work are suggested:

- The creation of more annotated data sets for evaluation of different types of annotated abnormal behavior will allow for a better understanding of the trained normalcy models and potentially highlight limitations in the behavior which may be flagged using deep neural networks.
- Development of additional models using different architectures may detect abnormal behavior different from previously suggested networks and im-

prove the detection performance of the collective ensemble. For example, the use of graph neural networks may allow modeling of vessel-2-vessel interactions, which could help detect collision and rendezvous scenarios.

- The use of historical AIS has a clear disadvantage, as abnormal behavior of the highest priority often occurs without AIS. Therefore, AIS-based detection systems will fail to detect these abnormalities, which may have large societal consequences.

- Enhance the AIS data set by including relevant external parameters. Including additional parameters might allow for a more complete normalcy model which may account for different contextual situations. Before inclusion of these additional features it is important to evaluate their relevancy e.g. using automated detection. Additionally it is important that they can be collected and used in real time and still allow training to scale to large data sets.

- The ship type is very important for the detection of abnormal behavior, as a certain type of behavior is allowed for some ship types but not for others. Additionally, reconstructive performance were found to differ heavily between ship types, resulting in some ship types being prone to false negative/positives. Therefore, it is important to incorporate the type of vessel into the detection process and possibly into the normalcy model.

- Although the normal maritime picture of deep neural networks was found to transfer to new ROIs without the need of extensive fine-tuning, the ROIs studied share many similarities in terms of the expected behavior. Therefore, the generalization to new ROIs with major changes in expected behavior remains to be studied.

- Deep learning methods trained for joint reconstruction and clustering. Such models may be used for automated detection using reconstruction errors and facilitate extraction and man-in-the-loop style detection using clustering.

- Deep learning architectures using global and local latent variables may facilitate a two-fold normal picture which can be used to detect global and local abnormalities.

- This thesis suggests that automated tools for deriving a maritime picture should enhance level 2 situation awareness of maritime surveillance operators. However, this claim remains unverified. Experiments involving surveillance operators should be conducted in order to verify the impact of automated tools for deriving the maritime normal picture and how these are best presented. Additionally, involvement of surveillance operators would be beneficial in defining user and operational requirements.

Chapter 6

A Review of Current Deep Learning Techniques for Maritime Abnormality Detection and Directions for Future Progress

---

Kristoffer Vinther Olesen[a] · Anders Nymark Christensen[a] · Sune Hørlück[b] · Line Katrine Harder Clemmensen[a]

[a] Technical University of Denmark, DTU Compute, Kgs. Lyngby, Denmark
[b] Terma A/S, Lystrup, Denmark

Publication Status: Paper is submitted and under peer-review for publication in *Transport Reviews*

Abstract: Increasing worldwide maritime traffic makes maritime safety and security of growing importance. Surveillance operators monitor and predict emerging critical situations based on a wide range of data sources from many vessels within a large sea area. This makes operators prone to mistakes due to cognitive overload, time pressure, and fatigue. To support operators, methods and systems capable of performing anomaly detection have received increased attention. In this work, we provide a review of deep learning techniques applied to the problem of maritime abnormality detection. We find a general lack of evaluation of proposed methods caused by the lack of established state-of-the-art baseline methods and open source benchmark data. We compare the performance of unsupervised anomaly detection based on reconstruction/prediction using five different Recurrent Neural Network architectures from the literature. We find poor performance of all suggested approaches for detection of real-life anomalies. Reconstruction-based (AutoEncoder) methods outperform predictive methods, but still suffer from too many false positives to be practical in operational use. We recommend future research focus on evaluation and comparison of proposed methods, development of larger benchmarks, and aims to establish ensemble models for anomaly detection and incorporation of the ship type into the abnormality detection.

6.1 Introduction

The oceans remain one of the most important ways to connect and sustain a growing globalised world. Maritime shipping is the most efficient and cost-effective form of long-distance transportation and is responsible for 80% of the world's trade [IMO, 2021]. The recent geopolitical issues in Ukraine have shown the necessity of free trade and transportation. The restriction of free trade and transportation of food, feed materials, fuel, energy, etc. have propelled into a food- and energy-crisis due to rising commodity costs [World Food Program, 2022], and even after an agreement to open Ukraine's Black Sea Ports for food transports the conflict still prevented shipowners due to fear risking their vessels [Bloomberg News, 2022]. Furthermore, oceans currently contribute 17 % of the world's production of edible meat, and it is estimated that this production could increase by $36 - 74$ %, compared to current yields, by 2050 [Costello et al., 2020]. Various threats such as collisions, smuggling, piracy, overfishing, maritime pollution, sabotage of infrastructures, etc. endanger activities at sea and an increasing usage of the oceans will only increase the risk of these threats [Statista Research Department, 2021]. This makes maritime safety and security key issues and, for this purpose, real-time delivery of maritime situational maps is a necessity for Search-and-Rescue (SAR) and law enforcement activities.

Anomaly detection is one of the most important tasks within the domain of maritime safety and security. Current operational systems rely strongly on human experts. Surveillance operators monitor and predict emerging critical situations mentioned above from a large number of vessels within a large sea area. As the amount of data increases, it becomes increasingly difficult for operators to process the amount of information due to a number of factors such as cognitive overload, time pressure, fatigue, and uncertainty in addition to the complex and heterogeneous nature of the data. In order to provide support for the operators, methods and systems capable of performing anomaly detection is a very active research area.

Initially designed for collision avoidance, the Automatic Identification System (AIS) has quickly become the main source of trajectories for maritime surveillance. Every day, AIS provides on a global scale hundreds of millions of messages [MarineTraffic, 2016], which contain the identifiers of the ships, their coordinates from the Global Positioning System (GPS), their speed, course, etc. In many areas this data is freely available and may be collected into large amounts of historical maritime trajectories. Previous works [Riveiro et al., 2018, Sidibé and Shu, 2017] highlight the need for scaleable models that are applicable to and may improve with very large training sets. Thus, it seems there is a great potential in combining large amounts of AIS trajectories with deep learning techniques. However, application of deep learning techniques for detection of abnormal maritime trajectories remains limited and is yet to be fully utilised.

A major barrier to the application of deep learning to maritime abnormality detection is the lack of large, freely available annotated data sets [Riveiro et al., 2018]. Since AIS was originally designed only for collision avoidance, it does not contain any metadata that could be used for anomaly detection. In addition, most surveillance operations are conducted by law enforcement or military personnel. Therefore, any information related tomanual detection of abnormal behavior is often classified for security purposes. This makes labeling of a large amount of abnormal trajectories very difficult and limit the possible approaches to unsupervised methods. The lack of annotated data sets limits evaluation and comparison of different models to qualitative comparison of the discovered abnormalities. This approach is highly problematic since it ignores potentially important false negatives.

In this work, we provide a review of recent deep learning methods applied to maritime abnormality detection and compare the abnormality detection performance of the most prevalent methodologies. The aim of this review is to provide an early idea of working techniques and identify short-comings and open questions for future work addressing the challenges identified in [Riveiro et al., 2018, Sidibé and Shu, 2017]; scaleability, interpretation of alarms, online detection, and incorporation of relevant external features. Pang et al. [2021]

provides a general review of deep learning techniques for abnormality detection and highlights AutoEncoders [Nguyen et al., 2021, Murray and Perera, 2021] and predictability modelling [Forti et al., 2020, Capobianco et al., 2021b] as widely used methods for learning generic feature representations useful for abnormality detection. These models are tested and compared on data in which abnormal trajectories related to a collision accident have been manually labelled. Based on our findings we propose guidelines for the future work of maritime abnormality detection.

Our contributions are:

- A review of existing deep learning approaches for maritime abnormality detection.
- Comparison of maritime deep learning methodologies for reconstruction/predictive based outlier detection.
- We provide empirical evidence towards the short-comings of the investigated models.
- Guidelines for future work within the field of deep learning based maritime abnormality detection.

The paper is organised as follows: In section 6.2 we discuss previous applications of deep learning for abnormality detection of maritime trajectories. Section 6.3 is a more general discussion of deep learning methods for the analysis of maritime trajectories. Section 6.4 compares the abnormality detection performance of five deep learning model on real-life abnormal trajectories from a ship collision. Finally, we present our conclusion in Section 6.5.

6.2 Maritime Abnormality Detection

Maritime anomaly detection can be separated into two main approaches: knowledge based or data driven. Knowledge-based approaches look for predefined patterns in the data source, while data-driven approaches derive a model representing the normal picture and evaluate anomalies as observations deviating from normalcy.

The majority of knowledge-based approaches implement well-defined rule systems. Having rules makes detections very easy to interpret and analyse since operational systems may output which rules were broken and supply operators

with a clear reason for an alert in real time. However, rules may vary heavily between time of year or geographic location, making a relevant and exhaustive list of rules difficult to construct and implement in operational systems.

Traditionally, the most common way of expressing the normal maritime picture is through density based clustering. Pallotta et al. [2013b] presented the widely used TREAD method to cluster trajectories into traffic routes used for anomaly detection and trajectory prediction. TREAD is a point-based method and extracts coordinates of new entries, exits, and stops within the region of interest (ROI). These points are clustered using density-based clustering (DBSCAN Ester et al. [1996]), to form waypoints in which ships enter, exit, or stop within the ROI. A route between waypoints is formed when a certain number of transitions between the waypoints have been observed. Yang et al. [2022b] and Zhao and Shi [2019a] take another approach to density-based clustering. Here, the trajectories are reduced using the DP algorithm [Douglas and Peucker, 2011] and similarity is measured using the Hausdorff distance or dynamic time warping before being clustered by DBSCAN.

6.2.1 Deep Learning for Maritime Abnormality Detection

The wide availability of AIS based maritime trajectories have recently spurred an increase in applications of deep learning techniques to describe the normal maritime picture for the purposes of maritime abnormality detection. While the applications of deep learning are still new and remains limited, we believe the field would benefit from a discussion of the diverse approaches to abnormalities, evaluation, and detection models. The problem of maritime abnormality detection can be separated into three main elements; What constitutes an abnormality, how the normalcy model is constructed, and how detections are made. A summary of how previous deep learning applications have approached these three elements can be found in Table 6.1 and will be discussed in detail in the following subsections.

6.2.1.1 Abnormalities and Evaluation

A precise definition of what constitutes maritime abnormalities is very difficult to make. Today, most maritime surveillance operations are conducted manually by military or law enforcement. Thus, the reasons for detection of an abnormality are often classified information. Additionally, abnormal behaviour is highly dependent on factors such as time, location, and ship type. For this reason, it

Work	Abnormalities	Normalcy model	Detection
Singh and Heymann [2020a]	Labelled abnormal inputs	Artificial Neural Network	Prediction
Wang [2020]	Simulated abnormalities	Artificial Neural Network	Prediction
Liu et al. [2022a]	Labelled abnormal inputs	Recurrent Neural Network	Prediction
Zhang et al. [2021b]	Labelled abnormal inputs	Recurrent Neural Network	Prediction
Hu et al. [2022]	Labelled abnormal inputs	Ensemble*	Prediction
Protopapadakis et al. [2017]	Unsupervised	Density based clustering	clustering outliers
Huang et al. [2021]	Unsupervised	Recurrent Neural Network	Global threshold
Zhao and Shi [2019b]	Unsupervised	Recurrent Neural Network	Global threshold
Singh et al. [2021]	Unsupervised	Recurrent Neural Network	Global threshold
Xia and Gao [2020]	Unsupervised	Bayesian Recurrent Neural Network	Global threshold
Liu et al. [2022b]	Unsupervised	Graph Neural Network	Global threshold
Eljabu et al. [2021]	Unsupervised	Graph Neural Network	Global threshold
Nguyen et al. [2019, 2021, 2020]	Unsupervised	Variational Recurrent Neural Network	A-Contrario
Zang et al. [2021]	Unsupervised	Variational Recurrent Neural Network	A-Contrario

Table 6.1 – Summary of previous deep learning based approaches to maritime abnormality detection. $\star$: Graph Neural Network and LSTM based Variational AutoEncoder ensemble members.

is often not possible to obtain a large list of annotated maritime trajectories for training or evaluation.

In the work of data-driven automated abnormality detection there are generally three ways of defining abnormalities; Simulated abnormalities Wang [2020], self annotated abnormalities Singh and Heymann [2020a], Liu et al. [2022a], Hu et al. [2022], and unsupervised anomalies Nguyen et al. [2021], Zhao and Shi [2019b]. The general lack of labelled data sets causes the majority of previous work to approach maritime abnormality detection as an unsupervised problem. In order to avoid unsupervised methods, some use self-labelling based on added noise, rules, or clustering. The use of self-labelling allows direct classification of trajectories as normal or abnormal, and networks can be trained using the cross-entropy loss.

Wang [2020] simulate outliers by adding noise to their normal observations and predicts the probability of abnormality directly using the simulated labels. Singh and Heymann [2020a] suggest a model for detection of AIS on/off switching. They propose to re-sample incoming AIS messages to every two seconds and treat missing data as abnormal cases of on/off switching. Trajectories with continuously missing data are labelled as abnormal. Liu et al. [2022a] self annotate the dataset based on extreme position, speed, or course values decided using the COLREGS international collision avoidance rules for the area. Zhang et al. [2021b] cluster the data using DBSCAN and propose a Long-Short Term Memory (LSTM) model to classify abnormal trajectories identified by DBSCAN. Hu et al. [2022] self annotate their data set based on large deviations in position, speed, or course.

Self-annotated or simulated data sets may fully utilise the power of supervised deep learning models and are trained purposefully for abnormality detection.

Additionally, evaluating their performance is trivial. However, it is important that the models are trained on meaningful outliers, and while extreme values or observations with added noise are outliers compared to the training set, they might not define abnormal trajectories of operational interest to surveillance operators. In addition, the process of manual annotation or annotation using clustering suffers from scaleability issues negating the modelling power of deep neural networks applied to very large data sets.

To utilise the immense modelling power of deep neural networks applied to large training sets, an unsupervised approach is needed. In unsupervised models, a normalcy model is learnt on the training set. It is then assumed that abnormal observations are poorly modelled by the normalcy model and abnormal trajectories can be detected based on the loss. Unsupervised approaches can be easily trained on very large data sets and use big data to learn the normalcy model. A drawback of unsupervised methods is that they flag statistical outliers as abnormal. This means that they can potentially learn to model abnormal data if it occurs frequently, causing false negatives. Thus, evaluation and comparison of unsupervised methods is an important aspect for them to find practical use.

The most common way to evaluate unsupervised methods is through qualitative examples and cases, but simulated anomalies using added noise or comparison with rule-based systems have also been suggested. Nguyen et al. [2021] provide a qualitative description of the trajectories flagged as abnormal. This is further expanded in Nguyen et al. [2020] where experts affirm the flagged trajectories and provide contextual explanation for the abnormalities. Zhao and Shi [2019b] provide evaluation of their proposed algorithm by a walk-through of a hand full of detections. Singh et al. [2021] simulate a single case by translating normal data to another position and provide a qualitative description of the detections. Huang et al. [2021] simulate abnormal trajectories by adding noise to the position and speed inputs and compute accuracy and false positive rate. Eljabu et al. [2021] and Liu et al. [2022b] compute the accuracy of the detection using labels obtained from real world context and news or by a rule based anomaly detection toolkit, respectively.

Maritime abnormality detection suffers from the lack of standardised data sets that authors may use to evaluate and compare models. Evaluation using simulated data with noise suffers from the same issues as a model trained using simulated data. There is an inherent risk that the simulated abnormalities have lost physical meaning, which makes the task trivial and without practical application. Similarly, an evaluation using the detections of rule-based systems as the ground truth would result in models that are only capable of reproducing rule-based detections and not providing any new information about the problem. Verification of detections by the means of qualitative examples or cases works to illustrate the potential type of abnormal behaviour that could be flagged.

However, it completely negates the issue of false negatives, which in a military or law enforcement operation is of far greater importance. In this case, false negatives could potentially be both trajectories showing similar behaviour to the ones detected, but also trajectories with a completely different behaviour that is of interest for operational use. For this purpose, the generation of smaller manually annotated data sets for evaluation may be the superior approach. These data sets should ideally be annotated by subject matter experts but using contextual knowledge of the environment and news events, one might be able to create hand annotated data sets that is expected to have an operational interest. In section 6.4, we compare different unsupervised models for abnormality detection on a manually labelled data set with verified abnormal activity.

6.2.1.2 Models and Detection

Several different neural network architectures have been suggested for the detection of maritime abnormalities. However, as discussed previously, the lack of a standardised data sets for evaluation purposes makes it very difficult to compare the performance of different models. For supervised tasks, both feed-forward artificial neural networks (ANN) [Singh and Heymann, 2020a, Wang, 2020] and recurrent neural networks (RNN) [Liu et al., 2022a, Zhang et al., 2021b] have been suggested. Hu et al. [2022] train an ensemble of Variational AutoEncoders (VAE) based on LSTM and a Graph Variational AutoEncoder based on trajectory similarities. Each member of the ensemble is trained to reconstruct the input trajectory. Reconstruction errors are then combined using the Twin Delayed Deep Deterministic Policy Gradient Algorithm (TD3) to make a final binary prediction of the abnormality.

Protopapadakis et al. [2017] suggest using AutoEncoders (AE) as feature extractors for density-based clustering. They find that features extracted using AEs are superior to principal component analysis or the raw input. Huang et al. [2021], Zhao and Shi [2019b] train LSTM models for one-step predictions of the position and kinematic values and abnormal trajectories are detected by a global threshold on the prediction error. Liu et al. [2022b] also base detection on one-step predictions, but suggest a self-attention graph neural network in which each trajectory is transformed into a graph with every node corresponding to one time stamp. The attention scores are then used to update a Gated Recurrence Unit for one-step prediction. Singh et al. [2021] suggest training a bidirectional LSTM to output the parameters of a Normal inverse-Gamma distribution of future position and kinematics. The predictive probability of the location, speed, and course is then used to optimise the network. Abnormalities are defined as those that experience a significant increase in predicted variance. Xia and Gao [2020] suggest a Bayesian RNN to model the probability of reconstruction and

detect abnormal trajectories based on a global threshold. Nguyen et al. [2019] suggest a Variational Recurrent Neural Network (VRNN) approach in a multi-task fashion. The proposed network is used for trajectory prediction, ship type classification, and anomaly detection. Anomaly detection is further expanded in Nguyen et al. [2021, 2020]. Instead of a global threshold for detection, an A-Contrario detection is suggested taking into account the model performance in a local area to determine detection thresholds. Zang et al. [2021] test the VRNN approach suggested in Nguyen et al. [2021] in a case study in another maritime area and based on qualitative investigations find the method generalizes here.

Instead of modelling individual trajectories, Eljabu et al. [2021] detect abnormalities in traffic patterns over larger time scales of weeks or months. They propose a graph neural network where semantic stopping points in the ROI, e.g. ports, form nodes in a graph. The number of transitions between these nodes over a long period of time is then embedded using graph convolutions. The embeddings are assumed to be normally distributed and outliers declared as abnormal transition patterns.

The use of a global threshold for abnormality detection may cause a bias in sparse regions in the input space. Since the model will naturally have better performance in regions with more training data, the use of global thresholds will cause an increase of detection in regions with lower modelling performance. The A-Contrario detection methods overcome this issue by only considering the reconstruction/predictive errors of observation in the local vicinity when deciding on the threshold. The time complexity of A-Contrario detection is quadratic wrt. to the trajectory length. Thus, there may be a need for real-time detection methods.

Interpretation of detections has been largely ignored by previous works. We suggest to investigate four approaches. **1.** For automated models to be useful in an operational setting, they must be able to explain why a certain trajectory was flagged as abnormal. Deep neural network are infamous for their general lack of interpretability but this is a very active research area and we recommend state-of-the-art methods (see e.g., Samek et al. [2019]) be adapted for maritime abnormality detection. **2.** Generative distributions, as suggested in Nguyen et al. [2019], Singh et al. [2021], provide uncertainty estimates on reconstruction/prediction for operators to assess model confidence. **3.** Latent variable models Nguyen et al. [2021], Hu et al. [2022] could be used to extract physical interpretable information from the latent space which could be used to explain model predictions. **4.** Statistical detection methods such as A-Contrario can be applied on top of other models and tuned with a specific false positive ratio in mind, giving operators a tool to control the number of detections.

6.3 Applications of Deep Learning for Analysis of Maritime Trajectories

Deep Learning methods have been applied to a wide variety of problems associated with maritime trajectories. The most common problems are trajectory prediction and anomaly detection, but other applications include; shipping lane recognition [Yao et al., 2017], collision risk [Dan Vukša et al., 2022, Namgung and Kim, 2021], track association [Yu et al., 2020], traffic flow prediction [Zhou et al., 2020, Mandalis et al., 2022], ship type classification [Duan et al., 2022, Liang et al., 2021], prediction of vessel service time [Abualhaol et al., 2019], and classification of vessel states [Chen et al., 2020, Mantecon et al., 2019, Ferreira et al., 2022]. Table 6.2 summarises recent applications of Deep Learning methodologies. A more detailed table of all references can be found in the supplementary material.

Problem - # Works	Model - % of works	Input Features - % of works
Abnormality Detection - 16	RNN - 37.5% AutoEncoder - 37.5% ANN - 12.5% Graph Neural Network - 12.5%	Position - 100% Kinematic - 93.8% Time - 25% Environment - 6.2% Size - 6.2%
Track association - 1	AutoEncoder - 100%	Position - 100% Time - 100%
Collision risk - 2	RNN - 100%	Feature Engineering - 100%
Shipping lane recognition - 1	AutoEncoder - 100%	Feature Engineering - 100%
Traffic flow prediction 2	RNN - 100%	Flow matrices - 50% Position - 50% Time - 50%
Trajectory prediction - 31	RNN - 48.4% ANN - 19.4% Seq-2-Seq - 16.1% AutoEncoder - 9.7% Transformer - 3.2% Graph Neural Network - 3.2%	Position - 96.8% Kinematic - 61.3% Time - 16.1% Destination - 9.7% Vessel-Vessel Interaction - 6.5% Size - 3.2% Draught - 3.2% Radar Imagery - 3.2%
Vessel classification - 2	RNN - 50% AutoEncoder - 50%	Position - 100% Kinematic - 50% Size - 50% Draught - 50%
Vessel service time prediction - 1	RNN - 100%	Feature Engineering - 100%
Vessel state classification - 3	CNN - 66.7% RNN - 33.3%	Kinematic - 100% Position - 66.7%

Table 6.2 – Summery of deep learning methods and input features used in analysis of problems with maritime trajectory data.

In the previous section, we discussed how the normalcy model for unsupervised abnormality detection can be constructed both as a predictive or reconstructive model. Thus, any model that can be converted into a trajectory prediction or

reconstruction model can potentially be used for abnormality detection. In this section, we review and discuss the wide range of deep learning models and the related learnings for analysis of maritime trajectories.

6.3.1 Application to Abnormality Detection

As discussed previously, trajectory prediction can be extended to the detection of abnormal trajectories based on predictive errors. Unsupervised abnormality detection using reconstructive or predictive methods may in theory be able to detect all kinds of abnormal maritime behaviour. Other applications may be extended similarly or directly interpreted as detection of specific types of maritime abnormalities. The risk of collision [Dan Vukša et al., 2022, Namgung and Kim, 2021] itself can be considered an abnormality and is closely related to the rendezvous case where two or more vessels meet at sea. This form of behaviour is expected in port areas or other locations where vessels may accumulate, but in other areas only pilot boats are expected to approach other vessels. Classification of the ship type [Duan et al., 2022, Liang et al., 2021] based on the trajectory may be used to detect vessels that behave unexpectedly for their shipping type and are trying to hide their true identity. Prediction of the vessel state Chen et al. [2020], Mantecon et al. [2019], Ferreira et al. [2022] may be used to detect illegal fishing activities in exclusion zones. Track association [Yu et al., 2020] may be used to verify incoming AIS messages and detect cases of spoofing. Finally, while the methods discussed so far detect abnormalities on an individual scale, traffic flow prediction [Zhou et al., 2020, Mandalis et al., 2022] may be used to detect potential abnormalities on a population scale.

6.3.2 Deep Learning Models

Early applications of Deep Learning were mainly focused on feed forward ANN's for prediction of the future position [Zissis et al., 2015, Gan et al., 2018, Wen et al., 2020]. Since then, several different architectures have been suggested for various applications. Convolutional Neural Networks (CNN) have been the most widely applied model for vessel classification [Duan et al., 2022, Nguyen et al., 2019] or vessel state classification [Chen et al., 2020, Mantecon et al., 2019], suggesting that the entire trajectory is needed for a precise classification.

For trajectory prediction, the most applied architecture has been RNNs either as one-step predictions [Liu et al., 2021, Sørensen et al., 2022], direct multi-step predictions [Chondrodima et al., 2022, Mandalis et al., 2022, Spadon et al., 2022], or iterative one-step prediction [Forti et al., 2020, Capobianco et al., 2021b, Dijt

and Mettes, 2020] in a Sequence-2-Sequence fashion. Liu et al. [2022d] suggest a Graph Convolutional Network. At each time step the authors construct a graph in which nodes denote the current location of vessels in the ROI and the future position is then predicted using graph convolutions. Nguyen and Fablet provide an initial preliminary study on the feasibility of Transformers for trajectory prediction. Transformers have become state-of-the-art within Natural Language Processing but applications for trajectories are still limited.

AutoEncoder models have been suggested for different purposes of maritime trajectory analysis but are especially appropriate for abnormality detection and feature extraction for further downstream tasks. Yao et al. [2017] and Murray and Perera [2021] suggest using a Recurrent AutoEncoder (RAE) for extraction of features useful for clustering. In a RAE the encoder and decoder are LSTM/GRU cells with the last hidden state being used as the latent variable. However, Murray and Perera [2021] impose an additional variational constraint, (RVAE), on the latent space assuming it can be modelled using a standard normal distribution. Yu et al. [2020] focus on the reconstruction errors of the trajectories for the purpose of verifying new AIS updates. If a newly received update is a true continuation of the previous trajectory, then the trajectory can be reconstructed with a very low error, whereas false updates will cause a large reconstruction error. Hu et al. [2022] suggest a similar approach for abnormality detection; using a VAE-LSTM model each trajectory update, x_t, is encoded into a latent variable, z_t. The sequence of latent variables is then used as inputs for a decoder reconstructing the original trajectory. Nguyen et al. [2021], Ding et al. [2020] suggest the closely related Variational Recurrent Neural Network (VRNN) model. In the VRNN model, both the encoder and decoder depend on the same hidden state of a unidirectional LSTM. In addition, the encoder depends on the current input x_t and the decoder on the current latent variable z_t. At each time step the recurrence is then updated using x_t and z_t. Although abnormality detection using AEs has mainly focused on the reconstruction error as in Hu et al. [2022] and Nguyen et al. [2021], RAE and RVAE models as suggested in Yao et al. [2017], Murray and Perera [2021] may also be able to detect outliers depending on latent variables.

Several different combinations of recurrent architectures have been suggested in the literature. Although GRU cells have been suggested in both unidirectional [Suo et al., 2020] and bidirectional settings [Wang et al., 2020], most works suggest either unidirectional [Gao et al., 2021, Qian et al., 2022] or bidirectional [Park et al., 2021, Yang et al., 2022a] LSTM cells for temporal modelling. A few extensions to the basic RNN structure have also been proposed. Spadon et al. [2022] suggest a hybrid solution where each recurrence is preceded by a 1-dimensional convolution. Capobianco et al. [2021b] suggest the use of an attention mechanism to allow the decoder to focus more easily on specific time updates of the input during prediction. Liu et al. [2021, 2022c] suggest an

encoder/decoder structure in which the encoded latent vector from multiple ships is averaged and used to initialise the decoder. This is supposed to make the decoder able to take into account not just the previous trajectory of the current ship being predicted but also all other vessels in the vicinity.

While RNNs allow for variable-length input sequences, it seems valid to assume that the position multiple hours ago has very little impact on predictions in the short term. Therefore, many works in literature limit the input length and treat the window size as a hyperparameter to tune. The limitation on the input window is often referred to as the context length. Gao et al. [2018], Qian et al. [2022] find a context length of 3 time steps sufficient to accurately predict the next position, while Sørensen et al. [2022], Forti et al. [2020] suggest a much longer context length of 20 time steps. Other works [Liu et al., 2021, Yang et al., 2022a, Suo et al., 2020] completely remove the windowing constraint and instead use the entire previous trajectory to predict the next position. In addition to the context length, Seq-2-Seq models vary by the number of time steps predicted, referred to as the prediction length. Capobianco et al. [2021b], Forti et al. [2020], and Dijt and Mettes [2020] suggest a prediction length of 12, 20, and 50 time steps, respectively.

The effect of increased context/prediction length on the prediction error remains unclear as there is no established state-of-the-art nor baseline data sets used for comparison of models. In addition, most papers differ in their proposed preprocessing or data filtering, making it even more difficult to evaluate the effect of context length. Spadon et al. [2022] investigate different values of context/prediction length with the purpose of finding the optimal model architecture for different data complexities, but do not investigate the effect of varying the context length on the prediction error. For the purpose of anomaly detection, the context/prediction length can have quite an impact. An increased context length may allow us to successfully predict more complicated manoeuvres and behaviour, reducing the number of false alarms. On the other hand, a very large context length may oversaturate the models with data unrelated to the abnormality decision. Similarly, the prediction length must have a suitable value. A very long prediction length is unwarranted because it requires the operator to wait a long time before a decision can be made. However, a very short prediction length may result in noisy prediction and more false alarms.

While many different architectures and hyperparameters have been proposed, in particular for trajectory prediction, it remains difficult to compare their performance against each other due to the lack of baseline data sets. Most works in the literature conduct ablation studies and compare their proposed architecture with basic RNNs using unidirectional LSTMs or GRUs. In Nguyen and Fablet, Transformers are found to outperform the Seq-2-Seq and RVAE models suggested in Forti et al. [2020], Capobianco et al. [2021b], and Murray and Perera

[2021] but the performance on large complex data sets remains to be studied in depth. Spadon et al. [2022] compare performance of different recurrence cells in stacks up to three, as well as in unidirectional and bidirectional settings. They find that the proposed CNN-RNN hybrid stabilises the performance across data complexities and improves feature extraction for multiple vessels of different types. However, in general, they find little difference in terms of the Root Mean Squared Error (RMSE) between models, and the variance over multiple trainings is larger than the differences between models.

6.3.3 Input Features for Deep Neural Networks

Using the AIS system, ships are required to broadcast dynamic information related to movement, such as vessel position, speed, course, heading, rate of turn, and current navigational status. This information is submitted every 2-10 seconds while underway and every 3 minutes while at anchor. In addition, static information about the ship is broadcasted every 6 minutes. This static information contains the call sign and name of the vessel, the type of ship, the dimension of the ship, the position of the positioning system on the ship, the draught of the ship, the current destination of the ship and the estimated time of arrival.

As stated above, the AIS information is transmitted at irregular intervals depending on the speed of the vessel. Additionally, the AIS system is inherently noisy and some messages may not be received causing gaps in the data stream. In order to account for the irregular sequences, most works resample and interpolate the AIS data-stream to some predefined frequency. For the purpose of abnormality detection, this resampling frequency is expected to be very important. At longer time steps between updates we might fail to detect small maneuvers but a fast resampling might cause trajectories to be very long and training to be computationally heavy. Additionally, a fixed interval sampling makes it difficult to detect trajectories with missing fragments as the missing segment is simply interpolated between normal segments. Instead of resampling the AIS data stream to regular time intervals, some works suggest incorporating the time into the input features of the models. This can be done either by the raw time [Yu et al., 2020, Gao et al., 2018] or by the time step from the previous update [Spadon et al., 2022]. This requires some form of trajectory compression or sampling in order to avoid computationally infeasible training due to long trajectories. Gao et al. [2018] suggest a trajectory compression algorithm in which only relevant AIS updates are retained. Yu et al. [2020] sample a predefined number of AIS messages from each trajectory and Spadon et al. [2022] suggest sampling a set of windows of consecutive AIS updates of a predefined length from each trajectory. For trajectory prediction tasks, the window sam-

pling technique may be practical, since we are interested in predicting the future in the same time-step size as the input. However, in abnormality detection we may be interested in evaluating the trajectories with different step sizes, thus the random step size of the fixed number sampling may be preferable.

The dynamic information from AIS messages can be processed into trajectories of latitude and longitude, however, it varies whether or not kinematic information such as speed, course, and heading is included. Many trajectory forecasting problems suggest only using positional information [Forti et al., 2020, Capobianco et al., 2021b, Chondrodima et al., 2022] whereas other works also include speed, course and/or heading [Murray and Perera, 2021, Gao et al., 2018, Zhang et al., 2021a]. For trajectory prediction tasks, the inclusion of kinematic information is logical if the objective includes the prediction of future speed and course; however, in the case of abnormality detection, this distinction is not as clear. Zhao and Shi [2019b] and Huang et al. [2021] both use kinematic features for trajectory prediction, but only consider the prediction error of the positional features for the purpose of abnormality detection. At low speeds the prediction error of the position must be expected to be very low. This could lead to an increase in false negatives unless other kinematic features are included in the abnormality detection. Conversely, the course over ground is known to be noisy at low speeds which could lead to an increase in false positives. To allow for easier modelling of the course, it has been suggested to preprocess the feature into sine and cosine values [Murray and Perera, 2021] which potentially could be extended to velocity vectors. We believe that there is a lack of study into the impact of different input features and their preprocessing on the abnormality detection results, particularly using only positional features vs including speed and course.

In most applications, static information is disregarded, and only dynamic information is used. However, a few previous works discuss the value of contextual information added from static messages. Capobianco et al. [2021b] include the final destination as an optional categorical variable in the decoder of a Seq-2-Seq model predicting the future position and report an approximate 50 % increase in performance. Capobianco et al. [2021b] consider only two possible shipping lanes and use the exit points of these shipping lanes as the final destination categories. This distinction may not be scaleable to global or large ROIs with many different shipping lanes, harbours, and traffic that does not follow the well-defined shipping lanes. First, the discovery of exit points requires a clustering of AIS messages similar to the procedure suggested in TREAD [Pallotta et al., 2013b]. Secondly, in order to be used in real-time it requires a mapping from the final destination reported using AIS to the discovered exit points. Duan et al. [2022] find that the size and draught of the ship are important features in classifying the type of vessel. Inclusion of ship dimensions and/or external weather data for trajectory prediction or abnormality detection might seem in-

tuitive as smaller vessels are capable of tighter maneuvers and extreme weather might lead to a higher number of abnormal trajectories. However, according to our personal feedback from maritime surveillance operators, the effect of the environment on maritime trajectories is minor. Huang et al. [2021] include ship dimensions and weather data for each time step in the input trajectory to detect abnormal trajectories based on one-step predictions of position and speed. Detecting trajectories with added noise, Huang et al. [2021] report a minor increase in performance.

While static information from AIS messages may be generalised to full trajectories, the acquisition of external weather data is less trivial. Weather data are generally not recorded at sea and is typically reported less frequently than AIS messages using average and extreme values for the previous time-period. Thus, weather data may require both spatial and temporal interpolation in order to be included as features in the input trajectory as suggested in Huang et al. [2021]. Alternatively, weather data could be included using a separate RNN encoding external weather data with a different update frequency or as categorical variables. Additionally, real-time trajectory prediction applications might require access to weather forecasts for long-term predictions. More studies are needed to fully evaluate the effect of including external weather data for trajectory prediction and abnormality detection. For this purpose, the validation data should consist of a large ROI and time period in order to ensure significant variation in the weather.

Most applications limit themselves to individual trajectories and ignore vessel interactions with each other. However, a couple of different methods for including vessel interactions has been suggested. Liu et al. [2021, 2022c] suggest accounting for vessel interactions by incorporating a repulsive force from nearby vessels. At each time step a weighted sum of the LSTM hidden states of nearby vessels is calculated. This weighted sum is used as an additional input feature for the LSTM of the original vessel being modelled. While this makes it possible to model vessel interactions, it may be difficult to scale to very large data sets with long trajectories. The method requires a large batch size, preferably the entire data set, to ensure proper vessel-to-vessel interaction. Secondly, the inclusion of the hidden states as a feature makes it impossible to parallelise over time during training. Dijt and Mettes [2020] include a sequence of radar images centred on the modelled vessel to provide context regarding the local environment including the shoreline and nearby vessels. To ensure that the radar images encode physical information, the encoding of the radar images is used both as features for the LSTM encoder and to create a mask for land masses. This approach may work for large ships with on-board radar capabilities centered on their own position but applications to off-site surveillance centers when radar images are off center is less clear.

For the purpose of abnormality detection, it is necessary to study the effect of including additional input features on the type of behaviour flagged as abnormal. Features such as final destination, external weather data, nearby vessel positions, etc. all seem intuitive to include, but it is extremely difficult to evaluate their value without a publicly available baseline validation set. Even if these features are discovered to be useful for abnormality detection, it may not be trivial to include them in a real-life operational setting.

6.3.4 Limitations of the Training Data

To a large extent, the literature has focused on the design of increasingly deeper and more complex models. However, as highlighted by Spadon et al. [2022], the literature lacks investigation of the application of these models to streaming data reflective of data expected in real life operational settings. Many works limit the size of training data [Capobianco et al., 2021b, Dijt and Mettes, 2020], only consider a small ROI [Liu et al., 2022a, Zhou et al., 2020], or limit the analysis to a single-ship type [Sørensen et al., 2022, Forti et al., 2020]. From personal interviews with maritime surveillance operators, we learnt that models trained on such data sets generalise poorly to the data observed in real-life applications. In particular, this is an issue for unsupervised abnormality detection. Out-of-sample trajectories are expected to have higher error and tend to be flagged as abnormal, increasing the amount of false positives. Therefore, it is important to evaluate methods on data sets representative of expected traffic in terms of ROI size, time period, and ship types.

Other methods constrain the application compared to having more complete data sets. Zhao and Shi [2019b], Murray and Perera [2021] suggest clustering trajectories into clusters representing the major shipping lanes and train separate trajectory prediction networks for each cluster. In both methods, the outliers found in the clustering are disregarded in the following prediction model. If applied to the detection of abnormalities, this approach could be sensitive to minor abnormalities in the major shipping lanes, but it completely disregards vessels that do not conform to the shipping lanes. Additionally, the training of the clustering algorithm may not be scaleable with large amounts of data, making this approach less desirable. Nguyen et al. [2021], Lu et al. [2021] apply a discretization of the features. Each feature is converted into a one-hot encoded vector based on grid values, which are then concatenated to the final input. This discretization creates a trade-off similar to the resampling frequency of the trajectories. If the resolution of the grid is too coarse, we can not detect small maneuvers, but a very fine grid or very large ROI might cause the dimensionality of the data to be too large and computationally heavy. The proper size of the grid resolution is also connected to the resampling frequency. If we choose

a high resampling frequency, we might be forced into choosing a finer grid in order to ensure frequent movement between grid cells, further increasing the computational complexity.

Osekowska et al. [2017] finds that maritime normality pictures change significantly over time, which requires the abnormality detection models to be updated frequently. Deep learning models natively allow for retraining of model weights in light of a changed normalcy picture, new data, or for application to new ROIs. Hu et al. [2022] apply a linear data transformation to allow for transfer learning of trained abnormality detection models to new ROIs. They find that transfer learning can heavily reduce the amount of training needed.

6.3.5 Summary

The previous literature has mainly focused on the design of more complex models for both abnormality detection and other problems within maritime deep learning. Trajectory prediction often serves as the first step in an abnormality detection model and many other problems can also be interpreted as abnormality detection focusing on one special type of abnormal behaviour. However, the question of how these models compare with each other quantitatively and qualitatively when applied to abnormality detection remains open. Likewise, it remains an open question what impact the choice of architecture, hyperparameter values, resample period, and input features have on the abnormality detection. An issue of particular interest to surveillance operators is the general lack of interpretability of model detections. Based on these findings, we present a few guidelines for future work in the detection of maritime abnormalities using deep reconstructions/predictions.

- Models should be trained and evaluated on data sets representative of the expected traffic in real-time applications.
- Abnormality detection models should focus on unsupervised methods.
- Preprocessing steps should be kept to a minimum unless a change in the abnormality detection performance can be verified.
- We propose a study into the impact of additional inputs such as kinematic AIS information, weather data, vessel-to-vessel interaction, etc. on the behaviour flagged as abnormal.
- We suggest future model designs include ideas of how prediction interpretability may be achieved.

6.4 Experiments and Results

In order to make a through investigation of the current deep learning methods, we evaluate the abnormality detection performance of five different deep learning architectures suggested in the literature to review their abilities of abnormality detection in a real application. We evaluate a basic Seq-2-Seq predictive model Forti et al. [2020], a Seq-2-Seq model with attention Capobianco et al. [2021b], a RVAE reconstructive model Murray and Perera [2021] with (RVAE) and without (RAE) variational regularisation, and a VRNN reconstructive model Nguyen et al. [2021]. Unless otherwise detailed the architectures, preprocessing and hyperparameters used are the same as suggested in the original works. All networks are trained using Adam optimiser with a learning rate of 0.001 for 50 epochs using a batch size of 200 for the Seq-2-Seq models, 300 for the RVAE, and 32 for GeoTrackNet. The prediction length of the Seq-2-Seq models is limited to 25 updates to provide an adequate number of time steps for abnormality detection and keep detections close to real time. The context length is initially limited to 25 updates.

The models are trained and evaluated on the data sets released by Olesen et al. [2022a,b]. The test set labels abnormal traffic related to a collision accident. In addition to the colliding vessels, the abnormal trajectories include commercial traffic which had to deviate from the planned course to avoid the accident, Search-and-Rescue activities and law enforcement vessels responding to the accident, and any vessel taking part in the following search of two sailors thrown overboard. This is a small data set of only 521 trajectories and 25 labelled abnormalities, however, the abnormalities are confirmed to be of operational interest. In the future, we recommend the creation of additional data sets with labelled abnormalities of operational interest.

Abnormalities are discovered using the A-Contrario detection suggested in Nguyen et al. [2021]. Unlike Nguyen et al. [2021], we do not impose a fixed threshold for detection. Instead, we calculate the receiver operating characteristic (ROC) and measure the abnormality detection performance by the area under the ROC curve (AUC).

6.4.1 Comparison of Methods

Figure 6.1 compares the ROC curves of the five models. GeoTrackNet has the highest performance measured by the AUC followed by the RAE without variational constraint which achieves almost the same AUC. The two Seq-2-Seq models achieve a similar performance only slightly less that the AutoEncoder

models. The RVAE model achieves the worst performance, significantly underperforming compared to the other suggested architectures. However, all five methods suffer from false positives to a degree that would make them impractical in a real world setting. In order to detect 80% of all abnormal trajectories, validation suggests that a false alarm is triggered on every fifth trajectory.

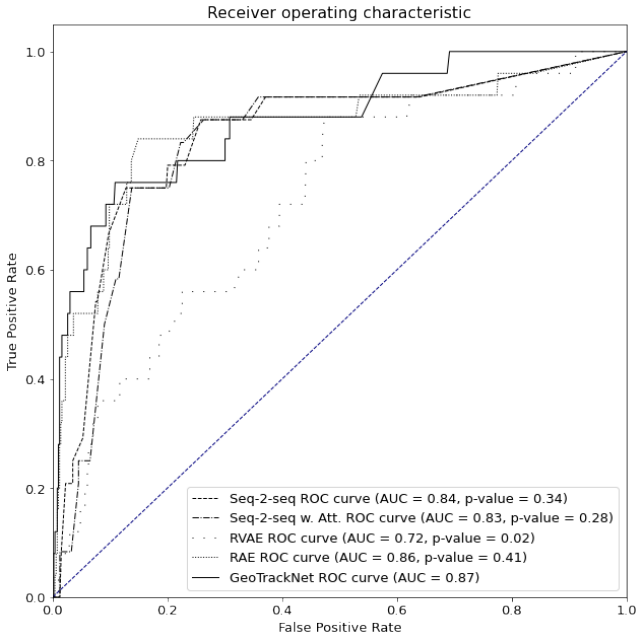
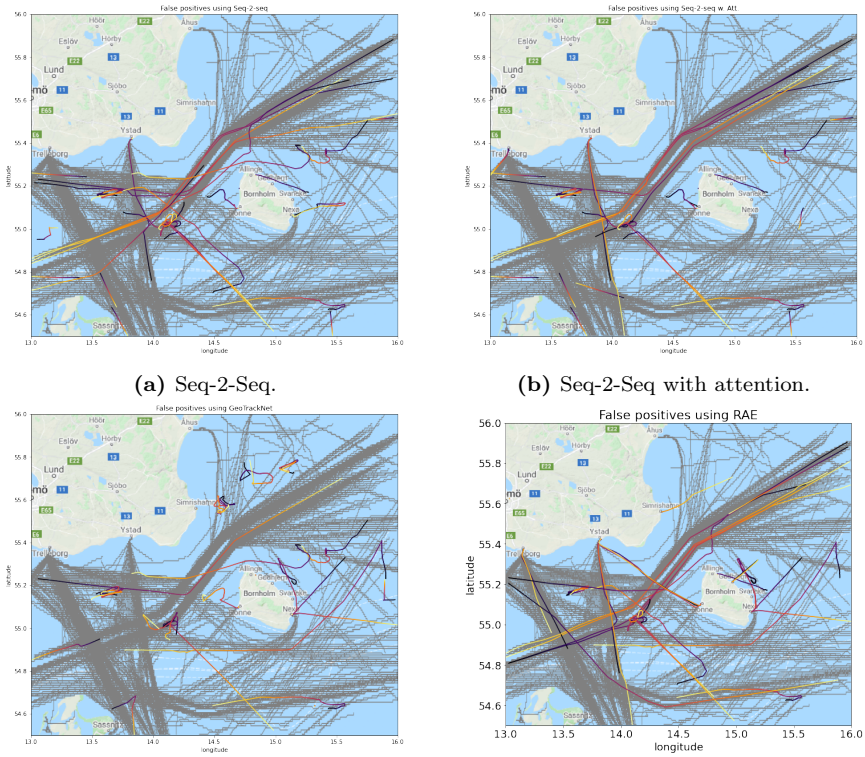


Figure 6.1 – Receiver operating characteristic (ROC) curves. P-values for the hypothesis of equality with the highest measured AUC is given [Hanley and McNeil, 1982].

Figure 6.2 shows 30 false positives from the four models with the highest AUC. False positives in the Seq-2-Seq models in figures 6.2a and 6.2b generally fall into the 3 categories; very short trajectories, trajectories with a highly varying course, or trajectories sailing along the shipping lanes but showing minor course adjustments. For very short trajectories, the separation into contextual and prediction parts may result in contextual parts that are too short for the model to learn meaningful behaviour resulting in poor predictions and trajectories being flagged as abnormal. The other two categories indicate that the models have learnt to linearly project the future positions as if the ship continues the current course. As such the models have not learned to model any turns apart from perhaps the well established turns in the large shipping lanes. The false positives using GeoTrackNet, Figure 6.2c, are generally trajectories with many course changes. GeoTrackNet also flags stationary trajectories as abnormal due

to their highly varying course when drifting. This indicates the GeoTrackNet model may be susceptible to flag trajectories with frequent course changes. The RAE model, Figure 6.2d, flags trajectories that follow the shipping lanes or has widely varying course in close proximity to the shipping lanes. This kind of behaviour may be of big interest to the maritime safety as ships drifting near the shipping lane increase the risk of collisions.



(a) Seq-2-Seq.

(b) Seq-2-Seq with attention.

(c) GeoTrackNet.

(d) RAE.

Figure 6.2 – The top 30 false positive using the suggested models.

Figure 6.3 shows all five reconstructions of two abnormal trajectories. The RVAE model makes the same reconstruction for both trajectories. Thus, the model has not learnt meaningful feature representations and has collapsed into a solution where all trajectories are mapped into the same point in the latent space. This results in large reconstruction errors for trajectories that cover a large distance and a long duration, while short trajectories close to the centre of the ROI have smaller reconstruction errors. The top case depicts a cargo ship sailing along the shipping lane. The ship suddenly makes a double u-turn before continuing out of the ROI. The Seq-2-Seq models predict the ship to continue

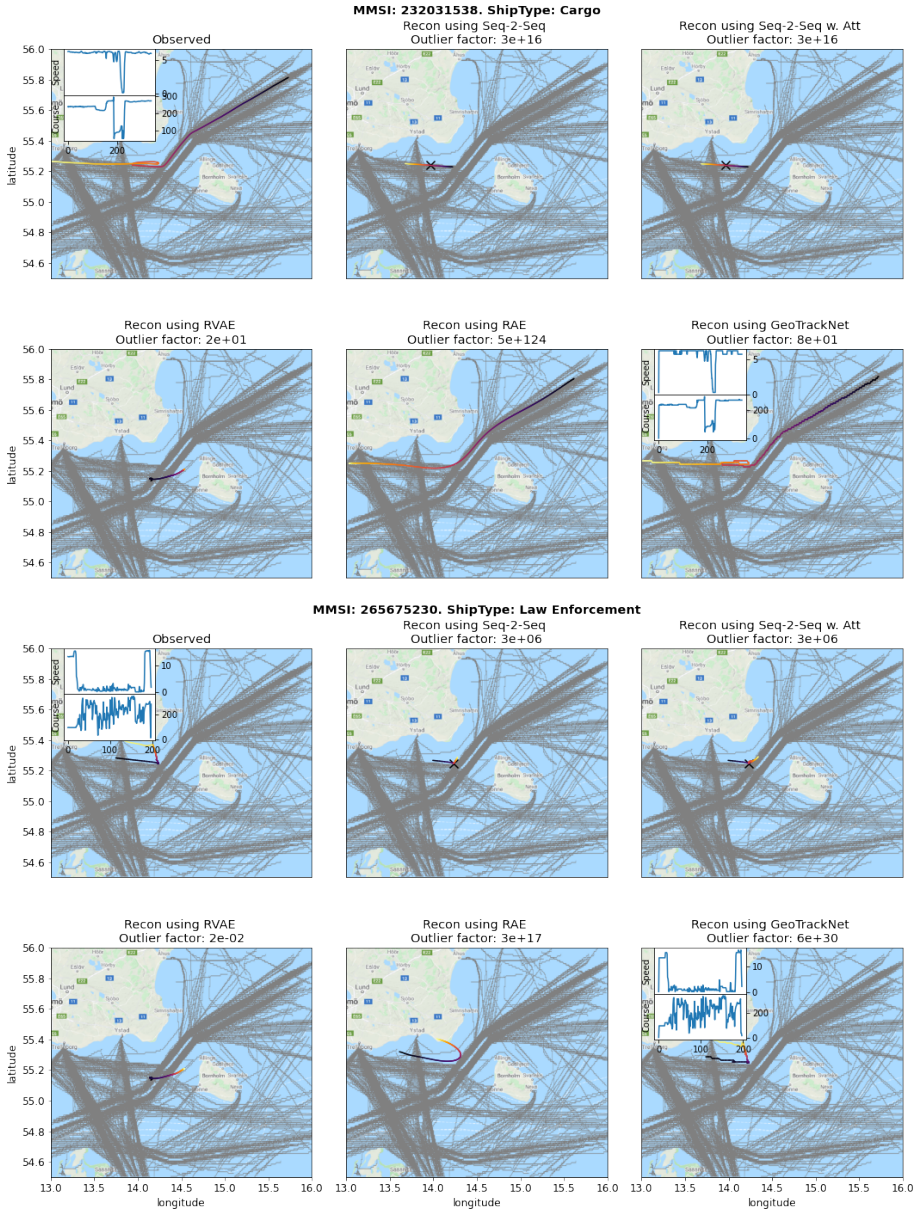


Figure 6.3 – Normal trajectories flagged as abnormal by the two Seq-2-Seq models. The models have learned to predict the vessels continue on the current course resulting in false positives. Trajectory origin is denoted in black.

on course which causes large prediction errors at the time of the u-turn and the trajectory is flagged as abnormal. Similarly, the RAE model is capable of mapping out the total route but the abnormal behaviour is not reconstructed causing a large reconstruction error and an alarm being raised. Contrarily, GeoTrackNet reconstructs the u-turn behaviour and does not flag the trajectory. A double u-turn in the shipping lane is clearly abnormal behaviour, and that it is not flagged by GeoTrackNet is very surprising. The bottom case shows a trajectory sailing parallel and close to the shipping lane at high speed before abruptly stopping. It then continues north at very low speed and turns along the coast when reaching land. The sub-segment with the highest prediction error using the Seq-2-Seq models is the segment in which the trajectory slows down near the shipping lane. Here the models predict the trajectory to continue along the shipping lane but at reduced speed causing the smaller prediction error resulting in a false negative prediction. This is an example of abnormal activity that will not be detected because it is conducted at slow speeds. The RAE model reconstructs the general route of the trajectory, but since this abnormality generally occurs away from the shipping lanes, the outlier score is much lower. GeoTrackNet has a high reconstruction error due to the highly varying course and correctly flags the trajectory as abnormal. The RAE and GeoTrackNet models generally detect very different behaviour as abnormal. Therefore, we hypothesise a combination or an ensemble of these two models would be a better outlier detector than each model individually.

Figure 6.4 shows box plots of the outlier score split by ship type. We identify two major issues with the proposed detection of abnormal trajectories. Different models will result in outliers scores on different scales maybe even with an order of magnitude difference. This makes it difficult to compare the degree of abnormality across models. Additionally, there are also large differences in outlier scores between ship types. Ship types with a less restricted behavioural patterns have higher outlier factors than ship types typically restricted to the shipping lanes. Therefore, deciding on a fixed threshold for detection across all ship types may result in some ship types being prone to false positives/negatives.

The difference in outlier scores may cause the creation of an ensemble to be a non-trivial task. Taking the average of the outlier scores may not be a fair measure due to the large-scale difference. The RAE obtains outlier scores that are one order of magnitude larger than GeoTrackNet, so this model may dominate when calculating the average. Majority voting of the models' binary abnormality decisions is also not trivial for the same reason. The differences in scale warrant the need for different detection thresholds for each model participating in the voting. Taking the average of the reconstruction loss is also not possible due to differences in loss functions. GeoTrackNet is trying to maximise the reconstruction loss, while the RAE model is trying to minimise the reconstruction loss. Thus, taking the mean of the reconstruction error would average out the

differences. The different loss functions are due to the different inputs used. Thus, in order to have an ensemble using the mean of the reconstruction errors the models would have to use the same input. However, deciding an input type for all models warrants a study of how each model interacts with different input types and if there are any pros/cons with each suggested input.

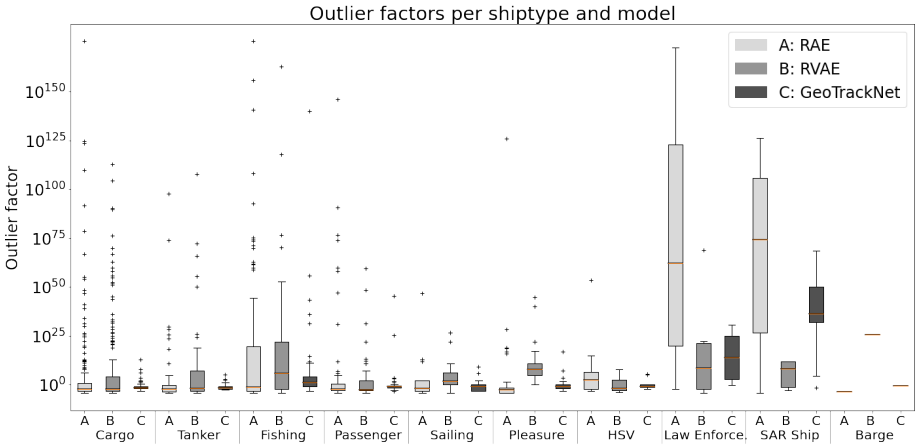


Figure 6.4 – The outlier scores calculated on the test set separated by ship type and model. HSV: High Speed Vessel

6.5 Conclusion

In this work, we review and discuss maritime abnormality detection using Deep Learning. We have provided an overview of methods applied in the literature along with the pros and cons of each method. We find a wide variety of different deep learning architectures have been applied for different problems using maritime trajectories. Most applications rely solely on positional and kinematic data from AIS messages, but other features including environmental information and radar imagery have been suggested. However, we find a general lack of evaluation of the proposed methods caused by no established state-of-the-art, baseline methods, or data sets and suggest the community to work on establishing benchmark data sets and baseline methods as a top priority. Supervised and unsupervised methods have been suggested for the problem of abnormality detection. Since large labelled datasets for training are difficult to achieve, labelled datasets are often based on simulated data or extreme values which might not reflect true abnormalities. Based on these findings, we recommend abnormality detection models should focus on unsupervised methods, preprocessing steps should be kept to a minimum, and the impact of additional external data

sources needs to be evaluated on an independent test set representative of the expected data in real-time applications. Additionally, there is a need to focus on the interpretability of detections for models to be applicable in real-life operations.

We compare the performance of maritime abnormality detection of five reconstructive / predictive models suggested in the literature. The AutoEncoder-based GeoTrackNet and RAE models have the highest performance measured by the AUC followed by the predictive Seq-2-Seq method. We hypothesise that an ensemble of the two different AutoEncoder models (RAE, GeoTrackNet) would have better abnormality detection performance than any of the individual models. We have shown that ship types and different models may have outlier scores on different orders of magnitude, meaning that some ship types may be prone to false positives/negatives, and an ensemble may be dominated by a single model. In future work, we recommend that the use of ensembles are studied further. This will require investigations into both inputs and outputs to make sure models are directly comparable.

6.6 Appendix

Work	Problem	Model	Inputs	Data Limitations
Singh and Heymann [2020a]	Abnormality Detection	ANN	Pos/Kin	-
Wang [2020]	Abnormality Detection	ANN	Pos/Kin/ΔT	Small dataset
Liu et al. [2022a]	Abnormality Detection	RNN	Pos/Kin	Small area
Zhang et al. [2021b]	Abnormality Detection	RNN	Pos/Kin/T	Small dataset
Hu et al. [2022]	Abnormality Detection	VAE-LSTM & Graph Neural Network	Pos/Kin/ΔT	Small dataset
Protopapadakis et al. [2017]	Abnormality Detection	VAE	Pos/Kin	-
Huang et al. [2021]	Abnormality Detection	RNN	Pos/Kin/T/Size/Enviromental	-
Zhao and Shi [2019b]	Abnormality Detection	RNN	Pos/Kin	Clustering
Singh et al. [2021]	Abnormality Detection	RNN	Pos/Kin	Limited ship types
Xia and Gao [2020]	Abnormality Detection	Bayesian RNN	Pos/Kin	-
Liu et al. [2022b]	Abnormality Detection	Graph Neural Network	Pos/Kin	Limited ship types
Eljabu et al. [2021]	Abnormality Detection	Graph Neural Network	Pos/Kin	Small dataset
Nguyen et al. [2019, 2021, 2020]	Abnormality Detection	VRNN	Pos/Kin	Discretization
Zang et al. [2021]	Abnormality Detection	VRNN	Pos/Kin	Discretization
Yao et al. [2017]	Shipping lane recognition	RAE	Feature Engineering	Few Trajectories
Dan Vukša et al. [2022]	Collision risk	RNN	Feature Engineering	Small dataset
Namgung and Kim [2021]	Collision risk	RNN	Feature Engineering	Small dataset
Yu et al. [2020]	Track association	RAE	Pos/T	-
Zhou et al. [2020]	Traffic flow prediction	RNN	Flow matrices	Small area
Mandalis et al. [2022]	Traffic flow prediction	RNN	Pos/ΔT	-
Wen et al. [2020]	Trajectory prediction	ANN	Pos/Kin/Static	Clustering
Zissis et al. [2015]	Trajectory prediction	ANN	Pos/Kin	Discretization
Gan et al. [2016]	Trajectory prediction	ANN	Feature Engineering	Small dataset
Xu et al. [2011]	Trajectory prediction	ANN	Kin	Small dataset
Zhou et al. [2019]	Trajectory prediction	ANN	Pos/Kin	Small dataset
Gan et al. [2018]	Trajectory prediction	ANN	Feature Engineering	Small dataset
Gao et al. [2018]	Trajectory prediction	RNN	Pos/Heading/T	Compression
Suo et al. [2020]	Trajectory prediction	RNN	Pos	Clustering
Gao et al. [2021]	Trajectory prediction	RNN	Pos	Limited ship types
Qian et al. [2022]	Trajectory prediction	RNN	Pos/Kin	Few trajectories
Wang et al. [2020]	Trajectory prediction	RNN	Pos/Kin	-
Yang et al. [2022a]	Trajectory prediction	RNN	Pos/Kin	Small dataset
Liu et al. [2021]	Trajectory prediction	RNN	Pos	-
Liu et al. [2022c]	Trajectory prediction	RNN	Pos	-
Zhang et al. [2020c]	Trajectory prediction	RNN	Pos/Kin	Small dataset
Lu et al. [2021]	Trajectory prediction	RNN	Pos	Discretization
Park et al. [2021]	Trajectory prediction	RNN	Pos/Kin	Clustering
Chondrodima et al. [2022]	Trajectory prediction	RNN	Pos/ΔT	-
Zhang et al. [2021a]	Trajectory prediction	RNN	Pos/Kin/T	Small dataset
Sørensen et al. [2022]	Trajectory prediction	RNN	Pos/Kin	Limited ship types
Dijt and Mettes [2020]	Trajectory prediction	Seq-2-Seq	Pos	Small dataset
Forti et al. [2020]	Trajectory prediction	Seq-2-Seq	Pos	Small dataset
Capobianco et al. [2022]	Trajectory prediction	Seq-2-Seq w. Att	Pos/Destination	Small dataset
Capobianco et al. [2021a]	Trajectory prediction	Seq-2-Seq w. Att	Pos/Destination	Small dataset
Capobianco et al. [2021b]	Trajectory prediction	Seq-2-Seq w. Att	Pos/Destination	Small dataset
Spadon et al. [2022]	Trajectory prediction	CNN-RNN Hybrid	Pos/Kin/ΔT	-
Murray and Perera [2020]	Trajectory prediction	AE	Pos/Kin	Clustering
Murray and Perera [2021]	Trajectory prediction	RVAE	Pos/Kin	Clustering
Ding et al. [2020]	Trajectory prediction	VRNN	ΔPos/Kin/ΔT	-
Nguyen and Fablet	Trajectory prediction	Transformers	Pos/Kin	Discretization
Liu et al. [2022d]	Trajectory prediction	Graph Neural Network	Pos	-
Duan et al. [2022]	Vessel classification	CNN-VAE	Pos/Kin/Size/Draught	Discretization
Liang et al. [2021]	Vessel classification	RNN	Pos	-
Abualhaol et al. [2019]	Vessel service time prediction	RNN	Feature Engineering	-
Chen et al. [2020]	Vessel state classification	CNN	Pos/Kin	Discretization
Mantecon et al. [2019]	Vessel state classification	CNN	Kin	-
Ferreira et al. [2022]	Vessel state classification	RNN	Pos/Kin	Clustering

Table 6.3 – Problems, models, and input features considered by references in this review.

Chapter 7

Detecting Abnormal Maritime Trajectories using Ensembles and Transfer Learning

---

Kristoffer Vinther Olesen[a] · Anders Nymark Christensen[a] · Sune Hørlück[b] · Line Katrine Harder Clemmensen[a]

[a] Technical University of Denmark, DTU Compute, Kgs. Lyngby, Denmark
[b] Terma A/S, Lystrup, Denmark

Publication Status: Paper is unpublished

Abstract: Detection of abnormal maritime behavior detection is a key component of ensuring maritime safety and security. Current operational systems are largely based on manual surveillance of a wide range of data sources from

many vessels within a large sea area. To support operators, methods and systems capable of performing anomaly detection have received increased attention. As abnormal maritime behaviour is a complex and ill defined concept, different model architectures or data preprocessing may be better suited for detection of certain types of abnormal behaviour. In this work, we suggest the use of ensembles consisting of different model architectures to improve detection of abnormal maritime traffic in the wake of a fatal collision accident. We find that ensemble members flag different types of abnormal behavior, and using them in an ensemble setting both increases the chance of detection and the reduction of false positives. We also investigate the possibility of transfer learning to reduce training times of models in new maritime regions and find that models trained for point prediction do not require fine-tuning to achieve the same abnormality detection performance.

7.1 Introduction

Maritime shipping is the most efficient and cost-effective form of long-distance transportation and is responsible for 80% of the world's trade [IMO, 2021]. This makes maritime security and safety crucial to the protection of the global supply chain. Every day, the Automatic Identification System (AIS) provides on a global scale hundreds of millions of messages [MarineTraffic, 2016], which contain identifiers of ships, their coordinates of their Global Positioning System (GPS), their speed, course, etc. In many areas this data is freely available and may be collected into large amounts of historical maritime trajectories for maritime surveillance. Anomaly detection is one of the most important tasks within the domain of maritime surveillance. Current operational systems rely strongly on human experts and to provide support for operators, methods, and systems capable of performing anomaly detection have been a very active research area Pallotta et al. [2013a], Nguyen et al. [2021], Singh et al. [2021].

Maritime abnormal behavior is complex, poorly defined, and may vary widely between regions of interest (ROI), time of year, and ship types. As a result, different models may be preferable to detect certain types of abnormal behaviour Singh et al. [2021], Hu et al. [2022], Olesen et al. [2022c]. Ensembles of different model architectures or data fusion using different feature extractors have been found to improve prediction/detection in many transport domains such as hot spot prediction Jin et al. [2020], Yao et al. [2018], flow prediction Zhang et al. [2020b], Yuan et al. [2020], vessel classification Zhang et al. [2019], and abnormality detection Singh et al. [2021], Hu et al. [2022]. Additionally, analysis of trajectories using different spatial and temporal resolutions may improve performance Zhang et al. [2019, 2020b]. Particularly, for abnormality detection it may

be interesting to evaluate the trajectories using different temporal resolutions Mascaro et al. [2014]. A low sampling frequency might be too slow to adequately capture the trajectories of vessels with frequent course changes, such as fishing or leisure vessels. However, a faster sampling might lead to noisy trajectories and have difficulty modeling frequent course changes leading to false positives.

As mentioned, the behaviour deemed as abnormal may vary heavily between ROIs and time of year. Additionally, patrol vessels, coast guard, or military war ships in large scale anti-piracy operations are not fixed to a predefined area but are moving platforms with a constantly changing ROI. Particularly, the case of international antipiracy operations may see a military warship deployed to an area where little to no historical data is available for model training. These problem require the use of transfer learning to apply large scale model between ROIs and occasional retraining of models for season based pattern fluctuations Osekowska et al. [2017]. Graph- or grid-based models and preprocessing as suggested in Nguyen et al. [2021] may render trained models inapplicable for transfer learning as the model architecture will be tired to chosen size and resolution of the input features.

In this work, we make an analysis and comparison of input features, resulting objective functions, and random temporal sampling of trajectories. The purpose of this analysis is to design ensembles of Sequential AutoEncoder based models with different temporal resolutions, model architectures, and learning objectives. These models will then be trained and applied on different ROIs to investigate the spatial generalization of models and the applicability of transfer learning to reduce training times on new ROIs. We summarize our contributions as follows:

- Analysis of input features and objective functions of deep learning models for the detection of abnormal maritime trajectories.
- We suggest the use of ensembles of different temporal resolutions, models, and objective functions for the detection of abnormal maritime trajectories improving the current state-of-the-art.
- Investigation of the applicability of transfer learning to reduce model training times in new maritime regions.

The paper is organized as follows: Section 7.2 discuss previous work on maritime abnormality detection, particularly related to ensembles and transfer learning. Section 7.3 outlines and discuss the models, learning objectives, and temporal sampling applied in this paper. In Section 7.4 presents our experiments and results within ensemble creation and transfer learning. Finally, we present our conclusion in Section 7.5.

7.2 Related Work

Historically, detection of abnormal maritime behaviour has been implemented using knowledge based approaches that look for predefined patterns in the data source. The majority of knowledge-based approaches implement well-defined rule systems Lauro Snidaro et al. [2012], Snidaro et al. [2015], Neves et al. [2019]. Having rules makes detections very easy to interpret and analyze, since operational systems may output which rules were broken and supply operators with a clear reason for an alert in real time. However, rules may vary heavily between time of year or geographic location, making a relevant and exhaustive list of rules difficult to construct and implement in operational systems. The recent wide availability of open AIS data has made data-driven maritime anomaly detection using AIS trajectories a popular research question for detection of abnormal maritime behaviour. Data-driven approaches derive a model representing the normal picture and evaluate anomalies as observations deviating from normalcy. Traditionally, the most common way of expressing the normal maritime picture is through density-based clustering Pallotta et al. [2013b], Liu et al. [2015b]. Density-based clustering is used to find waypoints within the ROI, and a route is formed between waypoints whenever a certain number of transitions have been observed. Density-based clustering may form the basis for anomaly detection. Pallotta and Jousselme [2015] proposes a two-stage anomaly detection scheme using the extracted routes. First, a trajectory is associated with a route using only the positional part. Afterwards, kinematic outliers are found by comparing the speed and course to the average behaviour of the route. Liu et al. [2015b] proposes to divide the routes into smaller geographical regions and compute the average kinematic values for each split. These values are then used to detect abnormalities Liu et al. [2015a]. Yang et al. [2022b] and Zhao and Shi [2019a] take another approach to density-based clustering. First trajectories are reduced using the DP algorithm [Douglas and Peucker, 2011] and the similarity is measured using the Hausdorff distance or dynamic time warping before being clustered. Zhao and Shi [2019b] extend this clustering for anomaly detection. The discovered clusters are used to train LSTM models for one-step predictions of position and kinematic values, and abnormal trajectories are detected by a global threshold on the prediction error. Recently, deep learning has been proposed as a feature extractor for clustering of maritime trajectories Yao et al. [2017], Murray and Perera [2021]. Clustering in the latent space of recurrent AutoEncoders were find to discover clusters representative of the major shipping routes through the ROI.

Deep learning has similarly been applied to explicit detection of abnormal trajectory. Singh et al. [2021] suggests training a bidirectional LSTM to output the parameters of a Normal inverse-Gamma distribution of the future position and kinematics. The predictive probability of the location, speed, and course is

used to optimise the network and abnormalities are defined as trajectories which experience a significant increase in predicted variance. Nguyen et al. [2019] suggests GeoTrackNet, a Variational Recurrent Neural Network (VRNN) approach in a multitask fashion. The proposed network is used for trajectory prediction, ship type classification, and anomaly detection. The anomaly detection is further expanded in Nguyen et al. [2021] and Nguyen et al. [2020]. Instead of a global threshold for detection, an A-Contrario detection is suggested taking into account the overall model performance in the local area to determine detection thresholds. Zang et al. [2021] test the VRNN approach suggested in Nguyen et al. [2021] in a case study in another maritime area.

Olesen et al. [2022c] provide a general review and comparison of deep learning methods for the detection of abnormal maritime behavior and find that AutoEncoder approaches such as GeoTrackNet Nguyen et al. [2021] or recurrent AutoEncoders Murray and Perera [2021] are preferable to predictive models. Analysis of the type of behaviour flagged as abnormal by each method finds that the two models flag different behaviour and Olesen et al. [2022c] suggest an ensemble of these models may be a better detector than each model individually. The idea of ensembles is also investigated in Hu et al. [2022]. The proposed ensemble consists of a Variational LSTM AutoEncoder (VAE-LSTM) and a Graph Variational AutoEncoder. Each ensemble member is trained to reconstruct the input trajectory, and the reconstruction errors are then combined using the Twin Delayed Deep Deterministic Policy Gradient Algorithm (TD3) to make a final binary prediction of the abnormality. They find ensembles may effectively improve the accuracy of abnormal trajectory detection, however, the proposed fusing method is prone to being dominated by a single member.

Most works employ a resample and interpolation strategy to create a regularly sampled trajectory from AIS data Nguyen et al. [2021], Zhao and Shi [2019b]. For the purpose of abnormality detection this resampling frequency is expected to be very important. A longer time step between updates we might fail to detect small maneuvers but a fast resampling might cause trajectories to be very long and training to be computationally heavy. Additionally, a fixed interval sampling makes it difficult to detect trajectories with missing fragments as the missing segment is simply interpolated between normal segments. Alternatively, the time may be incorporated into the input features of the models Spadon et al. [2022], Yu et al. [2020]. Yu et al. [2020] sample a predefined number of AIS messages from each trajectory and Spadon et al. [2022] suggests to sample a set of windows of consecutive AIS updates of a predefined length from each trajectory. For trajectory prediction tasks, the window sampling technique may be practical since we are interested in predicting the future in the same timestep size as the input. However, in abnormality detection we may be interested in evaluating the trajectories with different step sizes, thus the random step size of the fixed number sampling may be preferable. Mascaro et al. [2014] use

a Bayesian Network to investigate the importance of including a wide variety of factors and the importance of the time scale. Two different networks are trained; one on the track data in its original time series form and one on a track summery, which includes included average speed and course, number of stops, major stopping points, and percentage of time traveling straight. They find both methods are successful in detecting anomalies although they conclude the methods focus on different variables and thus best used in conjunction with one another.

The application of transfer learning in for maritime abnormality detection is very limited. Hu et al. [2022] suggests a data transformation strategy. A simple linear transformation of the AIS trajectories from the target region to the source region is calculated. This is found to speed up the training time in a new ROI; however, it is not investigated whether transfer learning affects trajectories flagged as abnormal.

7.3 Methodology

In this section, we introduce the neural networks, learning objectives, and temporal sampling used in this paper to construct the different ensembles. We then briefly present the anomaly detection method suggested in Nguyen et al. [2021]. First, we define an AIS trajectory:

Definition 1: An AIS trajectory $\mathbf{x}$ of length L is represented by a sequence of time-stamped points collected from the AIS system, that is, $\mathbf{X}_{1:L} = (\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L)$ where $\mathbf{x}_t = (time_t, longitude_t, latitude_t, speed_t, course_t)$ denote the positional and kinematic features extracted from an AIS message at time t.

7.3.1 Neural Networks

In this paper, we shall use two sequential AutoEncoder models previously suggested for analysis of maritime trajectories; the Recurrent AutoEncoder, suggested in Murray and Perera [2021], and the Variational Recurrent Neural Network suggested in Nguyen et al. [2021]

7.3.1.1 Recurrent AutoEncoder

The Recurrent AutoEncoder (RAE Srivastava et al. [2015]) is an AutoEncoder model in which both the encoder and decoder consists of RNN's. The RNN encoder reads an input sequence, $\mathbf{X}_{1:L}$, of length L. The RNN decoder is initialized using the final hidden state of the encoder network, $\mathbf{h}_L$, and outputs a reconstruction of the input sequence.

The encoder is capable of processing variable-length time series, e.g., AIS trajectories, and compressing them into a fixed size vector, $\mathbf{h}_L$. From this representation, the decoder must reconstruct the input sequentially. The network must learn to retain as much mutual information as possible between the input sequence and the compressed representation, $\mathbf{h}_L$. Since $\mathbf{h}_L$ has a fixed lower dimensionality, it is unlikely that the model can learn trivial identity mappings for input sequences of arbitrary length.

The reconstruction can be done either as point predictions $\hat{\mathbf{x}}_t$ or as parameters $\theta_{x,t}$, of probability distribution, $P(\theta_{x,t})$, describing the input. In the case of point predictions, we have

$$\hat{\mathbf{x}}_t = \phi^{pred}(\mathbf{u}_t) \tag{7.1}$$

$$\mathbf{u}_t = g_\theta(\hat{\mathbf{x}}_t, \mathbf{u}_{t-1}), \tag{7.2}$$

where ϕ^{pred} denotes a linear neural network of one layer. In the case of probability distributions, we have

$$\mathbf{x}_t \sim P(\theta_{x,t}) \tag{7.3}$$

$$\theta_{x,t} = \phi^{pred}(\mathbf{u}_t) \tag{7.4}$$

$$\mathbf{u}_t = g_\theta(\mu_{t-1}, \mathbf{u}_{t-1}), \tag{7.5}$$

where μ_{t-1} denote the mean of the distribution $P(\theta_{x,t-1})$. In both cases the decoder hidden state $\mathbf{u}_t$ and input is initialized as $\mathbf{u}_0 = \mathbf{h}_L$ and $\hat{\mathbf{x}}_0 = \mu_0 = \mathbf{0}$ respectively. The decoder RNN, g_θ, is modelled by a Gated Recurrent Unit (GRU) [Chung et al., 2014] as suggested in Murray and Perera [2021]. Learning is done by maximizing the probability of reconstruction.

$$\mathcal{L}(\mathbf{X}_{1:L}) = \sum_{t=1}^{L} p(\mathbf{x}_t|\theta_{x,t}) \tag{7.6}$$

7.3.1.2 Variational Recurrent Neural Network

The Variational Recurrent Neural Network (VRNN, Chung et al. [2015]) is an RNN that at each time step consists of a Variational AutoEncoder (VAE) condi-

tioned on the recurrence model. The prior distribution on the latent stochastic variable at time t, $\mathbf{z}_t$, is given by a Gaussian with parameters $\boldsymbol{\mu}_{0,t}$ and $\boldsymbol{\sigma}^2_{0,t}$, obtained by a neural network ϕ^{prior} taking as input the recurrent hidden states, $\mathbf{h}_{t-1}$.

$$\begin{aligned} &\mathbf{z}_t \sim N(\boldsymbol{\mu}_{0,t}, \mathrm{diag}(\boldsymbol{\sigma}^2_{0,t})), \\ &\text{where } [\boldsymbol{\mu}_{0,t}, \boldsymbol{\sigma}^2_{0,t}] = \phi^{prior}(\mathbf{h}_{t-1}) \end{aligned} \tag{7.7}$$

Similarly the generating distribution and the approximate posterior $q(\mathbf{z}_t|\mathbf{x}_t)$ is also modelled using neural networks, ϕ^{dec} and ϕ^{enc}, taking $\mathbf{h}_{t-1}$ as an input. The generating distribution also depends on the latent variable $\mathbf{z_t}$, which first passes through a feature extractor $\phi^{\mathbf{z}}$.

$$\begin{aligned} &\mathbf{x}_t|\mathbf{z}_t, \mathbf{h}_t \sim P(\theta_{x,t}), \\ &\text{where } P(\theta_{x,t}) = \phi^{dec}(\phi^{\mathbf{z}}(\mathbf{z}_t), \mathbf{h}_{t-1}) \end{aligned} \tag{7.8}$$

The approximate posterior depends on the input $\mathbf{x_t}$ that is first passed through another feature extractor $\phi^{\mathbf{x}}$.

$$\begin{aligned} &\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_t \sim N(\boldsymbol{\mu}_{z,t}, \mathrm{diag}(\boldsymbol{\sigma}^2_{z,t})), \\ &\text{where } [\boldsymbol{\mu}_{z,t}, \boldsymbol{\sigma}^2_{z,t}] = \phi^{enc}(\phi^{\mathbf{x}}(\mathbf{x}_t), \mathbf{h}_{t-1}) \end{aligned} \tag{7.9}$$

At each time step the recurrence $\mathbf{h_t}$ is updated according to

$$\mathbf{h}_t = f_\theta(\phi^{\mathbf{x}}(\mathbf{x}_t), \phi^{\mathbf{z}}(\mathbf{z}_t), \mathbf{h}_{t-1}) \tag{7.10}$$

The recurrence function f_θ is modelled using an LSTM unit as suggested in Nguyen et al. [2021]. The neural networks ϕ^{prior}, ϕ^{dec}, ϕ^{enc}, $\phi^{\mathbf{x}}$, $\phi^{\mathbf{z}}$ are modelled using two layer linear networks with ReLU activation functions. Learning is done by maximizing the time-stepwise ELBO.

$$\mathcal{L}(\mathbf{X}_{1:L}) = \frac{1}{L}\sum_{t=1}^{L} \mathbb{E}_{q(\mathbf{z}_t|\mathbf{x}_t,\mathbf{h}_{t-1})} p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{h}_{t-1}) - \beta KL\left[q(\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_{t-1})|p(\mathbf{z}_t, \mathbf{h}_{t-1})\right] \tag{7.11}$$

7.3.2 Temporal Sampling

The transmission frequency of the AIS system is related to the current sailing speed, resulting in AIS trajectories with irregular samples. Previous works generally adopt a resample and interpolation strategy to create a regularly sampled trajectory. This approach resamples trajectories to some lower update frequency

using the average value in each interval. Virtual messages obtained by linear interpolation are inserted in case of missing values after resampling. This strategy can introduce uncertainty in trajectories when the gap between two consecutive AIS messages is too large or if the resampling frequency is set too low. In this case the data distribution of the trajectory would be changed and picture an inaccurate trajectory. This would especially be the case for vessels with sailing patterns different from a constant velocity such as fishing and military vessels. To avoid large gaps between consecutive AIS messages, we split trajectories into two if the gap between consecutive AIS messages exceeds 15 minutes.

Using the resample and interpolation strategy, the combination of different temporal resolution would require the training of different models for each frequency. Inspired by Yu et al. [2020] we instead propose to train a single model by sampling trajectories with varying time steps between messages. We propose to sample n points from each trajectory where n is decided based on the length of the trajectory. For short trajectories with fewer than 360 messages we use all messages. The limit of 360 messages was selected because this is the number of messages expected during one hour when using a sampling period of 10 seconds, which is the upper limit of the AIS transmission rate for vessels under way using engine. For trajectories longer than 360 messages, we sample 360 messages pr. 12 hour of duration with a minimum of 360 messages. The purpose of this is to simulate real-time use in which the model may initially be applied to streaming data without temporal sampling. As the trajectory evolves, the sampling period may be increased to compensate for the increased trajectory length. Using this approach the expected time between AIS messages from long trajectories becomes two minutes.

$$n = \max(\text{dur}_{\mathbf{x}}/43200, 1) \cdot 360 \tag{7.12}$$

As discussed in Spadon et al. [2022] modelling of random time stamps is theoretically infeasible since they have a discrete probability distribution. Instead, we consider the elapsed time Δt since the previous AIS message.

At evaluation time, we propose to use the resample and interpolation strategy with a constant time step between AIS messages. In this way, we can evaluate trajectories using the same model with different temporal resolutions.

7.3.3 Data Representation and Objective Functions

Several different data representations have been suggested for the analysis of AIS trajectories. Nguyen et al. [2019, 2021] suggest a discrete encoding based

on concatenated one-hot encodings. Other works in the literature consider continuous inputs with point predictions [Murray and Perera, 2021, Capobianco et al., 2021b] or parameters of distributions [Capobianco et al., 2021a, Sørensen et al., 2022] for prediction/reconstruction. In this section, we shall briefly introduce each data representation and the resulting generative model and objective function.

7.3.3.1 Discrete Encoding

The discrete 4-hot encoding suggested in Nguyen et al. [2019, 2021] discretizes latitude, longitude, speed, and course into grids using one-hot encoding and concatenates these vectors to form a 4-hot encoding. Thus, the 4-hot encoding becomes a concatenation of four categorical distributions. This can be modelled as a special case of a multivariate Bernoulli distribution in which exactly four of the binary variables takes the positive value. In this case, the generating distribution $P(\theta_{x,t})$ in Eq. (7.3) and Eq. (7.8) is given by a multivariate Bernoulli distribution $P(\theta_{x,t}) = \textsc{Bernoulli}(\theta_{x,t})$ with logits $\theta_{x,t}$.

The resolution of the imposed grid naturally plays an important role. If it is too high, the network may require too many computational resources to run and is sensitive to noisy inputs, which could lead to overfitting. Contrary, If the resolution is too low, we may lose critical information. In this work, we use resolutions suggested in Nguyen et al. [2021]. Using the discrete 4-hot-encoding the inputs will be centered using the mean computed by

$$\mu = \frac{\sum_i^N \sum_t^{L_i} \bar{\mathbf{x}}_{i,t}}{\sum_i^N L_i} \tag{7.13}$$

where $\bar{\mathbf{x}}_{i,t}$ denotes the 4-hot encoded vector of the AIS message at time step t from the trajectory i.

The discrete 4-hot-encoding has two major drawbacks. First, it is only applicable to regularly sampled sequences. The discretization of the time feature is computationally infeasible, since irregular sampled AIS trajectories may have very large time steps, causing the dimensionality of the associated one-hot encoded vector to be very large. The temporal resolution could be lowered in order to reduce the dimensionality, but this would cause critical information loss related to the relationship between temporal and spatial features. Secondly, the input features and size of the network weights becomes tied to the design of the grid making it impossible to transfer learned networks onto new area of different sizes.

7.3.3.2 Standardization

The restrictions of the 4-hot encoding related to the discretization of the inputs can be relaxed by using continuous inputs. In this case we assume the generating distribution $P(\theta_{x,t})$ in eq. (7.3) and eq. (7.8) is given by a multivariate normal distribution $P(\theta_{x,t}) = \mathrm{N}(\mu_{x,t}, \sigma_{x,t})$ with mean and covariance matrix $[\mu_{x,t}, \Sigma_{x,t}] = \theta_{x,t}$. We shall impose a regularization and only consider diagonal covariance matrices assuming that the inputs are uncorrelated. Since we assume the inputs are normally distributed, we standardize to zero mean and unit variance using the empirical mean μ_x and standard deviation σ_x. In the case of the course feature, we convert to radians and only center the data since this feature is cyclical.

$$\mathbf{x}'_t = \frac{\mathbf{x_t} - \mu_x}{\sigma_x} \tag{7.14}$$

In the case of irregular sampled data the empirical mean and variance are not given as the trajectories in each epoch varies due to the random temporal sampling. In this case, we estimate the empirical mean and variance by repeating the temporal sampling as described in Section 7.3.2 on each trajectory 10 times, calculating the empirical mean and variance of the sampled trajectories.

In this study, we shall only consider a multivariate normal distribution as the generating distribution for all features. However, the course would be better described by cyclical distribution such as a warped normal distribution or a Von Mises distribution. However, these distributions have computationally expensive probability densities that significantly increase training time.

7.3.3.3 Normalization

Instead of considering complete generative distributions, models may output point predictions, $\hat{\mathbf{x}}_\mathbf{t}$, of the predictive mean. This corresponds to the maximum likelihood estimate of the reconstructed value. In this case, the reconstruction probabilities $p(\mathbf{x}_t|\theta_{x,t})$ and $p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{h}_{t-1})$ in Eq. (7.6) and eq. (7.11) are exchanged for the maximum likelihood estimate using the MSE loss.

$$\mathcal{L}(\mathbf{X}_{1:L}) = \begin{cases} -\frac{1}{L}\sum_{t=1}^{L} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 & \text{for RAE} \\ -\frac{1}{L}\sum_{t=1}^{L} \|\hat{\mathbf{x}}_t - \mathbf{x}_t\|^2 + \beta KL\left[q(\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_{t-1})|p(\mathbf{z}_t, \mathbf{h}_{t-1})\right] & \text{for VRNN} \end{cases} \tag{7.15}$$

As mentioned above, the course feature is cyclical, which could cause major reconstruction errors at the discontinuity. Therefore, we split the course into coordinate directions taking sine and cosine, $(sin(course_t), cos(course_t))$ making the features continuous. Since we make no assumptions as to the underlying distribution of the data, we normalize all inputs to be between $[-1, 1]$. The course values are already normalized to this interval due to being processed using the sine and cosine.

$$\mathbf{x}'_t = 2\frac{(\mathbf{x}_t - X_{\min})}{X_{\max} - X_{\min}} - 1 \tag{7.16}$$

$X_{\min}$ and $X_{\max}$ denote the minimum and maximum feature values observed. As previously, we sample each trajectory 10 times to find the minimum and maximum values when using irregular samples.

7.3.4 A-Contrario Detection

In this work, we shall use A-Contrario detection suggested in Nguyen et al. [2021] for the detection of abnormal behavior based on the reconstruction error. A-Contrario detection divides the ROI into geographical cells C_i. In order to determine whether an AIS message is abnormal, only the local reconstruction errors within the same cell is considered, $l^{C_i}_{\mathbf{x}'_t}$. An AIS message is considered abnormal if the reconstruction error is worse than the $1/p$ -quantile of the local reconstruction errors distributed according to $l^{C_i}_{\mathbf{x}'_t} \sim P^{C_i}$.

Assuming that the events "$\mathbf{x}_t$ is abnormal" are independent in a trajectory $\mathbf{X}_{1:L}$, the probability that at least k out of n AIS messages are abnormal follows the tail of a binomial distribution.

$$B(n, k, p) = \sum_{i=k}^{n} \binom{n}{i} p^i (1-p)^{n-i} \tag{7.17}$$

A segment of a trajectory is considered abnormal if the probability of the observed binomial tail is lower than some threshold. If an abnormal segment exists the entire parent trajectory is denoted as abnormal.

$$\mathbf{X}_{1:L} \text{ is abnormal } \Leftrightarrow \exists (n, k), N_s \cdot B(n, k, p) < \epsilon \tag{7.18}$$

The scaling factor N_s accounts for the number of different subsegments that can be created from the trajectory $\mathbf{x}_{1:T}$ of length T. Its value can be calculated using $N_s = \frac{T(T+1)}{2}$. For more details on A-Contrario detection, we refer to Nguyen et al. [2021]

The quantity $N_s \cdot B(n, k, p)$ denotes the outlier score of the trajectory $\mathbf{x}_{1:T}$ that we can use to gauge the degree of abnormality. We define the outlier factor of the trajectory $\mathbf{x}_{1:T}$ as the inverse of this quantity.

$$OF(\mathbf{X}_{1:L}) \equiv \frac{1}{N_s \cdot B(n, k, p)} \tag{7.19}$$

7.3.5 Ensembles

In the previous sections, we discussed different models, objective functions, and temporal sampling strategies. We propose to create ensembles consisting of these different models, learning objectives, and temporal resolutions in order to provide a more detailed situational picture. We propose to average the outlier factor (7.19) of the ensemble members to evaluate the outlier score of an AIS trajectory $\mathbf{X}_{1:L}$.

7.4 Experiments

We test the models on AIS data from Danish waters around the island of Bornholm. The ROI is a rectangle bounded by $(54.5^oN, 13^oE)$ to $(56^oN, 16^oE)$. The training data were collected from June 1st to November 30th 2021 and contains traffic from 8 different ship types ranging from commercial cargo and tanker traffic to fishing vessels and private sailing and pleasure vessels. The speed was truncated to $20m/s$. Tracks shorter than 10 minutes were discarded, and tracks exceeding 12 hours were divided into smaller tracks, each between 10 minutes and 12 hours. The test data were collected on December 13th 2021. On this day there was a fatal collision accident between two ships that caused several abnormal trajectories. Data from this day have been manually labeled, finding 25 abnormalities out of 521 trajectories. In addition to the colliding vessels, abnormal trajectories include commercial traffic, which had to deviate from the planned course to avoid the accident, SAR and law enforcement vessels responding to the accident, and any vessel taking part in the following search of two sailors thrown overboard. The test data was sampled at Models are evaluated using the receiver operating characteristic (ROC) obtained by varying ϵ in eq. (7.18) and measure the abnormality detection performance by the area under the ROC curve (AUC).

All networks are trained using Adam optimizer with a learning rate of 0.0003 for 30 epochs. After 15 epochs the learning rate was further reduced by a factor of 0.3. The models are trained using a batch size of 300 for the RAE model and

32 for GeoTrackNet. The dimensionality of the stochastic latent space z_t and recurrent hidden state h_t are set to 100 for discrete input features and 20 and 50 respectively for continuous input features.

7.4.1 Analysis of Objective Functions

We compare the abnormality detection performance of GeoTrackNet trained using different generative distributions as described in section 7.3.3 and different inputs. Figure 7.1 shows the ROC curves for the GeoTrackNet model trained using a diagonal Gaussian generative distribution with point predictions of the mean (left) or prediction of the distribution parameters (right). Additionally, the models are trained using only position (green) or position and kinematic input features (orange). Models trained using a discrete multivariate Bernoulli distribution as the generative distribution are shown in blue and serve as a baseline performance.

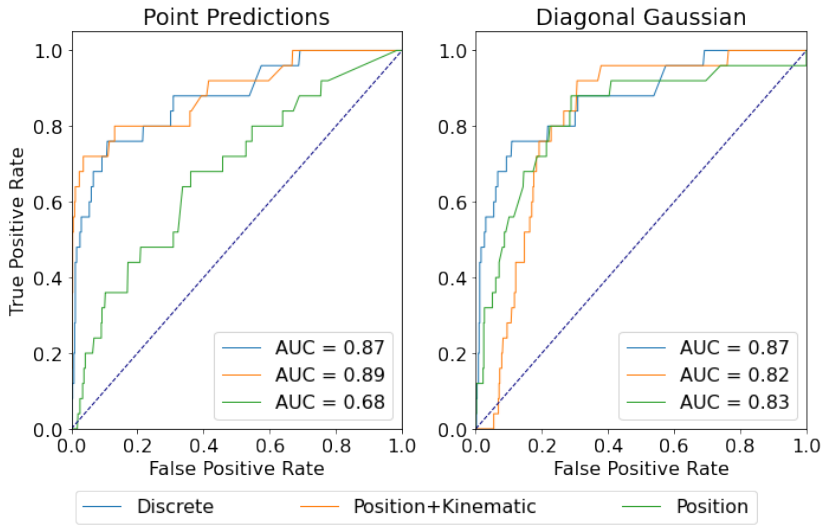


Figure 7.1 – ROC curves from varying the network inputs and generative models. Left: Using point predictions. Right: Using diagonal Gaussian distribution.

The model trained for point predictions using both position and kinematic inputs obtains the highest AUC and has the fewest amounts of early false positive. It achieves a True Positive Rate (TPR) of 60% and a False Positive Rate (FPR) of 0%, i.e., it detects 60% abnormal trajectories before making one false alarm. However, we see the last 5 abnormal trajectories are detected very late. In order

to obtain a TPR of 90% the FPR increase to approx. 45%. The model trained for prediction diagonal Gaussian parameters performs worse than the discrete model and suffers heavily from false positives early in the detection. The FPR is 6% before the first correct detection is made. Interestingly, the model achieves a TPR of 90% before the point prediction model already at a FPR of approx. 30%.

We also note the models trained using only the position as input features decrease in performance for point predictions, but have similar performance and fewer early false positive when prediction distribution parameters. This may indicate that most abnormal behavior is defined by kinematic behavior, and removing these as explicit features may hurt detection performance. On the contrary, providing estimates of the standard deviation may reveal abnormal behaviour due to increased uncertainty in the reconstructed position. Additionally, much normal traffic, such as fishing or sailing vessels, may have frequent speed and course changes that are difficult to model and subsequently flagged as false positives.

Figure 7.2 shows all three model reconstructions (red) of a false negative trajectory (black) using the point prediction model. This is the trajectory of the vessel responsible for the collision during the tow to the port of Ystad. The behavior is fairly predictable; either at drift with little to no course changes or steaming with a constant speed. This is typical behaviour for the false negatives using the point prediction model. Both the multivariate Bernoulli model and the point prediction model reconstruct the trajectory without major errors. However, the diagonal Gaussian model has a high uncertainty in the reconstruction, particularly, of the course, which causes an alarm to be triggered.

Figure 7.3 shows all three model reconstructions (red) of an early false positive (black) using the model predicting the distribution parameters. All three models make good reconstructions of the positional features, and the point prediction model reconstructs the speed and course with only minor errors that cause the trajectory not to be flagged as abnormal. Although the diagonal Gaussian model correctly reconstructs the mean, the standard deviation is very large during the last section of the trajectory. The added uncertainty makes the reconstruction loss high and causes an early detection. The reason for the increased uncertainty may be due to the very high speed of the trajectory. Our best empirical guess of the usual speed of the ships in this section of the shipping lanes is slower at about $6 - 8m/s$. The resolution of the multivariate Bernoulli model causes the speed time-series to be fairly constant. This reduces the complexity of the problem and removes the uncertainty of the generative model triggering no alarm.

The large amount of early false positive trajectories generally fall into two categories. Vessels at high speed with very high uncertainty as in Figure 7.3 or

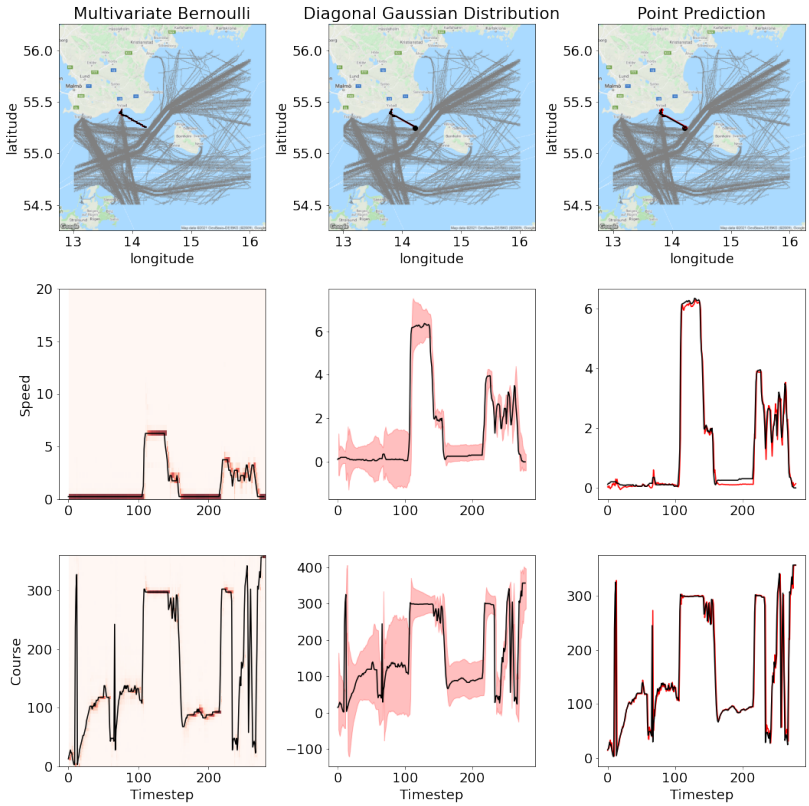


Figure 7.2 – Reconstruction of a false negative using the point prediction model. The reconstruction is shown in red and true values in black. For the diagonal Gaussian distribution the shaded area denote plus/minus 2 standard deviations.

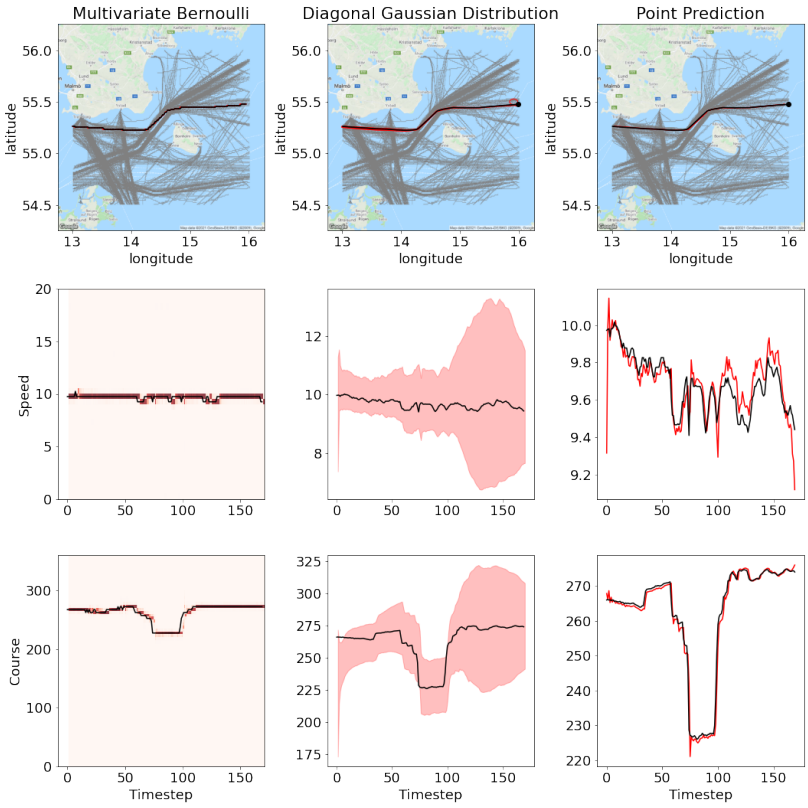


Figure 7.3 – Reconstruction of an early false positive using the model predicting the distribution parameters. The reconstruction is shown in red and true values in black. For the diagonal Gaussian distribution the shaded area denote plus/minus 2 standard deviations.

fishing vessels with frequent speed and course changes. The fishing vessels were also reconstructed with very high uncertainty in speed and course. The discrete multivariate Bernoulli and point prediction models also have higher reconstruction errors for the fishing vessels, and they are the first trajectories of normal behavior to be flagged as abnormal.

Most abnormal behaviour is defined by the kinematic behaviour and require explicit inclusion of these feature. The behaviour flagged as abnormal is overall very similar between the three learning objectives considered. The model trained for point predictions achieves the overall best abnormality detection performance. However, introduction of the uncertainty prediction by using the full generative diagonal Gaussian distribution may make the prediction more sensitive to frequent changes. This may help detect some abnormal trajectories, but at the cost of more false positives and a general sensitivity to noise. This sensitivity can be reduced by implementing a discretization of the inputs. However, this makes the models tied to the ROI and makes application to new regions or implementation on moving platforms impossible. Furthermore, the discretization may make irregular sampling impractical due to difficulty selecting a proper time resolution and lack of well defined upper limit on the time step.

7.4.2 Irregular Sampling

In this section, we implement the irregular sampling strategy outlined in section 7.3.2. We train GeoTrackNet and RAE models for point predictions and prediction of the distribution parameters. The models are then evaluated on the testset with a regular sampling of 2 and 10 minutes. Figure 7.4 shows the ROC curves for these models. The GeoTrackNet model is generally not affected by random irregular samples based on the AUC and TPR/FPR values. The model trained for the point predictions has the same AUC and same general shape of the ROC curve. The AUC of the model trained to predict the distribution parameters increases slightly. The large number of early false positives is generally removed, but the following detections are also pushed later. As a result, the TPR does not reach 90% until a FPR of 50%, which is higher than the model trained for point predictions.

Training using irregular sampled trajectories may act as a regularizer. Compared with the regularly sampled case, the diagonal Gaussian model reconstruct the speed of fast going vessel without incurring a large uncertainty as we noted previously. This has removed a large portion of false positives. However, it is also the cause of the later detection of the abnormal commercial traffic that mainly follow the shipping lanes.

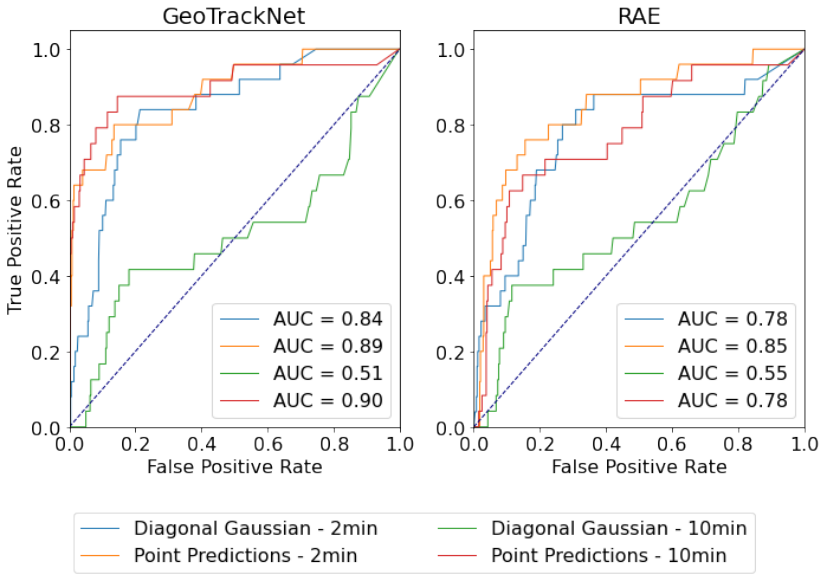


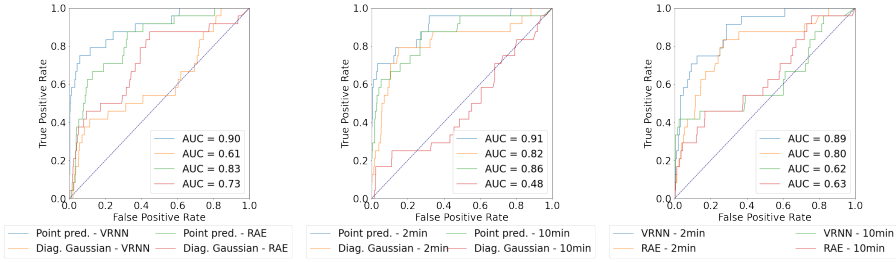
Figure 7.4 – ROC curves from varying the generative distribution and temporal resolution using GeoTrackNet (left) and RAE (right).

Increasing the sampling period to 10 minutes during the evaluation slightly increases the performance using the GeoTrackNet model for point prediction, but heavily decreases the performance of the diagonal Gaussian model. The increased sampling period accentuates the reconstruction error of minor deviations from normal steaming behaviour. This behaviour was typical for false negatives using point predictions when evaluated using a two min sampling period and thus increasing the sampling period causes them to be detected earlier. The false positives using 10 min sampling tend to be fishing ships with a varying course; however, we see the increased sampling period may help reduce detection of stationary vessels with highly varying course due to drift at low speeds. The reduced sampling rate may act as a low-pass filter, smoothing course changes and reducing the reconstruction error. This has the adverse effect for diagonal Gaussian model. In this case, the smoothing of the course changes results in Search-and-Rescue activities being detected later because the sampled course generally being within the two standard deviation interval of the reconstructed mean. On the contrary, we see high reconstruction loss for traffic following the shipping lanes and making course changes. Due to the time passed since the previous sample, the new course is not within the two standard deviation interval, and if multiple such course changes occur within a short time, the trajectory may be flagged as abnormal.

Using the RAE model, the AUC slightly decreases for all models. The trajectories from Search-and-Rescue activities are generally detected later, and the first detections made are false positives that fall into one of two types; fishing vessels that are very close to the shipping lanes or commercial traffic following the shipping lanes. Additionally, the detection of abnormal trajectories of commercial traffic that deviate from the shipping lanes is made earlier compared with the VRNN model. The same results hold for the RAE model evaluated using a 10 minute sampling period, but they are less pronounced than using a 2 min sampling period. This is also reflected in the AUC which decreases when increasing the sampling period.

Training using random irregular samples does not worsen the abnormality detection capability of the GeoTrackNet or RAE model when using point predictions. However, it negatively affects the reconstruction capabilities of the RAE models. RAE models trained using irregular samples are unable to correctly reconstruct the temporal aspect of trajectory. As a result, the reconstructions show the general shape of the trajectory, but the origin and end locations are not correct. Using the MSE loss (Eq. (7.15)) this does not affect the abnormality detection results. However, using the full probability loss (Eq. (7.6)) the abnormality detection results are worse due to systematic errors at the origin and end of trajectories with large spatial extent. Training using random irregular samples makes the models generalizable to different resample periods. These different resample periods can be sensitive to different types of behavior due to longer sampling periods acting as a smoothing low-pass filter. Therefore, depending on the type of behavior of interest, different sampling periods may be useful, and a model trained on random irregular samples can be used to evaluate all sampling periods.

7.4.3 Ensembles



(a) Different sampling periods. (b) Different models. (c) Different objective functions.

Figure 7.5 – Ensembles of different sampling periods, models, or objective functions.

In the previous section, we highlighted how the GeoTrackNet and RAE model flag different types of behaviour as abnormal and found that different resampling periods may also change the type of behaviour flagged. In this section we investigate the performance of ensembles of either different sampling periods, models, and/or objective functions. The ROC curves of ensembles of models with different sampling periods, model architectures, and learning objectives are shown in Figure 7.5.

We see combining detections with different sampling periods using the VRNN model for point predictions results in superior detection to each individual model. The AUC is the same, however, the ensemble model reaches a TPR of 90% before the individual models. As we discussed above, the behavior of detections made using a resampling period of two and ten minutes, respectively, differed, which is why we see a small boost in performance. Ensembles of the RAE model for point predictions has a lower AUC than the individual model sampled at a period of two minutes due to later detection of the first few abnormal trajectories. Ensembles of diagonal Gaussian models are much worse than individual models sampled at a period of two minutes. As discussed above, the increased sampling period caused the diagonal Gaussian models to not detect Search-and-Rescue activities, which is also reflected in the poor performance of ensembles with these members.

Ensembles with different models trained point predictions with a resampling period of two minutes are better than each individual model and are in fact the best performing detector as measured by AUC. This is due to the difference in behavior detected by the ensemble members, making the ensemble good at detecting both Search-and-Rescue activities using GeoTrackNet and deviations from the shipping lanes using the RAE model. With a sampling time of ten minutes the ensemble performs a little worse than the GeoTrackNet model by itself. This is mainly due to the later detection of abnormal trajectories due to the RAE model, which is a worse predictor. As discussed above, the two models with a sampling period of ten minutes flag similar behavior. Therefore, they are not able to complement each other as in the case for sampling period of two minutes Ensembles of the diagonal Gaussian models using a sampling period of two minutes is a little worse than the individual GeoTrackNet model mainly due to the very late detection of the last three abnormal trajectories. These three trajectories are detected very late in the RAE model, which dominate the ensemble.

The ensemble of GeoTrackNet models trained for different objectives with a resampling period of 2 minutes has the same AUC as the best performing ensemble member. The ensemble struggles with early false detections. However, it already reaches a TPR of 90% at a FOR of 30% and detects all abnormal trajectories with an FPR of 60% which is the best of any ensemble or individual

model.

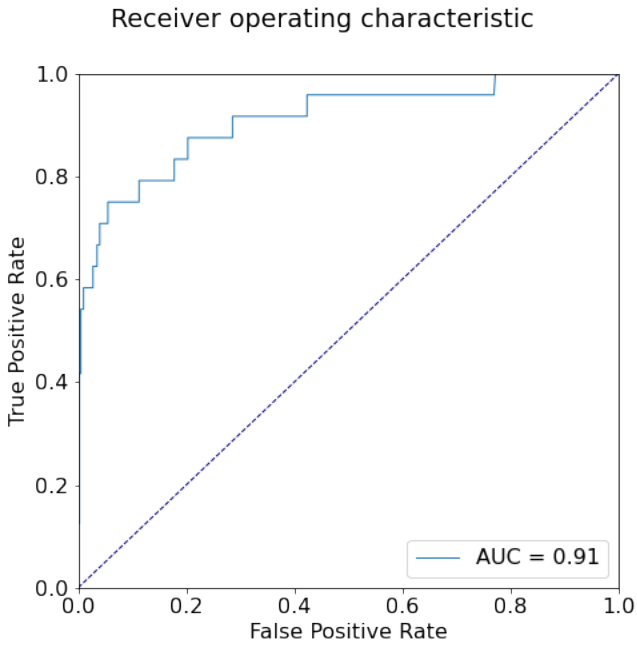


Figure 7.6 – ROC curve of an ensemble with members hand picked due to type of behaviour flagged as abnormal.

Based on the discussion above, we identify two ensembles which noticeably improved the performance over the three members individually (GeoTrackNet for point predictions and a sampling of two minutes was a common member in both ensembles). Common for these two ensembles were that we had identified they flag different types of behaviour. Thus, the ensembles are capable of flagging both kinds of behaviour and the ensemble may help suppress some false positives. Figure 7.6 show the ROC curve of an ensembles with these three members. This ensemble is the best performing detector of abnormal maritime trajectories compared with all other ensembles and models considered.

7.4.4 Transfer Learning

In this section, we evaluate the abnormality detection performance of models trained on a different ROI. We consider GeoTrackNet and RAE models trained for point prediction and distribution parameters evaluated using a resampling period of two minutes. The models will be trained on the waters around the

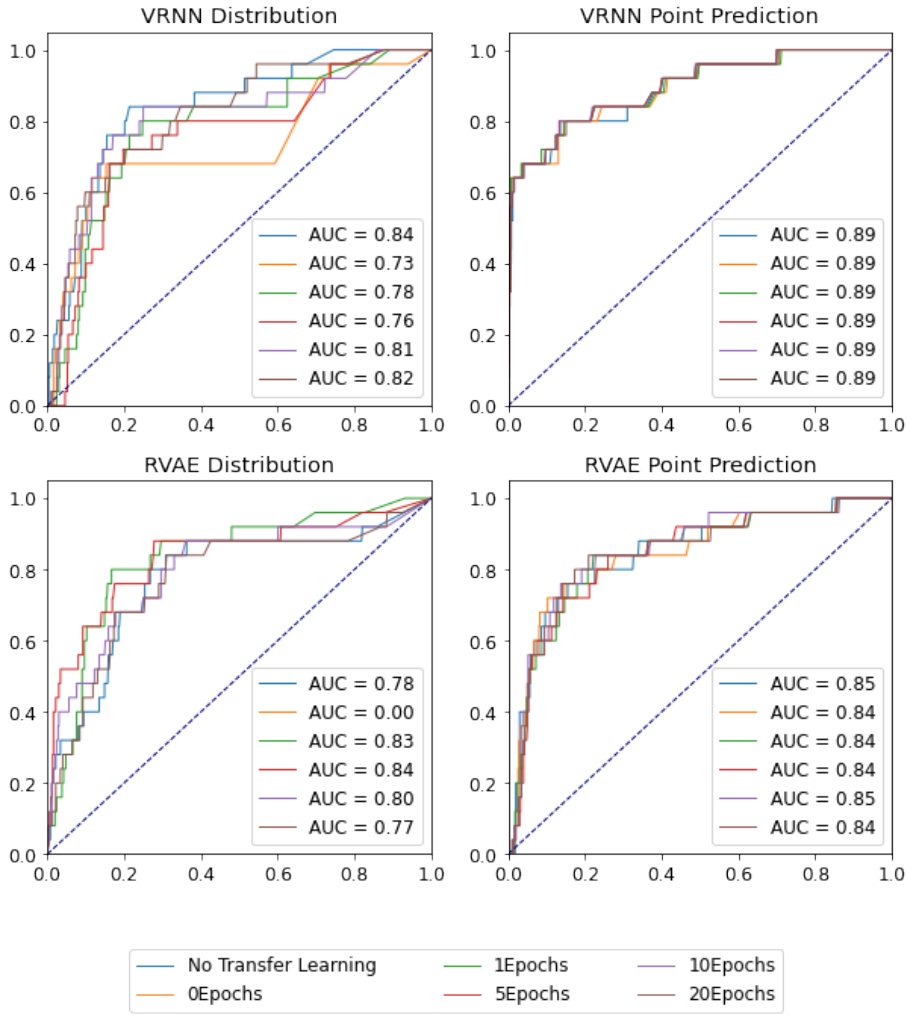


Figure 7.7 – ROC curves of models trained using trajectories from Anholt and fine-tuned using trajectories from Bornholm.

Danish island of Anholt bounded by (56^oN, 10.7^oE) to (57.5^oN, 13^oE) using the same collection period and preprocessing. This ROI was selected because it is of similar size and expected behaviour as the Bornholm area, however, the directions of main shipping lanes around Anholt are in a north/south direction opposite the primary east/west direction around Bornholm.

Figure 7.7 shows the ROC curves for all models with varying amounts of fine-tuning on the Bornholm area. The baseline (blue) is trained fully on data from the Bornholm area. We note here that the A-Contraio detection is grid based and computation of the local area distribution P^{C_i} requires evaluation of a training set from the target domain. For this reason we also use the values of the mean/std/min/max on the target domain to preprocess the data on the target domain.

The models trained for point predictions have the same abnormality detection performance as models trained on the target domain and require no fine-tuning on the target domain. There are no significant changes to the outlier factor and the reconstructions are only slightly affected by the transfer to another ROI. Particular, the speed feature is reconstructed with a small error without any fine-tuning, However, it does not affect the abnormality detection. Additionally, the reconstructions are similar to models trained fully on the target domain after just a single epoch of fine-tuning. We see two possible explanations for this result. One explanation could be the two ROIs are not sufficiently different to produce poor reconstructions when weights are transferred between ROIs. Secondly, the suggested preprocessing to the interval $[-1, 1]$ causes inputs from different ROIs to be mapped to the same domain before the model. This in turn make the models capable of reconstructing trajectories sufficiently regardless of the target domain.

The abnormality detection performance of the GeoTrackNet model trained for prediction of distribution parameters is worse and requires at least 10 epochs of tine-tuning before the performance is comparable to models trained on the target domain. There is a general increase of the uncertainty levels, particularly on the reconstruction of the course. Even after 20 epochs of fine-tuning the reconstruction of the course is nearly constant for all trajectories and 95% confidence level cover all possible values of the course. The reason for the nearly constant course reconstruction is unknown and is also seen for vessels with a primarily north/south bound direction. As a consequence of the constant course reconstruction, the relative weight of the course when determining abnormalities is reduced since all reconstructions have comparable reconstruction errors. Therefore, stationary trajectories at drift with frequent course changes are removed as false positives and the SAR vessels has generally a lower outlier factor, but are still flagged as abnormal. Furthermore, trajectories steaming at a constant course is now flagged as abnormal due to the relative higher weight on the

position error, particularly outside the shipping lanes.

The RAE model for prediction of distribution parameters perform better than the target domain model after a few epochs of fine-tuning and then decrease in performance with additional fine-tuning. The model may not be applied without any fine-tuning as the reconstruction quickly diverge and result in log-probabilities of negative infinity. The increase in detection performance is due to a poor reconstruction of the position of most SAR vessels. Since most SAR vessels originate in port the reconstruction of the position stay stationary over the port resulting in large positional reconstruction errors. With more fine-tuning the reconstruction of the position becomes better and detection of the SAR vessels are conducted later similar to the model trained on the target domain.

We find it overall interesting that the models for point predictions generally does not require any fine-tuning. Whether, this is a result of the suggested preprocessing or the similarity of the ROIs remains unclear. To further test the reason, the reason requires data from an area significantly different from Danish waters. Using the models for prediction of distribution parameters requires a large degree of fine-tuning and may not save much training time in practice. During and after the fine-tuning process, abnormality detection results may be significantly different. However, whether this is for the better or worse depends on the behaviour of interest and source/target domains.

7.5 Conclusion

In this work, we suggest the use of ensembles to improve the detection of abnormal maritime traffic in the wake of a fatal collision accident. We find that different model architectures and sampling periods may flag different types of behavior as abnormal. Additionally, ensembles consisting of members that flag different behavior can improve the detection rate and reduce the number of false positives. In future work, we wish to design and add ensembles members that alert to different kinds of behaviour. For instance, an ensemble member accounting for ship-to-ship interactions might be able to detect possible collision before they happen or serious maritime threats such as piracy or smuggling. However, detection of most other types of abnormal behaviour requires development of data sets labeled for each type of behaviour.

We investigate the applicability of transfer learning to reduce training times of models on new maritime regions and find that models trained for point prediction require no fine-tuning to achieve the same abnormality detection per-

formance. However, it still remains unclear whether this is a result of the suggested pre-processing or the relatively close similarity of the ROIs. Using models trained for prediction of distribution parameters the type of behaviour flagged as abnormal may change significantly, and thus may not be very applicable for transfer learning.

Chapter 8

Towards latent representation interpretability for maritime anomaly detections

Kristoffer Vinther Olesen[a] · Anders Nymark Christensen[a] · Line Katrine Harder Clemmensen[a]

[a] Technical University of Denmark, DTU Compute, Kgs. Lyngby, Denmark

Publication Status: Paper is unpublished

Abstract: The increase in worldwide maritime traffic makes maritime safety and security ever increasing key issues. For surveillance operators, the interpretability of the predictions of any system are of absolute importance in order to adequately gauge security risks. For these reasons, we argue that recently proposed anomaly detection models are impractical for operational use. We show that the generating models of state-of-the-art sequential AutoEncoders

disregard the stochastic latent variables. We propose to induce physical interpretable information in the stochastic latent space through the use of Kullback-Leibler (KL) annealing as well as sparsity and temporal invariance enforcing losses. We test our proposed losses on maritime trajectories extracted from AIS data from Danish waters around the island of Bornholm. We find that the proposed losses increase the interpretability of latent vectors in terms of geographical position. However, due to overlap in the latent space, we doubt the proposed self-supervised methods can learn a sufficiently disentangled latent space to generate realistic trajectories of modified latent variables.

8.1 Introduction

According to the International Maritime Organization, international shipping is currently carrying 80% of the world's trade. It is the most efficient and cost-effective form of long distance transportation [1]. These factors make maritime safety and security key issues. For this purpose, real-time delivery of maritime situation maps is a necessity for activities like search-and-rescue, smuggling detection, piracy detection etc. Current operational systems rely strongly on human experts, and surveillance operators monitor and predict emerging critical situations within a large sea area. Meanwhile, data are collected from a wide range of sensors, like radar, sonar, and Automatic Identification System (AIS) as well as intelligence reports and weather data. As the number of data sources increases, it becomes increasingly difficult for operators to process the amount of information due to factors like cognitive overload, time pressure, fatigue, and uncertainty due to the complex and heterogeneous nature of the data. In order to provide support for the operators, methods and systems capable of performing anomaly detection have been an active research area; Pallotta et al. [2013b], Nguyen et al. [2021], Ding et al. [2020], Protopapadakis et al. [2017].

A key requirement of any surveillance system is interpretability in order for operators to analyse the results, how they are derived, and decide on which actions to take next. This is further underlined by the inclusion of many external data sources, which operators might need to evaluate. Furthermore, the role of the external factors may be better understood using a stochastic latent space. In the ideal case, the random latent space would help explain the role these external factors play when generating maritime trajectories whose kinematics are described by a recurrence model.

In this paper we propose loss adjustments to GeoTrackNet, a state-of-the-art, anomaly detection algorithm based on sequential Variational AutoEncoders,

[1] About IMO

and compare the abilities to encode physical interpretable information in the stochastic latent space. Our contributions include:

- Evidence that the GeoTrackNet model does not fully utilize the stochastic latent layer.
- A comparison of loss functions in terms of latent space interpretability, and a demonstration of the limitation of the current approaches.

The paper is organised as follows: In section 8.2 we provide an overview of related work. Section 8.3 argues for the underused stochastic latent space in the GeoTrackNet model, in 8.4 we describe our suggested modifications to the loss function and our experimental results and conclude in Section 8.5.

8.2 Related Work

Maritime Anomaly Detection

Previous work within the domain of maritime trajectory analysis can be divided into *knowledge-based approaches* and *data-driven approaches*. *Knowledge based approaches*, Snidaro et al. [2015], Thomopoulos et al. [2019], form the basis of most operational anomaly detection systems currently in use due to a large degree of interpretability, human interaction, and capabilities for online detection. However, these models suffer from scaling issues when adding new data sources, and currently, no studies have focused on redefining the rule basis due to seasonality, geographical or data source related changes. Due to lack of labelled anomalies in public data sets *Data-driven approaches* have mainly focused on unsupervised methods. However, supervised methods such as: Support Vector Machines Sfyridis et al. [2013], self-organizing maps Venskus et al. [2019], Gaussian Mixture Models and Kernel Density Estimation Anneken et al. [2015] have been suggested using a combination of authors' own labelling and simulated outliers. Unsupervised methods first attempt to construct a normalcy model based on historical positional and kinematics data obtained from AIS messages. Outliers are then determined based on this model. The most popular approach for constructing a normalcy model has been using density based clustering; Pallotta et al. [2013b] Radon et al. [2015]. More recently, several deep learning based methods have been suggested. Zhao and Shi [2019b] suggest training a Long-Short-Term Memory (LSTM) network on trajectories clustered by DBSCAN Ester et al. [1996], Gao et al. [2018] use a bidirectional LSTM to predict future positions, and determine anomalies based on the prediction error, Protopapadakis et al. [2017] use stacked AutoEncoders in combination with density based clustering. Both Ding et al. [2020] and Nguyen et al. [2021] train a VRNN

Chung et al. [2015] in order to learn a probabilistic reconstruction of the trajectory. GeoTrackNet Nguyen et al. [2021] also suggested a preprocessing step converting the continuous kinematic data from AIS into four one-hot-encoded vectors and using the concatenated vector as inputs to the network.

Sequential Variational AutoEncoder

The VRNN can be considered a RNN which at each time step consist of a Variational AutoEncoder Kingma and Welling [2013] conditioned on the recurrence model. This means the prior distribution on the latent stochastic variable at time t, $\mathbf{z}_t$, is given by a Gaussian with parameters $\boldsymbol{\mu}_{0,t}$ and $\boldsymbol{\sigma}^2_{0,t}$, conditioned on the recurrent hidden states, $\mathbf{h}_{t-1}$.

$$\begin{aligned} &\mathbf{z}_t \sim N(\boldsymbol{\mu}_{0,t}, \mathrm{diag}(\boldsymbol{\sigma}^2_{0,t})), \\ &\text{where } [\boldsymbol{\mu}_{0,t}, \boldsymbol{\sigma}^2_{0,t}] = \phi^{prior}(\mathbf{h}_{t-1}) \end{aligned} \tag{8.1}$$

Similarly the generating distribution and the approximate posterior $q(\mathbf{z}_t|\mathbf{x}_t)$ will also be conditioned on $\mathbf{h}_{t-1}$. Using the four-hot-encoding suggested in GeoTrackNet, the generating distribution is given by a multivariate Bernoulli distribution with parameter $\boldsymbol{p}_{x,t}$

$$\begin{aligned} &\mathbf{x}_t|\mathbf{z}_t \sim B(\boldsymbol{p}_{x,t}), \\ &\text{where } B(\boldsymbol{p}_{x,t}) = \phi^{dec}(\phi^{\mathbf{z}}(\mathbf{z}_t), \mathbf{h}_{t-1}) \end{aligned} \tag{8.2}$$

and the approximate posterior is given by

$$\begin{aligned} &\mathbf{z}_t|\mathbf{x}_t \sim N(\boldsymbol{\mu}_{z,t}, \mathrm{diag}(\boldsymbol{\sigma}^2_{z,t})), \\ &\text{where } [\boldsymbol{\mu}_{z,t}, \boldsymbol{\sigma}^2_{z,t}] = \phi^{enc}(\phi^{\mathbf{x}}(\mathbf{x}_t), \mathbf{h}_{t-1}) \end{aligned} \tag{8.3}$$

At each time step the recurrence $\mathbf{h_t}$ is updated according to

$$\mathbf{h_t} = f_\theta(\phi^{\mathbf{x}}(\mathbf{x}_t), \phi^{\mathbf{z}}(\mathbf{z}_t), \mathbf{h_{t-1}}) \tag{8.4}$$

Learning is done by maximizing the timestep-wise Evidence Lower Bound (ELBO)

$$\begin{aligned} \mathcal{L}(\mathbf{x_{1:T}}) = \sum_{t=1}^{T} &\mathbb{E}_{q(\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_{t-1})} p(\mathbf{x}_t|\mathbf{z}_t, \mathbf{h}_{t-1}) \\ &- \beta KL\left[q(\mathbf{z}_t|\mathbf{x}_t, \mathbf{h}_{t-1})|p(\mathbf{z}_t, \mathbf{h}_{t-1})\right] \end{aligned} \tag{8.5}$$

It is a well documented issue that sequential VAEs may suffer from inactive latent variables Goyal et al. [2017], Bowman et al. [2015], Semeniuta et al. [2017]. Goyal et al. [2017] argues this because of strong local correlations at low level features. This causes the training objective to be insensitive to higher

level abstractions in the observations. Several solutions to this problem have been proposed; like reducing the capacity of the autoregressive decoder Bowman et al. [2015], Semeniuta et al. [2017], self-supervised learning objectives Zhu et al. [2020], Brito et al. [2020], auxiliary tasks Goyal et al. [2017], Figueroa and Rivera [2017], and sparsity enforcing penalty terms or priors Mathieu et al. [2018]. Furthermore, work in deep hierarchical Variational AutoEncoders Sønderby et al. [2016] shows annealing of the KL loss during the initial phase of training is essential for training deep latent encodings.

8.3 Encoding in GeoTrackNet

In recent work by Nguyen et al. [2020], the GeoTrackNet Model is extended with a post-processing step taking into account weather and other external sources. This post-processing step serves the purpose of potentially filtering out statistically unusual, but due to external conditions, not suspicious trajectories. Such a processing step would require knowledge of why a trajectory is considered statistically unusual. For this purpose, it may be beneficial to traverse a disentangled random latent space and condition generation on various physical conditions in order to test different hypotheses. In order to gauge the latent encoding in the GeoTrackNet model, we conduct a few different experiments.

Throughout training, the approximate posterior and prior distributions are very similar and show very little variation between observations. This causes the reconstructions generated using the approximate posterior and prior distributions to be very similar. Thus, the reconstructions will primarily be driven by the kinematic encoding in the recurrence and show very little variation due to the external random effects that might effect the trajectory of maritime traffic. This means that GeoTrackNet defaults into a LSTM-network trained to minimize the 1-step prediction error. Training such a network we found only a very small difference in the total reconstruction errors again confirming the vast majority of the information in GeoTrackNet passes through the recurrence.

In figure 8.1, we plot the latent encoding of the test set along the first two principal components found by Principal Component Analysis. The two first principal components describe 99.3% of the variance between them. We see the vast majority of latent vectors follow a linear relation between the first two principal components. The remaining 495 points show a larger variation along the first component especially. Plotting these points in the physical input space, figure 8.2. We note that the outlying points corresponds to a small region in the western part of the region of interest (ROI). The speed of these are among the highest observed and their course show large variation. These observations further confirm that the random latent space is mostly unused in the generative model. However, we do see that information may be encoded in extreme cases.

Figure 8.1 – Decomposition of the latent space in GeoTrackNet using Principal Component Analysis. We see the vast majority of encodings follow a linear relation (blue). The remaining points (red) corresponds to updates in a specific geographical area with high speeds and a large variation in course, see figure 8.2

This leads us to hypothesize that encoding relevant physical information in the latent space is possible, and that this information may be useful in evaluating detected abnormalities.

8.4 Experiments and Results

Implementation details

We tested the models on AIS data from Danish waters. We limited the study to AIS data from the area around the island of Bornholm. The ROI is a rectangle bounded by $(54.5^oN, 13^oE)$ to $(56^oN, 16^oE)$. The maritime traffic in this area

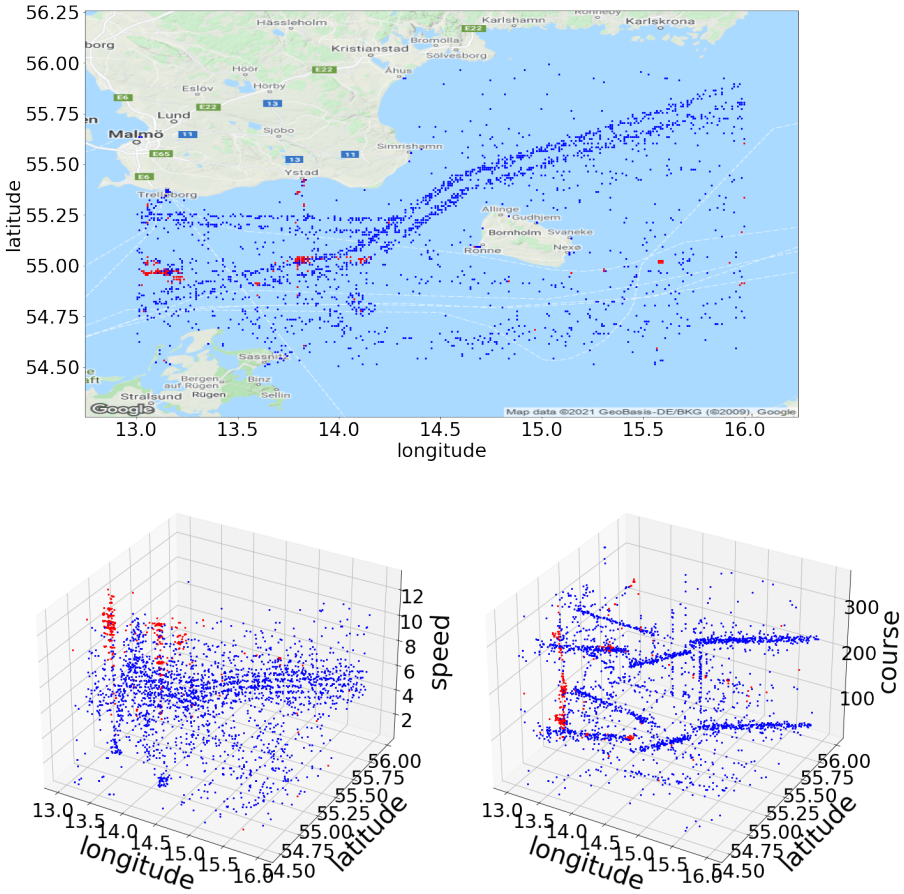


Figure 8.2 – Physical input space of AIS updates in the test set. Points showing a linear relation among the two first principal components of the stochastic latent space (blue) are spread through out the input space. Points showing a non-linear relation (red) are focused in two well defined geographical locations, and shows very high speeds (bottom left) and a large variation in course (bottom right).

resembles that used in Nguyen et al. [2021] and contains primary shipping lanes including forks and mergers as well as traffic to and from several ports. The data were collected from March 1^{st} to June 31^{st} 2021 and contains traffic from 8 different ship types. The speed was truncated to $15.5m/s$. If the time interval between two successive AIS messages exceeds 15 min, the track was split into 2 contiguous tracks. Tracks shorter than 10 minutes were discarded and tracks exceeding 12 hours were split into smaller tracks each between 10 minutes and

12 hours. All tracks are resampled every 120 seconds using linear interpolation. We then randomly set aside 20% of the generated tracks as a test set. For the four-hot-encoding we use a resolution of 0.01^o for the latitude and longitude, $0.5m/s$ for the speed, and 5^o for the course. This gives an input dimension of 553 for each timestep. For the functions $\phi^{\mathbf{x}}, \phi^{\mathbf{z}}, \phi^{\text{enc}}, \phi^{\text{prior}}, \phi^{\text{dec}}$ we use fully connected neural networks with one hidden layer of size 100. The recurrence f_θ is a LSTM with hidden size 100. The dimension of the posterior, $q(\mathbf{z}|\mathbf{x})$ and prior, $p(\mathbf{z})$ distributions is also set to 100. All inputs are normalized using the mean and the mean vector is used as a bias for the generative distribution. We use Adam optimizer with a learning rate of 0.001 which is decreased by 30% after 2, 5, and 10 epochs.

Improving latent information

We attempt to improve information in the stochastic latent space by linearly annealing the weight β of the KL loss in (8.5) from 0 to 1 during the first 10 epochs of training. This provides the posterior with additional flexibility in the early stages of training such that it is not overpowered by the recurrent part of the VRNN.

In order to improve disentanglement and ensure we encode physical interpretable information, we also implement an ElasticNet (EN) loss on the weights of the encoder, w_{enc}, and prior networks w_{prior}.

$$\mathcal{L}_{EN} = \alpha(|w_{\text{enc}}|_2 + |w_{\text{prior}}|_2) + \lambda(|w_{\text{enc}}|_1 + |w_{\text{prior}}|_1) \tag{8.6}$$

In order to make sure the stochastic latent space does not encode any dynamic recurrent information we suggest a self-supervised static consistency (SCC) loss inspired by Zhu et al. [2020]. For the sequence of latent vectors corresponding to input i, $z^i_{1:T}$, we compute the mean of the pairwise distance of all latent vectors

$$l_{\text{pos}} = \frac{\sum_{k=1}^{T-1}\sum_{l=k+1}^{T} D(z^i_k, z^i_l)}{T(T-1)/2} \tag{8.7}$$

This serves to drive the latent encodings at different time steps towards each other such that they may describe static information related to the entire trajectory. Direct minimization of (8.7) may result in trivial solutions that does not utilize the latent space as seen in section 8.3. Therefore, we also compute the pairwise distance of all latent vectors of other trajectories. We implement a batch-all strategy computing the average of trajectories in the current training batch i within the margin, M.

$$l_{\text{neg}} = \frac{\sum_{j=1, j\neq i}^{m} \frac{\sum_{k=1}^{T_i}\sum_{l=1}^{T_j} D(z^i_k, z^j_l)}{T_i T_j}}{m} \tag{8.8}$$

We use the euclidean norm as the distance measure $D(\cdot, \cdot)$. We wish to minimize the triplet loss using the margin $M = 1$ and weighting η.

$$\mathcal{L}_{SCC} = \eta \max\left(l_{\text{pos}} - l_{\text{neg}} + M, 0\right) \tag{8.9}$$

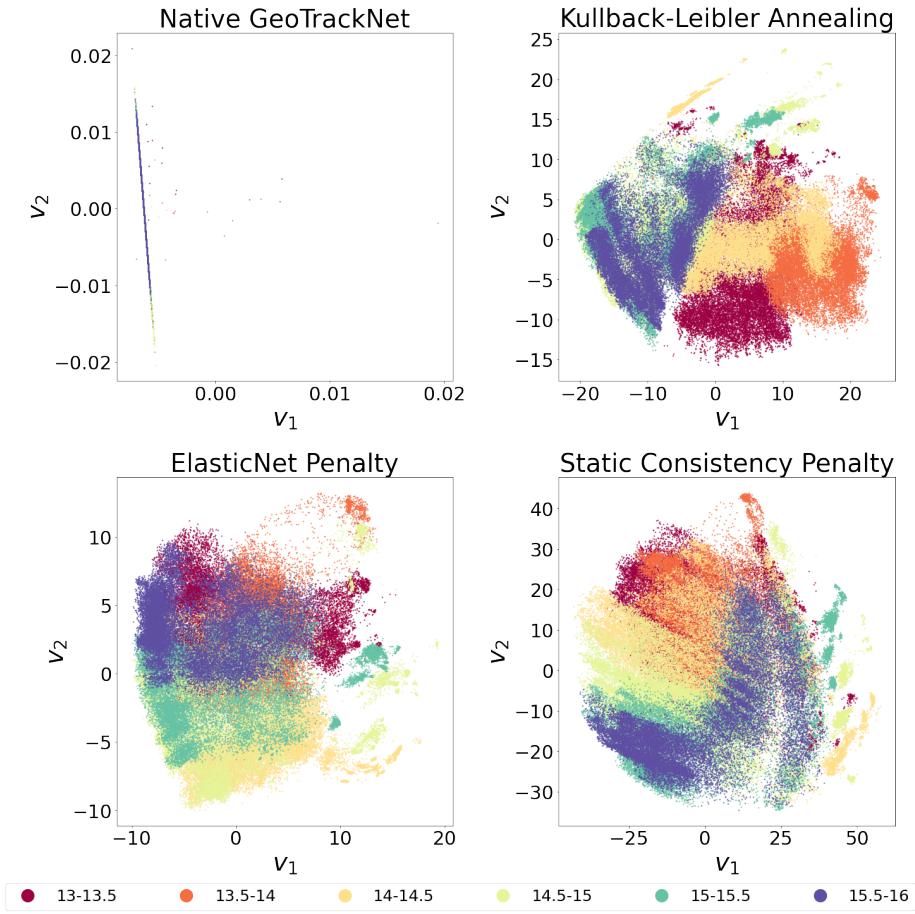


Figure 8.3 – PCA Decomposition of the latent encoding in topleft: native GeoTrackNet, topright: Using KL-annealing, bottomleft: Using Elstic-Net loss (8.6), bottomright: Using SCC-loss (8.9). Colors indicate the longitude of the corresponding input.

In figure 8.3 we plot the latent variables of the testset along the first two principal components. In all experiments, the first two principal components explain more than 98% of the variance. Thus, it does not seem the proposed extensions induce a higher dimensional encoding into the latent space. However, we note

a much larger variation along the first two principal components. The latent space is defined by a dense center with a couple of less dense outlying regions. The colouring indicates different intervals in the longitude of the corresponding input. For KL-annealing, we note the longitude intervals are not ordered the latent space. Thus, as a trajectory traverse the ROI in an east-west fashion, the trajectory of latent vectors will likewise traverse the entire latent space. This is not an unexpected behaviour as the shipping lanes in our ROI mostly travel east-west, and thus, more variance is associated with longitude compared to latitude. Introducing the EN- and SCC-losses cause a larger overlap between longitude intervals in the latent space. This indicates similar values in the latent space are now associated with a wider range of longitude inputs. Thus, we may hypothesize that each area in the latent space is associated with a segment of shipping lanes covering a larger interval of longitudes. The SCC-loss cause a more smooth transition between areas of latent space since latent vectors on the border are being driven towards similar values.

To verify our association hypothesis between segments of shipping lanes and areas in the latent space, we train a KMeans clustering in the unreduced latent space. In figure 8.4 we plot the geographical position in the input space coloured by the cluster affinity in the latent space. We see that clusters in the latent space tend to be associated with segments of shipping lanes, as we hypothesised. However, we note a significant overlap in geographical position between clusters associated with segments of shipping lanes. Occasionally, the clusters are spread to very wide geographical areas outside of the busiest shipping lanes. We further notice that segments overlapping in geographical position also often overlap in their speed associations. This is contradictory to our expectancy that overlapping geographical clusters showcase different speed profiles. We suspect the negative case of the SCC-loss cause separation in the latent vectors, which further increase the overlap between sections of the latent space.

We note the scale of the principal components significantly increase when implementing the SCC-loss (8.9). This might indicate that the latent vectors are being separated rather than driven towards temporal invariant encodings. Ideally, we would like trajectories sailing in the same shipping lanes with similar speed to have approximately equal encodings. However, since we treat such a pair of trajectories as negative cases we might inadvertently end up separating similar trajectories in the latent space.

The large overlap in latent space may indicate that the network has not achieved the degree of disentanglement we desired. In a fully disentangled latent space, we would expect to see a connection between updates in a trajectory. We would expect to see clusters of trajectories or segments thereof showing similar behaviour. However, due to the large overlap in the latent space, we see little separation in the latent space between different types of trajectories or ship

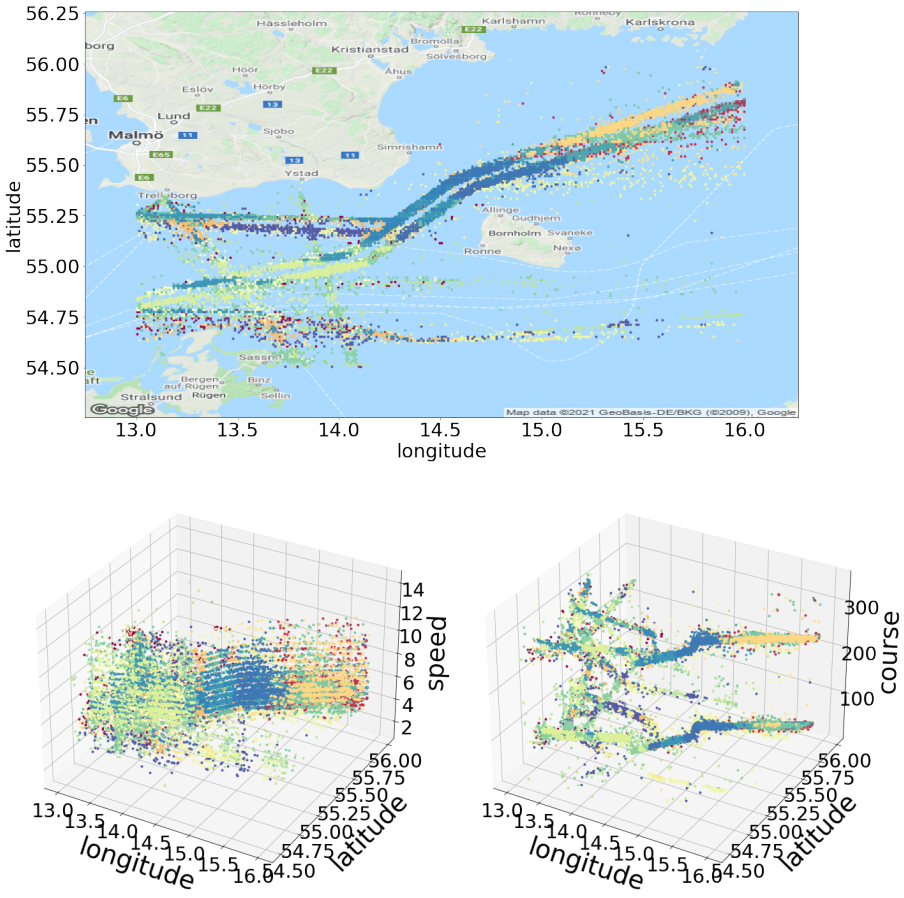


Figure 8.4 – Physical interpretation in input space of dense KMeans clusters of latent vectors using the SCC-loss (8.9). We note a significant overlap between clusters associated with shipping lane segments.

types. This indicates that the latent space may not be sufficient for anomaly detection. Sequences of latent vectors jump between different areas of the latent space as the trajectory moves through the ROI making it difficult to detect outliers based on latent space transitions. Thus, to do anomaly detection it is not enough to look at each update in isolation but we have to look at the trajectory as a whole. In GeoTrackNet this larger perspective is incorporated in the proposed A-Contrario Detector. We would argue that it would be possible to achieve a more interpretable prediction, if this perspective instead is incorporated into the normalcy model as a latent space incorporating knowledge from the entire trajectory.

Representation swapping

We investigate the possibility of swapping the latent vectors in the generating model between two trajectories in maritime shipping lanes. Suppose we have trajectories A and B, which both travel through the ROI in a westbound direction. Trajectory A sails through the westbound sea lane and trajectory B suddenly leaves the sea lane sailing due south before slowly returning to the appropriate shipping lane. The first column in figure 8.5 denotes generation using recurrent hidden states $\mathbf{h}_A$ and the latent vectors $\mathbf{z}_A$. Going to the right, the generation model is conditioned on $(\mathbf{h}_B, \mathbf{z}_B)$, $(\mathbf{h}_A, \mathbf{z}_B)$, and $(\mathbf{h}_B, \mathbf{z}_A)$. As expected from the inactive latent space, the swapping of latent vectors in the native GeoTrackNet does not affect the generated trajectory. For the KL annealing, EN-, and SCC-losses the generation does depend on the latent vectors, which causes the generation to break. The model is not capable to produce the wanted behaviour of transferring trajectories between input areas. We suspect this is because latent vectors $\mathbf{z}_t$ and recurrent hidden states $\mathbf{h}_{t-1}$ share dynamic information. This causes the mixing of potential contradicting local behavioral. We suspect representation swapping would be improved if the latent space encoded static knowledge from the entire trajectory.

8.5 Conclusion

In this work, we have studied the stochastic latent encodings of the VRNN based GeoTrackNet model for maritime anomaly detection. We find the generative model ignores the stochastic latent space making the model impractical for trajectory generation potentially conditioned on weather parameters or other external factors. We show the latent space can be activated through the use of KL annealing and that the latent space may encode noisy physical interpretable information using penalties enforcing sparsity and temporal invariance. However, these losses do not achieve a fully disentangled latent space. We suspect this is because the network is trained on each update in isolation and not conditioned on information related to complete trajectories. In future work, we

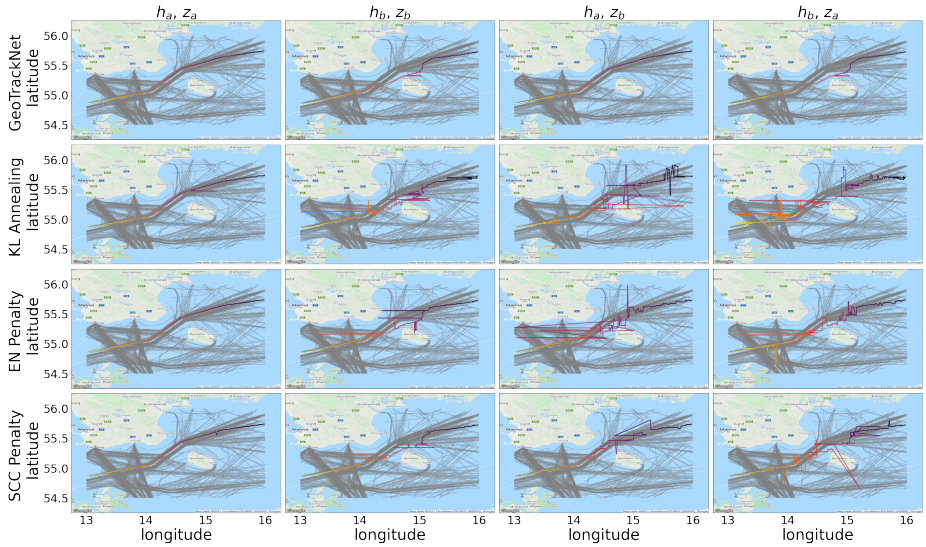


Figure 8.5 – Example of swapping of latent vectors in the generation model. For trajectories A and B the generating model is conditioned on (going left to right) $(\mathbf{h}_A, \mathbf{z}_A)$, $(\mathbf{h}_B, \mathbf{z}_B)$, $(\mathbf{h}_A, \mathbf{z}_B)$, and $(\mathbf{h}_B, \mathbf{z}_A)$.

wish to explore architectures which encoded static knowledge from the entire trajectory in the latent space.

Chapter 9

A two-step clustering method for maritime behaviour identification

Kristoffer Vinther Olesen[a] · Ahcène Boubekki[b] · Michael Kampffmeyer[b] · Robert Jenssen[b] · Anders Nymark Christensen[a] · Sune Hørlück[c] · Line Katrine Harder Clemmensen[a]

[a] Technical University of Denmark, DTU Compute, Kgs. Lyngby, Denmark
[b] UiT The Arctic University of Norway, Machine Learning Group, Tromsø, Norway
[c] Terma A/S, Lystrup, Denmark

Publication Status: Paper is submitted and under peer-review for publication in *IEEE Transactions on Intelligent Transportation Systems*

Abstract: The analysis of maritime traffic patterns for safety and security purposes is increasing in value. Expected maritime traffic patterns depend on several factors, such as position, kinematic behaviour, ship type, etc., which

complicates the analysis. It is beneficial for traffic pattern extraction methods to be able to disentangle these factors, whereof positional and kinematic behaviour are more complex and need to be derived from data. We propose a two-step clustering method to extract clusters that describe local kinematic behaviour. We design two similarity measures meant to capture positional and kinematic similarity, respectively. We separate trajectories into positional clusters and split them within each such cluster according to kinematic behaviour. We find that the proposed method results in clusters that have different kinematic behaviour while trajectories with the same behaviour but different local positions end up in the same cluster. We also extend the methodology for automatic anomaly detection and find that the performance is comparable to detections based on reconstruction using state-of-the-art deep neural networks.

9.1 Introduction

According to the International Maritime Organization, international shipping is currently responsible for 80% of the world's trade and is the most efficient and cost-effective form of long-distance transportation [IMO, 2021]. This makes maritime safety and security key issues.

During the last decade, research in maritime traffic patterns has been spurred along by easy access to large quantities of historical data from the Automatic Identification System (AIS). AIS is compulsory for all vessels exceeding a certain tonnage [Int]. Every day, AIS provides hundreds of millions of messages on a global scale [Mar] with static information such as ships' identifiers, size, as well as dynamic information of the Global Positioning System (GPS) coordinates, speed, course, etc. AIS forms the basis for extracting maritime trajectories used to model navigational characteristics and rules by analysing maritime traffic patterns. The most prevalent method for analysis of traffic pattern remains clustering either on the individual AIS updates [Pallotta et al., 2013a, Liu et al., 2015b] or using trajectory similarity measures [Zhao and Shi, 2019a, Yang et al., 2022b].

Real-life maritime laws and regulations are complex. Ship types such as cargo and tankers mainly follow well-defined shipping lanes with near-constant speeds, while other ship types, such as fishing vessels and sailing ships, have less constrained and more complex behaviour. The mixture of densely populated shipping lanes and more sparse regions with increased behavioural freedom complicates clustering. Most studies limit the data to focus on a single aspect of maritime traffic, for instance the behaviour of commercial merchant traffic [Wang et al., 2021] or port entry/exit ways [Liu et al., 2015a, Zhao and Shi,

2019a, Yang et al., 2022b]. However, in order for the discovered traffic patterns to be useful in a practical scenario, it is important that the underlying data set must be complete in terms of possible traffic. For this purpose, the training data has to contain trajectories from a Region of Interest (ROI) of a suitable size and all possible ship types. In ROIs of limited size, restrictions of ship types may be warranted, however, previous work on large ROI with complex behavioural patterns of various different ship types remains limited. In this work, we cluster traffic patterns in a large ROI with a number of different ship types with various priorities, tasks, and complex behavioural patterns.

Due to the simplifications discussed above, most previous research focuses heavily on clustering the position for the identification of shipping lanes [Pallotta et al., 2013a, Zhao and Shi, 2019a, Murray and Perera, 2021]. However, most ship types are generally not constrained to major shipping lanes. The exact route of these trajectories is less important than the local behaviour, i.e. speed and course changes, due to the additional freedom of movement. Thus, in order to better model the traffic outside the shipping lanes, we propose to look at the kinematic behaviour of the vessels using the speed and course. Naturally, the expected behaviour differs between locations and ship types. For instance, in specific locations, pilot boats steam between harbours and commercial traffic in the shipping lanes, but in other locations, this type of behaviour may be highly unexpected.

The discovery of maritime traffic patterns allows for the design of surveillance models for automatic maritime abnormality detection models, which may enhance maritime traffic safety. Anomaly detection of maritime behaviour is a difficult problem since there is no clear definition of what defines anomalous behaviour. Whether or not an event is abnormal is a combination of several factors: the location, the speed and course during the event, the ship type, the time of day/week/year, etc. Above, we mentioned that the behaviour of pilot boats might be considered abnormal outside specific locations or if being conducted by other ship types. Likewise, it may be expected that diving vessels perform frequent starts or stop in order to support divers in the water, but if this behaviour is seen from research vessels or happens near major shipping lanes, it may be considered abnormal unless permission has been obtained.

Recently, deep neural network models have been suggested for abnormality detection on large data sets of AIS trajectories covering multiple different ship types [Nguyen et al., 2021, Murray and Perera, 2021]. These models classify anomalies, as the trajectories for which they fail to predict the future or reconstruct. The main drawback here is a lack of any form of explainability as to why the networks fail to predict/reconstruct the correct trajectory. However, to be useful and accepted in an operational setting, abnormality detection models require a large degree of explainability to accommodate any scepticism

surveillance operators may have.

Our hypothesis is that separating position and kinematic information is key to describing different behavioural patterns in different local areas of a large ROI and detecting abnormalities. We propose a two-step clustering method that disentangles positional and kinematic behaviour. In the first step, trajectories are grouped based on their positional data. In the second step, the trajectories in each positional cluster are further clustered based on their speed and course. This final clustering is used to assess the typical behavioural patterns in the local area of a large ROI and to detect abnormalities. The latter can then be brought to the attention of an operator who can further assess the situation based on the expected behaviour of the given ship type in the area in question learned by the clustering.

We summarize our contributions as:

- We design positional and kinematic similarity measures that focus on different dimensions of maritime trajectories.
- We provide evidence that a multi-step clustering approach can help disentangle positional and kinematic information resulting in a better description of behavioural patterns in a large ROI.
- We provide an automated method for detecting abnormal maritime trajectories based on the designed similarity measures.
- We provide public access to data sets of preprocessed maritime trajectories in regions of Danish waters including one day with annotations for abnormal behaviour related to a search and rescue event.

The paper is organised as follows: In Section 9.2 we provide an overview of related work within trajectory clustering. In Section 9.3 we present our proposed two-step clustering. Section 9.4 shows the ability of our proposed method to disentangle maritime traffic patterns as well as detect real-life abnormal trajectories from a ship collision. Finally, we present our conclusion in Section 9.5.

9.2 Related Work

Clustering of maritime trajectories has been widely studied to explain typical traffic patterns and find abnormal trajectories. The type of behaviour discovered

by clustering spatio-temporal trajectories depends heavily on the chosen similarity measure. Laxhammar and Falkman [2015] suggest using the maximum Euclidean distance between each pair of coinciding points along two trajectories of the same length. The requirement of equal trajectory length can be relaxed by using either the Hausdorff distance, [Zhen et al., 2017, Yang et al., 2022b, Wang et al., 2021] or Dynamic Time Warping (DTW), [Klaas et al., Zhao and Shi, 2019a] to measure trajectory similarity. The Hausdorff distance is invariant to time flipping, meaning trajectories following the same route in opposite directions would not be distinguishable from one another. In addition, from the clusters reported in Wang et al. [2021] we notice that the Hausdorff distance may assign a large similarity to significantly different trajectories. If trajectories are spread over a large area, the Hausdorff distance and DTW may overestimate the distance between trajectories with similar behaviour. This makes clustering using these similarity measures prone to noise and may not prove useful over larger geographical areas containing many different ship types. Additionally, both of these methods have quadratic time complexity, and several works, therefore, suggest a compression using the Douglas-Peucker (DP) algorithm [Douglas and Peucker, 2011]. Klaas et al. propose a two-stage DP algorithm: first reducing the trajectory based on the speed time series and secondly based on the position. This two-stage approach is found to better retain periods of acceleration such as stops.

Several different clustering algorithms have been applied for clustering of trajectories. Methods such as K-means, [Klaas et al.] and K-medoids, [Zhen et al., 2017] have been suggested in collaboration with similarity measures. However, density-based clustering techniques have long been the predominant approach to data mining within maritime trajectory analysis. Pallotta et al. [2013b] proposed the widely used TREAD method to cluster trajectories into traffic routes, which can then be used for anomaly detection and trajectory prediction. TREAD is a point-based method that extracts coordinates of new entries, exits and stopping within the ROI. These points are clustered using DBSCAN [Ester et al., 1996] to form waypoints in which ships enter, exit or stop within the ROI. A route between waypoints is then formed whenever a certain number of transitions between them have been observed. Several works, [Yang et al., 2022b, Wang et al., 2021, Zhao and Shi, 2019a] combine the idea of a similarity measure and density-based clustering. First, trajectories are simplified using the DP algorithm. The similarities are then computed using the Hausdorff distance or DTW before being clustered by DBSCAN. Wang et al. [2021] considers a hierarchical search over the hyperparameters of DBSCAN, which allows for groups with different densities, and helps to find clusters in sparsely populated geographical regions. Murray and Perera [2021] cluster the latent encodings from a recurrent variational autoencoder (RVAE) trained for trajectory reconstruction using hierarchical DBSCAN and find clusters corresponding to the major shipping lanes. The clusters are then used to train neural networks for predicting

the future position.

The above-mentioned approaches have only considered the positional input, yielding clusters that mostly correspond to the primary shipping lanes. Zhen et al. [2017] introduce the difference of the average course in their similarity measure and Liu et al. [2015b] extends the DBSCAN clustering models to consider not only the geographical distance of the coordinates, but also the difference in speed and course. This allows them to distinguish between shipping lanes in opposite directions and find speed differences within the main shipping lanes. However, the work is limited to small geographical areas and a limited number of ship types.

Knowledge about maritime traffic patterns is useful for detecting abnormal activity, and several clustering methods have been extended with a detection step. Often this step includes knowledge about the kinematic behaviour. Pallotta and Jousselme [2015] proposes a two-stage anomaly detection scheme using the routes extracted by TREAD. First, a trajectory is associated with a route using only the positional part. Afterwards, kinematic outliers are found by comparing the speed and course to the average behaviour of the route. Liu et al. [2015b] propose to split clusters into smaller geographical regions and compute average kinematic values for each split. These values are then used to detect abnormalities [Liu et al., 2015a]. Zhao and Shi [2019b] use the initial clustering from Zhao and Shi [2019a] to train deep neural networks for trajectory prediction. Abnormalities are then detected based on the prediction error. Recently, Nguyen et al. [2021] has suggested a variational recurrent neural network (VRNN) for abnormality detection based on trajectory reconstructions. Nguyen et al. [2021] also suggests an A-Contrario detection methodology which is supposed to account for regional differences in the reconstruction accuracy. While results reported using VRNN look promising, our feedback from surveillance operators mentions the lack of explainability as a key limitation for operational use.

In this work, we introduce an alternative positional distance measure to efficiently capture trajectory similarities across a large, complex data set of maritime trajectories representative of real-life traffic. In addition, the proposed two-step clustering method is not limited to the identification of shipping lanes but focuses also on the diverse behavioural differences within the discovered positional clusters.

9.3 Proposed Method

To address the lack of focus on kinematic behaviour in previous clustering methods, we propose to separate the clustering of trajectories' positional evolution and kinematic time series using different similarity measures. Before presenting our algorithm in detail, we discuss candidate similarity measures for spatio-temporal trajectories.

Trajectories extracted from AIS data are multi-dimensional, including position (latitude, longitude) and kinematics (speed, course). To address the lack of focus on kinematic behaviour in previous clustering methods, we propose to cluster the trajectories' positions and their kinematics sequentially. Both dimensions carry different structures, which requires the clustering algorithm to employ different similarity measures. We will review first existing similarity measures for spatio-temporal data before introducing our two-step clustering algorithm, followed by a discussion on its application for anomaly detection.

9.3.1 Existing Similarity Measures

In this section, we assume that trajectories are regularly sampled without missing data and that data points lie in a euclidean space where d is a generic distance. Throughout the section, we consider two trajectories A and B of time duration T_A and T_B, both integers, at time points t and τ.

9.3.1.1 Hausdorff

The type of behaviour discovered by clustering spatio-temporal trajectories depends heavily on the chosen similarity measure. As discussed above, the Hausdorff distance and DTW are the most used trajectory similarity measures. The Hausdorff distance [Shamos and Preparata, 1985] between two trajectories corresponds to the maximum smallest distance realized by any pair of points each in one of the trajectories:

$$\text{Hausdorff}(A,B;d) = \max_{t\in[0,T_A-1]} \min_{\tau\in[0,T_B-1]} d\Big(A(t), B(\tau)\Big). \tag{9.1}$$

The computations require a comparison of all possible pairs of points resulting in quadratic time complexity. Furthermore, the Hausdorff distance disregards

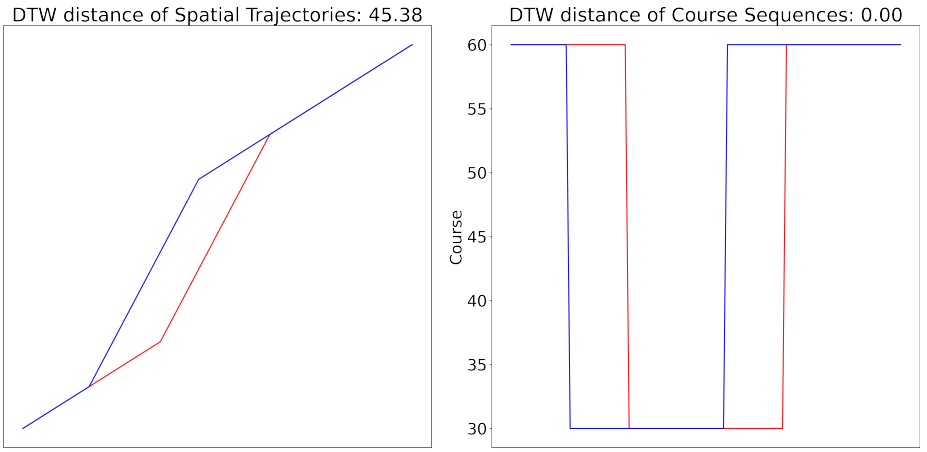


Figure 9.1 – Left: Two trajectories with the same origin and terminal point and similar behaviour through out the journey may obtain a large distance calculated by DTW. Right: The course sequences of the two trajectories may be warped perfectly onto each other and have zero distance between them calculated by DTW.

the time component. This means ships along parallel shipping lanes sailing in opposite directions are not distinguishable. Such a situation is studied in Wang et al. [2021].

9.3.1.2 Dynamic Time Warping

Dynamic Time Warping minimizes the pair-wise distance by re-indexing (alignment of) the data points in the trajectories according to certain rules. It can be defined as follows:

$$\mathrm{DTW}(A, B; d) = \min_{\pi \in \Pi(T_A, T_B)} \left(\sum_{(i,j)\in\pi} d\Big(A(t_i), B(t_j)\Big) \right) \tag{9.2}$$

where $\Pi(T_A, T_B)$ is the set of all possible alignments that are sequences pairs of indices $(i, j) \in [0, T_A - 1] \times [0, T_B - 1]$ satisfying three constraints: (1) the beginning and end of the time series must be matched together, respectively; (2) the sequence must be monotonically increasing in i and j; (3) all indices i and j must appear at least once. These ensure the sequences start and end together and that each point on either sequence is mapped onto at least one point of the other sequence without these mappings crossing in time.

Since DTW processes pairs of indices, it also has quadratic time complexity. The DTW alignment may overestimate the distance of trajectories with similar behaviour if this behaviour is spread over a large area. For example, consider two trajectories with the same starting point and sailing along the same direction as illustrated in Figure 9.1. At one point, trajectory A makes a 30 degree turn and continues in this direction, moving away from trajectory B. Later, trajectory B makes a similar 30 degree turn and continues parallel to trajectory A. Both vessels return to their initial course some time later, and the trajectories terminate at the same point. Since these two trajectories have the same origin and terminal location, and have similar behaviour throughout the journey, we would expect the distance between them to be very small. However, their distance, calculated by DTW on the sequence of geographical coordinates, may be significant. The re-indexing procedure of DTW aligns the course changes between the trajectories. However, due to the spatial nature of geographical coordinates DTW calculates the geographical distance between the location where the trajectories made the course changes. If we instead were to use the time series of the measured angles towards true north, the DTW distance would calculate the difference of the course values. Since these values are the same before and after the changes, the DTW distance between the two trajectories would be zero. By using the time series data, we remove the spatial dependence, and DTW can properly calculate the similarity of the course after aligning the changes. Therefore, DTW is a good candidate for a similarity measure for course and speed time series.

9.3.1.3 Average Haversine

The quadratic time complexity and the issues mentioned above make the Hausdorff distance and DTW unsuitable for measuring the similarity of many long and complex sequences of geographical coordinates. On the other hand, the average Haversine (AH) distance proposed in Nanni and Pedreschi [2006] to measure the positional evolution of AIS trajectories has a linear time complexity. It is defined as a continuous distance measure, but it can be approximated using the trapezoidal rule and assuming a regular sampling :

$$\mathrm{AH}(A,B;d_H)=\sum_{t=0}^{T-1}\frac{d_{\mathrm{H}}\big(A(t),B(t)\big)+d_{\mathrm{H}}\big(A(t+1),B(t+1)\big)}{2T}, \tag{9.3}$$

where $T=\min(T_A,T_B)$ and d_{H} is the Haversine distance. This similarity measure computes the geographical distance between trajectory points one by one in a linear fashion until the length of the shortest trajectory is reached. This means the measure puts an increased weight on the beginning of the trajectories. Thus, we expect the measure to be able to separate trajectories based on their

starting location. This is ideal in a real-time operational setting when observing new trajectories, as even short trajectories can very quickly be classified into a subset of historical trajectories with similar behavior.

9.3.2 Two Step Clustering

The rationale is first to make a rough separation of trajectories based on the global positional behaviour through the ROI, followed by a more fine-grained separation based on local kinematic behaviour. When used for anomaly detection, the first clustering step should not flag too many trajectories as outliers since this should be the purpose of the second step.

Our clustering algorithm consists of two steps:

1. Cluster all the trajectories based on their positional dimensions (latitude and longitude) using the average Haversine distance.
2. Cluster the trajectories within each positional cluster based on their kinematic dimensions (speed and course) using a custom similarity measure D_{kin}.

In both steps, we rely on hierarchical clustering with average linkage. The similarity measure D_{kin} is based on DTW:

$$\begin{aligned} D_{\text{kin}}(\tau_A, \tau_B) = \text{DTW}(s_A, s_B; d_{\text{speed}}) \\ + \text{DTW}(c_A, c_B; d_{\text{course}}), \end{aligned} \tag{9.4}$$

where s_A, s_B and c_A, c_B are, respectively, the speed and course sequences of trajectories A and B; d_{speed} is the standardized Euclidean distance and d_{course} the normalized angular difference in radians:

$$d_{\text{speed}}(x, y) = \frac{|x - y|}{\sqrt{V}}, \tag{9.5}$$

$$d_{\text{course}}(x, y) = \frac{1}{\pi} \cdot \begin{cases} |x - y| & \text{if } |x-y| \leq \pi \\ \pi - (|x-y| \ mod \ \pi) & \text{otherwise} \end{cases} \tag{9.6}$$

where V is the variance computed empirically from the speed time series s_A and s_B.

In practice, the distance thresholds for both hierarchical clusterings are set using Kneedles algorithm [Satopaa et al., 2011]. Also, in order to reduce the

time complexity of DTW, we compress the trajectories using the two-stage DP algorithm suggested in Klaas et al. and use the speed and course time series of the compressed trajectories. Note that although DTW-based similarity matrices are computed for each cluster, these involve much fewer trajectories than the whole dataset. The quadratic time complexity of DTW is thus less worrisome here.

9.3.3 Anomaly Detection

For anomaly detection, we exchange the kinematic clustering step with an outlier detection using our proposed kinematic distance measure, Eq. (9.4), as a basis for Local Outlier Factor (LOF) [Breunig et al., 2000] to detect trajectories with abnormal speed and course sequences.

The proposed method may also be used for anomaly detection on a new unseen trajectory. First, the trajectory is assigned to a positional cluster using a KNN classifier with $k = 3$ (hierarchical clustering cannot handle unseen data). Then, we compute the outlier scores using LOF within the assigned positional cluster.

LOF compares the local neighbourhood density of a point to that of its K-Nearest Neighbors (KNN). If the density of a point is significantly lower than its neighbours, the point is flagged as an outlier. In our experiments, we use $k = 5$ nearest neighbors. In practice, we did not see large changes in the number of outliers detected when varying that number. Nevertheless, we recommend a low value in order to capture information only from the local neighbourhood. LOF also has a hyperparameter related to the expected percentage of outliers. Since we only expect a small number of outliers, we recommend again to use small values for the hyperparameter. See Section 9.4.7 for an ablation study.

9.4 Experimental Results

9.4.1 Data Sets

For this work, we built two datasets of AIS data from Danish waters covering large ROIs and containing various types of vessels with different priorities and expected behavioural patterns. Both datasets are available for public use to facilitate reproducibility and to give researchers the ability to evaluate their proposed models on a complex data set representative of real-world setting [Olesen

et al., 2022b][1]. Complete AIS data from all Danish waters are available publicly [Soefartsstyrelsen], however, minor differences between the two sources may occur.

The first dataset covers a rectangle ROI over the Sjælland island bounded by $(54.4^oN, 10.5^oE)$ to $(56.4^oN, 13.5^oE)$. The data were collected during November 2021 and contain 18738 of trajectories from 11 different types of ships ranging from commercial cargo and tanker ships to private sailing and fishing boats. The second dataset covers a rectangle ROI over the Bornholm island bounded by $(54.5^oN, 13^oE)$ to $(56^oN, 16^oE)$. The data were collected during December 2021 and contain 12591 of trajectories from 8 different ship types. The speed was limited to $20m/s$, and updates with higher speeds were discarded. For both datasets, if the time interval between two successive AIS messages exceeds 15 min, the track was split into two contiguous tracks. Tracks shorter than 10 minutes were discarded and tracks exceeding 12 hours were split into smaller tracks, each between 10 minutes and 12 hours. All tracks are resampled every 120 seconds using linear interpolation.

We used the Sjælland data to evaluate the proposed two-step clustering algorithm. The entire data set was used for training. We test the proposed anomaly detection algorithm on the Bornholm data set. Data from December 13^{th}, 2021 was withheld as a test set, the rest serving as training set. On that day, a collision accident between two ships occurred, causing several abnormal trajectories. Trajectories from this day have been manually labeled, resulting in 25 abnormalities out of 521 trajectories. In addition to the colliding vessels, the abnormal trajectories correspond to commercial traffic, which had to deviate from the planned course, Search-and-Rescue and law enforcement vessels responding to the accident, and any other vessel taking part in the search of two missing sailors.

9.4.2 Experiments

We test our proposed distance measures and clustering method for both unsupervised clustering and abnormality detection. We evaluate clusterings using the Silhouette [Rousseeuw, 1987] score. In addition to our similarity measures, we include as baselines the Hausdorff and DTW distance using the Haversine distance as suggested in Yang et al. [2022b], Zhao and Shi [2019a]. In terms of clustering algorithm, we compare hierarchical clustering and DBSCAN. The linkage distance threshold for hierarchical clustering is decided using Kneedles algorithm to select the number of clusters. The hyperparameters of the DB-

[1]Datasets available at: `https://figshare.com/s/0012239be1c55e988a32`

Table 9.1 – Positional clustering performance in terms of silhouette score for various combinations of distance and clustering algorithms along with hyperparameters and characteristics of the clusterings. Our model corresponds to the last line.

Distance Measure	Clustering Algorithm	Eps-Threshold	MinSamples	# Clusters	% Outliers	Silhouette Score
Hausdorff	Hierarchical	9000	-	1515	-	0.535
Hausdorff	DBSCAN	12504	242	15	45.7	0.127
Hausdorff	DBSCAN	12504	25	53	10.7	0.376
Hausdorff	DBSCAN	27000	242	7	12.6	0.265
DTW	Hierarchical	140000	-	2862	-	0.349
DTW	DBSCAN	60941	91	7	68.8	−0.454
Avg. Haversine	DBSCAN	1.07	261	14	57.8	−0.033
Avg. Haversine	Hierarchical	10	-	52	-	**0.651**

SCAN are tuned by creating candidate lists of minimum distances and samples as suggested in Yang et al. [2022b]. The optimal value from these candidates is then determined using Kneedles algorithm [Satopaa et al., 2011]. For hierarchical clustering, we use Kneedles algorithm. We measure the performance of the proposed inductive clustering and anomaly detection using the area under the receiver operating characteristic (AUC). As baselines, we use the state-of-the-art VRNN [Nguyen et al., 2021] and RVAE [Murray and Perera, 2021] deep learning architectures.

9.4.3 Step 1: Positional Clustering

We begin by evaluating the first step of our clustering algorithm using the positional information of the Sjælland dataset. We consider different combinations of distance measures (Hausdorff, DTW based on Haversine distance and average Haversine distance computed using Eq. (9.3)) and clustering algorithm (hierarchical, DBSCAN). Our model combines the average Haversine distance with hierarchical clustering. We provide quantitative, qualitative, and runtime analyses.

9.4.3.1 Quantitative Analysis

In Table 9.1, we report clustering performance of each combination of distance measure and algorithm in terms of silhouette score (the larger, the better). We include hyperparameters selected using Kneedles algorithm, the number of clusters found and, when DBSCAN is used, the percentage of outliers. We observe that our combination of hierarchical clustering with the average Haversine distance achieves the best clustering.

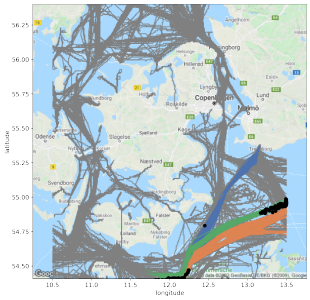
The silhouette scores show that DBSCAN generally performs worse than hierarchical clustering. One explanation could be the large number of trajectories flagged as outliers by DBSCAN. For all three distance measures, DBSCAN considers at least 45% of the data as outliers. This is too many false positives for an automated system to be useful. Despite the better silhouette scores, hierarchical clustering with the Hausdorff and DTW distances suffers from a similar phenomenon. Indeed, both combinations produce the largest number of clusters. Most of these clusters contain very few trajectories and, thus, serve a similar purpose as the outliers in DBSCAN.

On the other hand, our proposed combination finds a reasonable number of 52 clusters. The average number of trajectories in a cluster is 360 with a standard deviation of 380, meaning most clusters have roughly 100 trajectories assigned to them. These are reasonable numbers of trajectories per cluster to allow for further analysis in the second clustering step of the kinematic parts using DTW. Note that these numbers are comparable with the size of the full data sets used in most other works such as Yang et al. [2022b], Zhao and Shi [2019a].

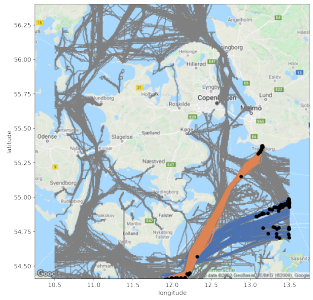
9.4.3.2 Qualitative Analysis

Clustering using DTW or the Hausdorff distance results in clusters corresponding to the well-defined shipping lanes as seen in Figure 9.2(a)-(b). However, they fail to cluster two types of trajectories: small groups of trajectories in less populated maritime routes, and trajectories which mostly follow the shipping lanes but make minor deviations from the other cluster members, see Figure 9.2(d). These trajectories are marked as outliers in DBSCAN or as single observation clusters in hierarchical clustering. Fine-tuning the hyperparameters may reduce the number of outliers by slowly admitting trajectories with similar positions into the clusters but with the risk of joining clusters of different shipping lanes as shown in Figure 9.2(c). This indicates that real, unfiltered trajectory recordings from a diversely populated ROI have too much randomness for these measures to find well-separated clusters using only the latitude and longitude without removing the majority of the data as outliers.

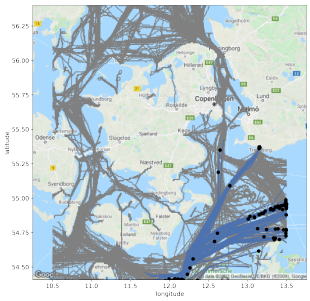
In Figure 9.2(e)-(f), we plot the 4 most populated clusters in the ROI around Sjælland using our positional clustering algorithm. Each cluster contains more than 1000 trajectories. We see that in each cluster, the trajectories begin in the same geographical area unique to each cluster. This is expected due to the increased attention by the average Haversine distance to the initial part of the trajectories. The discovered clusters contain trajectories from all different shipping lanes that originate in a given area. However, this is acceptable as we expect the second step clustering to separate the shipping lanes based on their



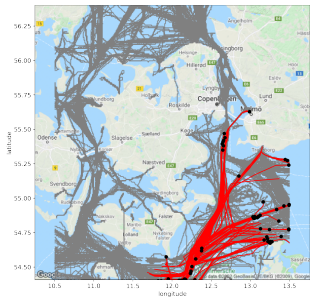
(a) DTW, hierarchical.



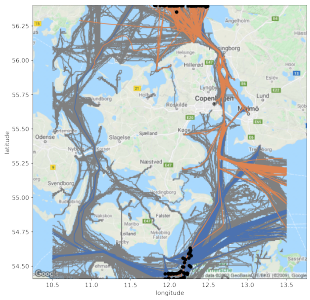
(b) Hausdorff, DBSCAN.



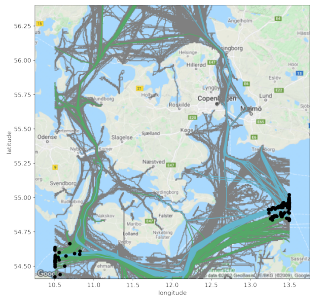
(c) Hausdorff, increased threshold.



(d) Hausdorff, DBSCAN outliers.



(e) Avg. Haversine, hierarchical.



(f) Avg. Haversine, hierarchical.

Figure 9.2 – Examples of clusters and outliers discovered in the first positional step using different similarity measures and clustering methods. Historical traffic is shown in grey and trajectory origins are denoted by a black circle.

common kinematic behaviour.

The merging of shipping lanes is similar to how clustering using the Hausdorff distance merged different shipping lanes when DBSCAN hyperparameters were optimized to reduce the number of outliers, as seen in Figure 9.2(c). However, shipping lanes are merged based on the origin and end location of the trajectory

because the Hausdorff distance is invariant to the direction of travel. This is a potential problem for real-time classification of incomplete trajectories into the discovered clusters.

9.4.3.3 Runtime Analysis

In Table 9.2, we report average runtimes for computing the average Haversine distance based on Eq. (9.3), DTW, and the Hausdorff distance. The distances were implemented in Python v. 3.8.11, programming language using Numpy v. 1.23.2. The Hausdorff and DTW distances were calculated using the *trajectory_ distance* library[2] implemented in Cython v. 0.29.24. With its linear time complexity, the average Haversine distance is undoubtedly the fastest to compute: it is 10 times faster than DTW and 100 times faster than the Hausdorff distance.

9.4.3.4 Summary

The traditional distance measures DTW and Hausdorff result in many outliers when applied to a complex, unfiltered dataset resembling trajectories expected in real-life applications. On the contrary, the average Haversine distance results in clusters that contain all the different routes that originate in a location that varies between clusters. Additionally, the average Haversine distance is much faster to compute, which allows for real-time assignment of unseen trajectories into precomputed clusters. Yet, these clusters do contain trajectories from many different maritime routes. Therefore, simply reducing the threshold in the hierarchical clustering does not equal a more detailed clustering describing their global positional or local kinematic behaviour. To achieve clusters that describe positional or kinematic behavior in more detail a different distance measure must thus be used during the second step.

Table 9.2 – Average time in seconds to compute a pair of trajectory similarities during computation of the distance matrix.

avg. Haversine	DTW	Hausdorff
$16.38\mu s$	$107.2\mu s$	$1084\mu s$

Table 9.3 – Hyperparameter values and clustering results of DBSCAN and hierarchical clustering using distances computed by Hausdorff, average Haversine distance, Eq. (9.3), or our proposed kinematic distance measure, Eq. (9.4), on trajectories assigned to positional cluster 0 in the first step.

Distance Measure	Clustering Method	Eps-Threshold	MinSamples	# Clusters	# Outliers	Silhouette Score
Avg. Haversine, Eq. (9.3)	Hierarchical	0.9	-	49	-	**0.558**
Hausdorff	Hierarchical	6250	-	134	-	0.533
Avg. Kinematic	Hierarchical	1.8	-	34	-	0.126
Kinematic	DBSCAN	12.0	46	2	821	0.036
Kinematic	DBSCAN	12.0	2	21	477	−0.221
Kinematic	Hierarchical	22.5	-	221	-	0.217

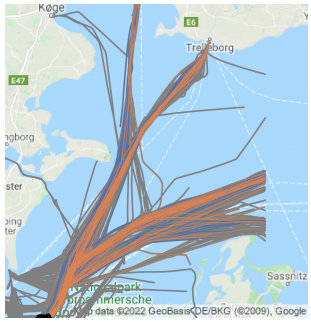
9.4.4 Step 2: Kinematic Clustering

In this section, we study the kinematic clustering of the second step using the positional cluster shown in blue in Figure 9.2(e). The trajectories in this cluster originate at the southeastern edge of the ROI and split into 4 major shipping lanes. One going west towards the Kieler Channel allowing passage to the Atlantic, one going north towards the North Sea, one going east towards the Baltic Sea, and one going northeast, terminating in the Swedish port of Trelleborg. In addition to these shipping lanes, the Danish port of Gedser (southern tip of the Lolland island) is a hub for pilot boats which often have to rendezvous with larger ships passing through the Fehmarn Belt between Denmark and Germany. These pilot boats form a triangle fanning outwards east from the port of Gedser seen in the bottom of Figure 9.2(e).

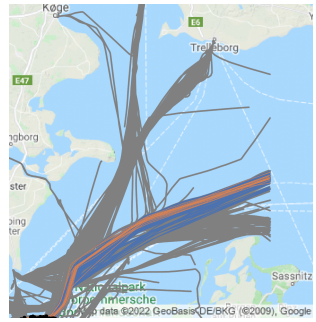
In order to evaluate the added value of kinematic features, we compare position and kinematic-based distance measures for the second step clustering. Regarding positional clustering, based on the results of Section 9.4.3, we consider only hierarchical clustering combined with the average Haversine distance and the Hausdorff distance. The former yielded better groupings, and the latter showed potential to further split clusters. As for the kinematic clustering, we test our proposed kinematic distance measure, Eq. (9.4) based on DTW. For completeness, we also include a DBSCAN clustering based on our kinematic distance. Finally, we also consider the average kinematic distance, Eq. (9.3) computed as the sum of speed and course differences computed by Eqs. (9.5)-(9.6).

We report in Table 9.3 hyperparameters and statistics about each clustering. The two positional-based clusterings find fewer clusters and obtain better silhouette scores than our proposed kinematic distance measure, Eq. (9.4). However, if we look at some of the clusters obtained in Figure 9.3(a)-(b), we notice that these former methods do not produce a more detailed clustering in terms of the local kinematic behavior of the trajectories.

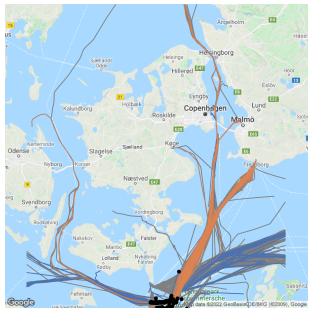
[2] `https://github.com/bguillouet/traj-dist`



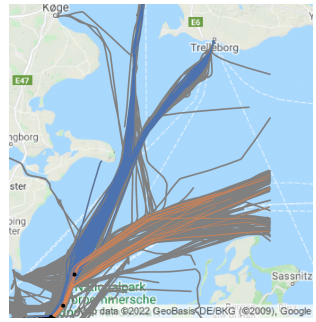
(a) Avg. Haversine.



(b) Hausdorff.



(c) Avg. Kinematic.

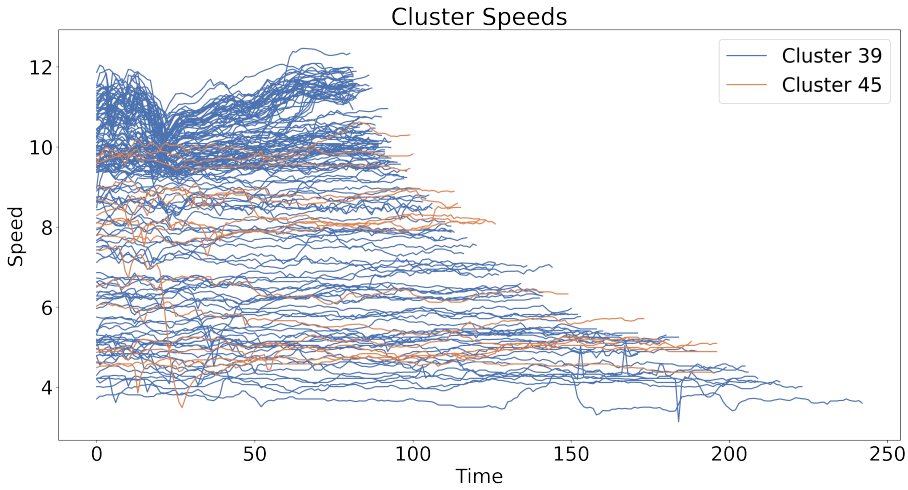


(d) Kinematic, Eq. (9.4).

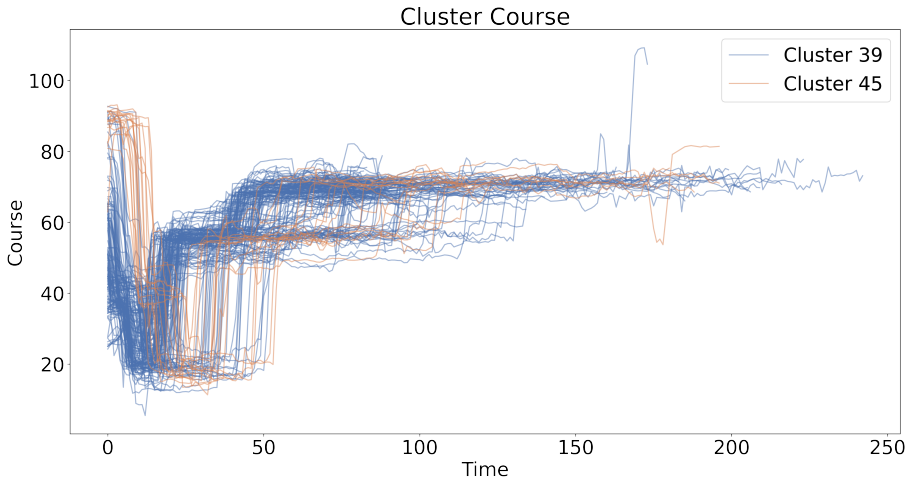
Figure 9.3 – Second step clusters obtained using hierarchical clustering and different similarity measures. Trajectories in grey denote trajectories in the same positional cluster and trajectory origins are denoted by a black circle.

9.4.4.1 Positional based Distance Measures

The clustering based on the average Haversine distance (Figure 9.3(a)) is unable to split the shipping lanes. We believe the focus on the initial part to be the cause. Using the Hausdorff distance for the second step clustering (Figure 9.3(b)) allows separating the maritime routes through the ROI. We also notice some maritime routes divided into two or more clusters, as seen in Figure 9.3(b). Thus, we have gained a more detailed clustering in terms of describing their global positional behaviour. In Figure 9.4, we show the speed and course of the trajectories assigned to the two clusters of Figure 9.3(b). We see that both clusters contain trajectories with different speeds and we notice trajectories that heavily decrease their speeds while in the shipping lane (blue trajectories with initial speed of about 12m/s). This is abnormal, and we would expect such trajectories to be outliers based on the kinematic features. Looking at the course, we see the two clusters generally have similar course changes, although they



(a) Speed.

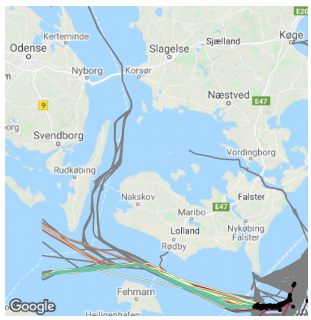


(b) Course.

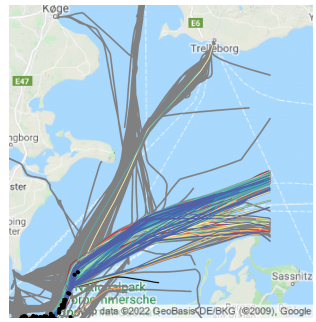
Figure 9.4 – Kinematic time series of the trajectories assigned to clusters obtained from the Hausdorff clustering shown in Figure 9.3(b).

happen at different times due to the time invariance of the Hausdorff distance. Based on the results above, we conclude that using the Hausdorff distance in the second step clustering results in a more detailed clustering regarding the global positional behaviour but not the local kinematic behaviour.

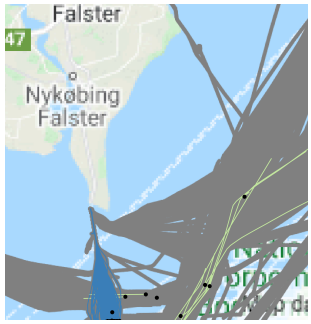
9.4.4.2 Kinematic based Distance Measures



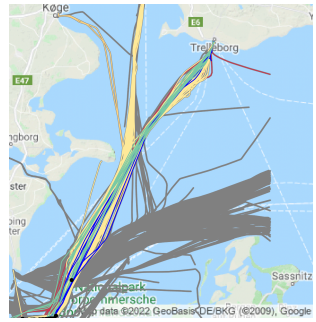
(a) Clusters going west.



(b) Clusters going east.



(c) Clusters to Gedser.

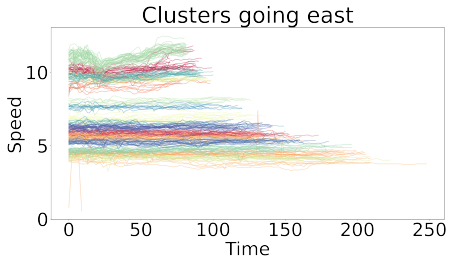


(d) Clusters to Trelleborg.

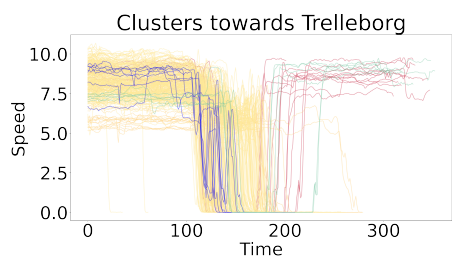
Figure 9.5 – Second step clusters with more than 5 assigned trajectories obtained using the kinematic distance matrix. Colours denote different clusters. Trajectories in grey denote trajectories in the same positional cluster.

Above we found positional-based distance measures unable to describe clusters with similar local kinematic behaviour. We now investigate the clusters obtained using kinematic-based distance measures, in particular our proposed kinematic distance measure (9.4). Our proposed distance measure obtains a higher silhouette score than the average kinematic distance, but the average kinematic distance finds fewer clusters. However, we see from Figure 9.3(c) that the average kinematic distance groups together trajectories that follow very different shipping lanes. Using Eq.(9.4) as the distance measure for hierarchical clustering, we find that a distance threshold of 22.5 is satisfactory using Kneedles algorithm. This threshold results in 221 different clusters. However, only 34 of these clusters have more than 5 trajectories assigned to them. These 34 clusters are pictured in Figure 9.5. The clusters are now separating the major maritime routes, and we notice clusters that are assigned trajectories that follow roughly the same route. Also, by looking at the speeds within these clusters (Figure

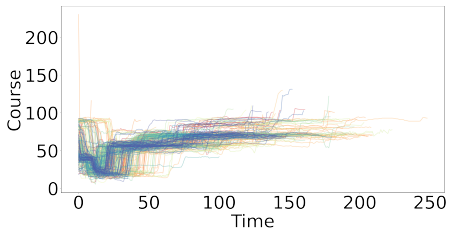
9.6(a-c)), we notice that trajectories within the same cluster tend to have the same speed and trajectories from different clusters have different speeds. In particular, the clusters distinguish five unique types of behaviour from the trajectories heading towards the port of Trelleborg, see Figure 9.6(b and d): slow speed returns south (green), fast speed return south (red), slow speed stops in port (orange), fast speed stops in port (yellow), fast speed stops in port with a spike in speed during slowdown (blue). Note that the high-frequency course changes at low speeds in Figure 9.6(d) are due to a vessel drifting in port. These random course changes may artificially decrease the similarity between trajectories of the same behaviour, but it is expected the two-stage DP compression [Klaas et al.] filter out the majority of these stationary periods at drift. Overall, the kinematic distance, Eq. (9.4) yields clusters with consistent kinematic behaviour. We notice another key difference with the Hausdorff case. Using the Hausdorff distance, the resulting clusters in Figure 9.3(b) become quite narrow, i.e. all the trajectories assigned to a cluster follow the same narrow maritime route. However, using D_{kin} Eq. (9.4) the resulting clusters have a larger geographical spread as seen in Figure 9.5(b). Thus, it indicates that the proposed kinematic distance measure, Eq. (9.4), is capable of clustering trajectories with similar kinematic behavior but in different geographical locations.



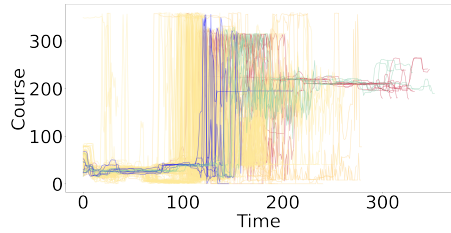
(a) Speed, eastbound.



(b) Speed, Trelleborg.



(c) Course, eastbound.



(d) Course, Trelleborg.

Figure 9.6 – Kinematic time series of all clusters following two different maritime routes, see Figure 9.5(b-d). Different colours represent different clusters.

All cases of pilot boats are not clustered together even though we know they must share some intrinsic behaviour. When inspecting the dendrogram, we ob-

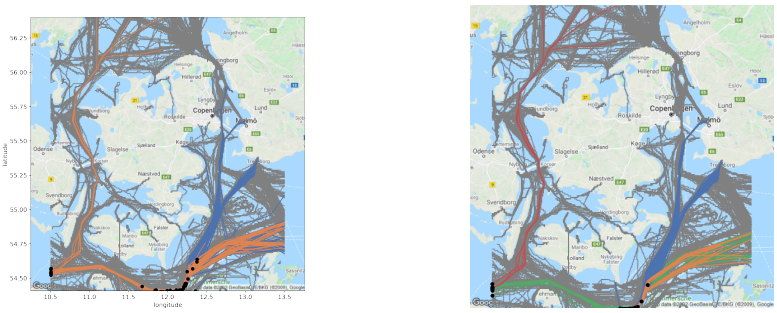
served that they are closer to one another than any other trajectories, but the distance between them is always above the selected distance threshold. Thus, there is a trade-off between clustering behaviour we know to be similar but that naturally has a higher distance, and clustering different behaviour that naturally is very close to one another. One might speculate whether a hierarchical density-based clustering approach as used in Murray and Perera [2021] would be able to correctly find clusters of varying density. However, most density-based approaches are based on some variant of single linkage, which we empirically find not to perform very well. We find our choice of hierarchical clustering with average linkage to be superior to variants of DBSCAN, as shown in Table 9.3. Using DBSCAN with the proposed hyperparameter selection, we find only two large clusters and the majority of trajectories assigned as outliers. Reducing the minimum number of sample required to define clusters increased the number of cluster to 21, but the vast majority of trajectories was assigned to the same cluster. This would not allow us to make any distinction between types of kinematic behaviour. We thus focused on hierarchical clustering with average linkage.

9.4.5 Single Step Clustering

We now compare the proposed two-step clustering to a single hierarchical clustering merging both steps using as similarity measure the sum of the average Haversine distance, Eq. (9.3) and of DTW based on D_{kin} of Eq. (9.4). The Kneedles algorithm for the single-step clustering obtains a silhouette score of 0.095 and finds 2880 clusters. For comparison, the two-step clustering approach achieves a silhouette score of 0.162 and a total of 6963 clusters when applied to the entire dataset. In both methods, the majority of the clusters contain only 1 observation, which we previously highlighted for the two-step algorithm. Using the sum Eq. (9.3) and Eq. (9.4) may average out some of the differences in either position or kinematics that may be captured by modelling the similarity measures individually. Therefore, it is expected that the two-step returns more clusters.

Using the single-step clustering, trajectories travelling along different shipping lanes may be clustered together (Figure 9.7(a)), while our two-step algorithm disentangles them (Figure 9.7(b)). For example, the single-step approach groups together (orange cluster of Figure 9.7(a)) trajectories with distant initial points and following different shipping lanes because the speed of the trajectories are very similar. Thus, differences in the position are compensated by similar speed behavior. The two-stage algorithm splits these two routes into multiple clusters since it is able to disentangle the positional information from the kinematic.

As discussed previously, the 2-step approach has a trade-off between clustering behaviour we know to be similar, but that naturally has a higher distance, and clustering different behaviour that naturally is very close to one another. Using the single-step approach seems to automatically push this trade-off towards the latter option. Thus, our proposed 2-step approach allows for a more detailed clustering in terms of the kinematic behaviour and a better disentanglement of position and kinematic behaviour.

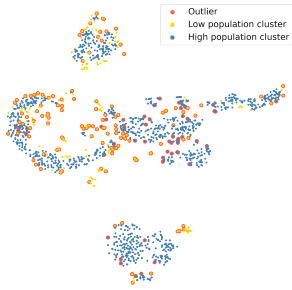


(a) Single-step clusters. (b) Two-step clusters.

Figure 9.7 – Trajectory clusters produced by the single-step clustering (a) that are split into multiple clusters (b) using the the two-step algorithm.

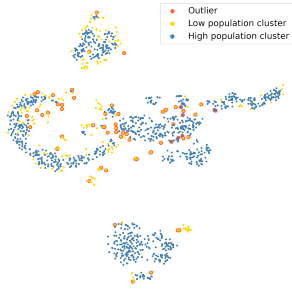
9.4.6 Outlier Contamination Analysis

In this section, we qualitatively explore how the kinematic distance measure may be used to find trajectories with similar kinematic behaviour regardless of the position and identify abnormal kinematic behaviour. The positional clustering is trained as described above, and the kinematic clustering is exchanged for outlier detection using LOF. In Figure 9.8, we plot a T-SNE embedding of the kinematic distance matrix and the results of LOF with different contamination levels of all trajectories of the blue positional cluster of Figure9.2(e). The contamination hyperparameter is related to the expected number of outliers in the data and is used to decide the threshold on the outliers scores. We compare (a) the `auto` option where the level is decided as suggested in Breunig et al. [2000], (b) 0.1 , (c) 0.05, and (d) 0.01. The `auto` options returned a contamination level decided from the data of 0.14. Points in clusters with at least 5 trajectories are in blue, and the others are in gray. Outliers are denoted by red borders. We see that the majority of the flagged outliers are observations that do not belong to highly populated clusters. However, some trajectories on the border of highly populated clusters (blue) are also marked as outliers. Most of the single cluster trajectories are marked as outliers except in the case where multiple of these trajectories join together in small groups of minimum 5 trajectories as,



(a) Contam. auto=0.14, 178 outliers.

(b) Contam. 0.10, 128 outliers.



(c) Contam. 0.05, 64 outliers.

(d) Contam. 0.01, 13 outliers.

Figure 9.8 – TSNE of the kinematic distance matrix of the blue cluster of Figure 9.2(e) with varying levels of contamination in LOF. Crosses denote detected outliers. Colors denote cluster assignments with grey being clusters with less than 5 members.

for instance, the majority of the pilot boats discussed previously. Fixing the contamination level acts as an upper bound on the number of marked outliers. Reducing the contamination level to 0.05 in Figure 9.8(c) seems to almost remove the outliers clustered around highly populated clusters.Reducing it further to 0.01 leaves almost no outliers. Thus, it seems that contamination levels from 0.05 to 0.15 would work best in practice.

9.4.7 Outliers and Embedding Analysis

In Figure 9.9, we plot a TSNE representation of the kinematic distance matrix of all trajectories in positional cluster 0, shown in blue in 9.2(e). Colours denote cluster assignments from the kinematic clustering, and grey dots are clusters with at most 5 trajectories. Red borders denote outliers flagged by LOF using a contamination level of 0.05. Squares denote embeddings of pilot boats, triangle

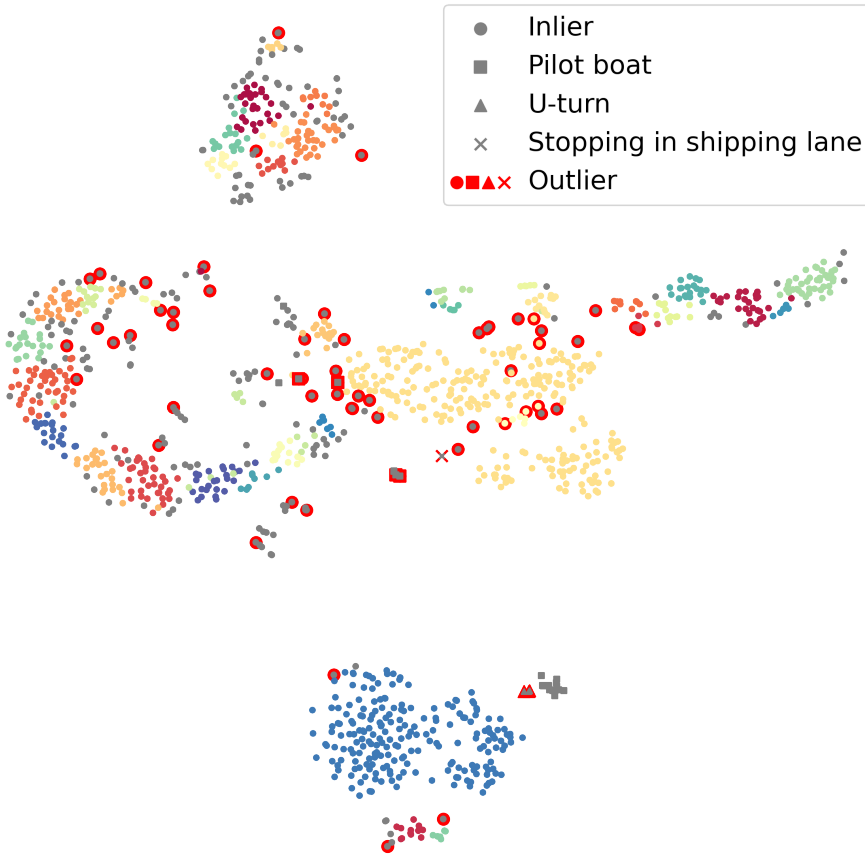
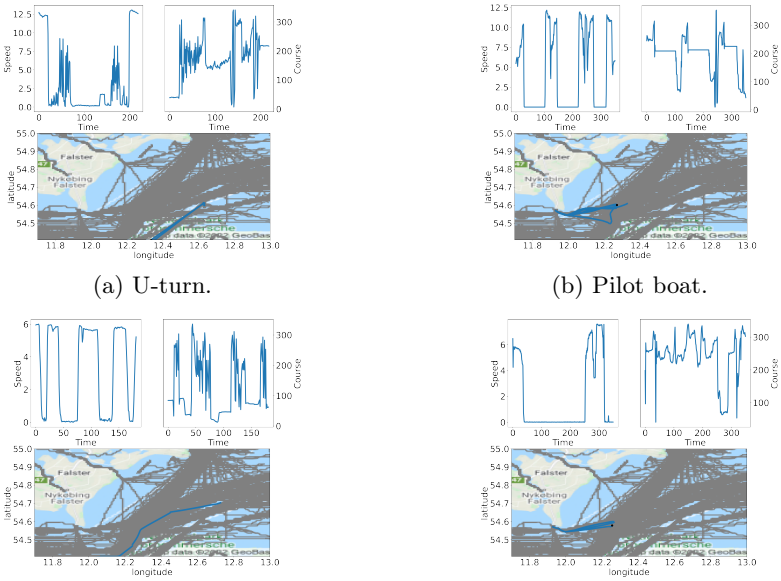


Figure 9.9 – TSNE of the kinematic distance matrix of the blue cluster in Figure 9.2(e). Colours denote the kinematic cluster assignment with grey being clusters with less than 5 members. Red borders denote outliers flagged by LOF. Shapes denote manually identified special trajectories.

trajectories with u-turns, and crosses trajectories with stops outside a port.

As we noted above, the majority of the pilot boats closest neighbours were other pilot boats which we also see in Figure 9.9. The majority of the pilot boats form a small unclustered group near a larger cluster (blue) in the bottom of the figure. This cluster represents ships coming from the south docking in the port of Gedser and leaving south. In the same area, we find cases of ships (red, green) sailing north and docking at Trelleborg before returning south. In between the pilot boats and the large blue cluster we notice three trajectories

discovered to be outliers by LOF. These three trajectories are boats coming from the south and making a stop and U-turn in the middle of the shipping lane very similar to the pilot boats. This is another example of the capabilities of our proposed kinematic distance measure to capture similar kinematic behaviour regardless of the position. However, small differences in the course time series lead the three U-turns trajectories to be flagged as abnormal. In Figure 9.10(a) and (b) we show an example of U-turn trajectories and the pilot boats nearby.



(a) U-turn.

(b) Pilot boat.

(c) Multiple stops on shipping lane.

(d) Pilot boat.

Figure 9.10 – Trajectories of abnormal activity such as U-turns (a) and stopping in the shipping lane (c) found to have similarity with the behaviour of pilot boats (b), (d).

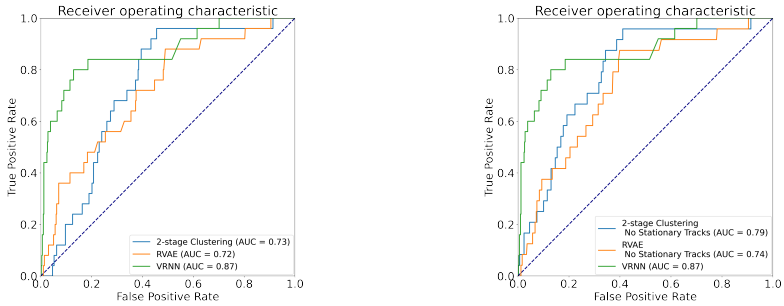
We also notice the tail clustered into many different clusters in the left of Figure 9.9. These clusters correspond to the traffic following the shipping lanes east (Figure 9.5(b)) at different speeds. To the right of this tail, we find some pilot boats (square) and a highlighted outlier corresponding to a trajectory following the shipping lanes going east with multiple sudden stops in the middle of the shipping lane (cross). This type of behaviour is of interest for certain ship types, such as research vessels and diving vessels. These two trajectories are shown in Figure 9.10(c) and (d). The large yellow cluster in the center of Figure 9.9 corresponds to trajectories going north and stopping in the port of Trelleborg. Contrary to the previous smaller red and green clusters in the bottom of Figure 9.9, these trajectories end in the port. This means that our proposed kinematic distance measure has found the start-stopping behaviour to be more akin to the

trajectories ending in port or pilot boats, however, not so much as to be part of their cluster due to differences in the course. Thus, we find our proposed distance measure is capable of capturing local kinematic similarities regardless of the geographical position and flag trajectories with abnormal local kinematic behaviour.

9.4.8 Anomaly Detection

We now study the proposed anomaly method (Section 9.3.3), which involves inductive clustering (i.e. clustering on new unseen data) and outlier detection on the data from the Bornholm area on December 13^{th}. The positional clustering and LOF are trained using data from December 2021 except that of December 13^{th}. Trajectories of that day are then processed as in a fully automated streaming scenario: A trajectory is first assigned to positional a cluster using a KNN classifier with $k = 3$, then kinetic similarities to other trajectories of the cluster are computed and fed to the LOF model which may trigger an alarm based on the outlier scores.

We investigate the precision of our model and discuss which value of contamination parameter to use based on the receiver operating curve from outlier detection on the December 13^{th} data as shown in Figure 9.11. We compare to the A-Contrario outlier detection method [Nguyen et al., 2021] using RVAE [Murray and Perera, 2021] and VRNN [Nguyen et al., 2021].



(a) All trajectories.

(b) Without stationary trajectories.

Figure 9.11 – ROC of the outlier detection on the Bornholm 13^{th} December data.

We see that our method outperforms detection based on RVAE reconstruction but not based on VRNN reconstruction. In particular, we see that our method suffers from false positives early in the detection. Looking at these trajectories,

we notice that the vessels are mostly stationary in port with a highly varying course resulting in a large directional distance to all other trajectories. By removing trajectories with an average speed of less than $0.3m/s$ from the detection, we can reduce the number of early false positives to similar levels as RVAE.

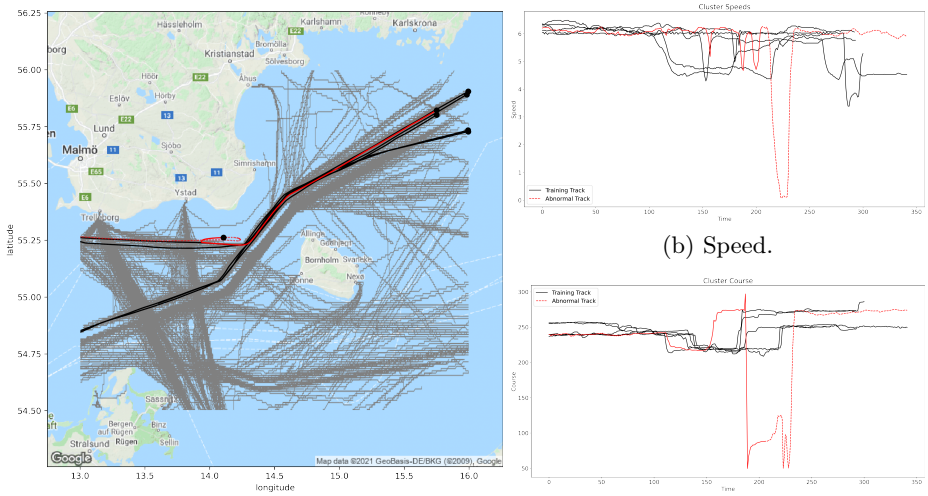
Previously, we counted outliers to infer the contamination parameter. Here, we show how the ROC curve may provide more intuition behind this hyperparameter. In order to detect 96% of the abnormal trajectories a false alarm is triggered on 40% of the normal trajectories which corresponds to a contamination rate of 0.43. A contamination level of 0.05 and 0.10 flag about 20% and 25% of the abnormal trajectories respectively, and the value discovered on the Sjælland data of 0.14 would result in a true positive rate of approximately 40%.

9.4.9 Anomaly Detection and Interpretation

The performance of our proposed method is below that of the VRNN. However, the time complexity of our proposed clustering methodology is much lower than using A-Contrario detection with the VRNN model. The evaluation time of VRNN and A-Contrario detection heavily depends on the resample frequency during preprocessing and the maximum sequence length evaluated during A-Contrario detection. In our experiments, the VRNN model and A-Contrario detection had an average evaluation time of 176 seconds pr trajectory. In comparison, our clustering approach average an evaluation time of only 6.8 seconds per trajectory. Furthermore, our method is based on similarities that allow for fast comparison with the expected behaviour and may be used to quickly identify the most similar trajectories. This could be used as an additional explainability to surveillance operators or be used to design rules for abnormal trajectories.

In Figure 9.12 we show in red the geographical evolution (a), speed (b) and course (c) time series of an abnormal trajectory making a double U-turn in the shipping lane. The trajectory was made by the sister ship of the vessel which caused a collision accident and was returning to the site of the collision perhaps to transfer crew to the collided vessel.

We plot in black the five most similar trajectories from the training set (different day) according to the proposed distance measures, Eq. (9.3) and Eq. (9.4) computed until time step $t = 205$ which corresponds to the period after the first u-turn when the vessel is travelling in the opposite direction of the shipping lane . All five trajectories originate in the northeastern part of the ROI, and all five vessels have extended periods of time in which they travel at reduced speeds. Surveillance operators may use the description of the most similar trajectories



(a) Geographical evolution.

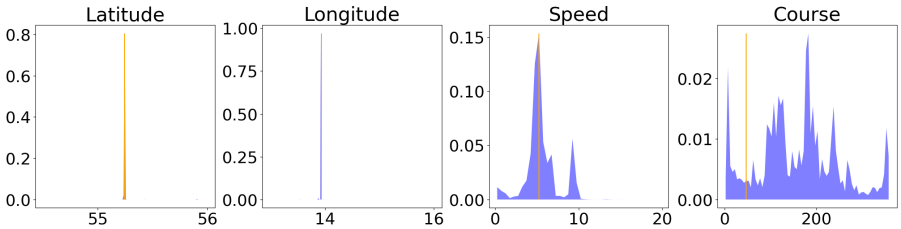
(b) Speed.

(c) Course.

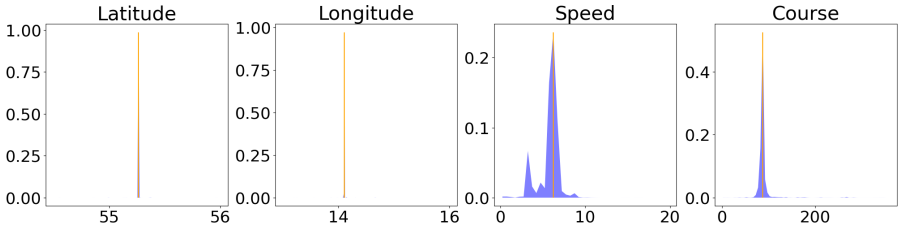
Figure 9.12 – Top five most similar trajectories to the abnormal trajectory in red determined by Eq. (9.4) at time step $t = 205$. The future trajectory is shown in dashes.

to categorize maritime behaviour in real time. Coupled with the operators' expectations towards the behaviour of the given ship type in the ROI, this may be used to identify vessels behaving abnormally. Additionally, surveillance operators could make predefined rules based on the clusters or trajectories of the training set. Alarms could be triggered if certain known abnormal trajectories are found as the nearest neighbours or ship types are classified to clusters with unexpected behaviour from the given ship type. Surveillance operators may prefer such human-in-the-loop decision making.

Information about the current behaviour may not be readily available in the VRNN model and may have to be discovered manually post detection. The VRNN outputs the probabilities of a multivariate Bernoulli distribution. We can evaluate how the actual values relate to the expected values by the VRNN models for each time update. This is illustrated in Figure 9.13 where we plot the multivariate Bernoulli distribution from the VRNN model at two time steps during and after the first u-turn. During the u-turn the speed and course are not reconstructed accurately. The multivariate Bernoulli distribution may be used to understand which input feature is abnormal, and surveillance operators may use this information to evaluate the behaviour. However, after the u-turn all four input features are reconstructed accurately. Thus, there is a risk that operators may disregard alarms thinking that everything went back to normal.



(a) During first u-turn $t = 189$.



(b) After first u-turn $t = 205$.

Figure 9.13 – multivariate Bernoulli distribution (blue) for the VRNN output at two time steps observed during and after the first u-turn. Actual values in orange.

9.5 Conclusion

In this work, we have proposed a 2-step clustering methodology for maritime trajectories. The two steps allow the clustering to better focus on the kinematic behaviour expected in a certain geographical position and help disentangle the geographical position with kinematic behaviour. This is useful information for surveillance operators in determining abnormal behaviour understood as combinations of physical position, kinematics, ship type, etc. We propose two trajectory similarity metrics. One metric allows for fast computation of positional similarity and reduces the need of time-costly hyperparameter optimization. The other metric captures the kinematic behaviour of the trajectories which improves clustering in terms of local kinematic behaviour.

We have also provided a method for automatic outlier detection using the designed similarity metrics and find it comparable to reconstruction based detection using deep neural networks. Although our proposed model achieves a lower area under the ROC than the VRNN model, it has a clear advantage in terms of explainability over current deep learning methods. In future work, we aim to compute the proposed similarity measures with neural networks and to utilize deep models for reconstruction based outlier detection that also align trajecto-

ries in the latent space in order to preserve a certain level of explainability.

Bibliography

International Convention for the Safety of Life at Sea (SOLAS), Chapter V: Safety of Navigation, Regulation 19, 13 December 2002. Technical report. URL www.imo.org.

MarineTraffic - A day in numbers - MarineTraffic Blog. URL https://www.marinetraffic.com/blog/a-day-in-numbers/.

I. Abualhaol, R. Falcon, R. Abielmona, and E. Petriu. Data-Driven Vessel Service Time Forecasting using Long Short-Term Memory Recurrent Neural Networks. *Proceedings - 2018 IEEE International Conference on Big Data, Big Data 2018*, pages 2580–2590, 1 2019. doi: 10.1109/BIGDATA.2018.8622626.

M. Ankerst, M. M. Breunig, H.-P. Kriegel, and J. Sander. OPTICS: Ordering Points to Identify the Clustering Structure. *SIGMOD Rec.*, 28(2):49–60, 6 1999. ISSN 0163-5808. doi: 10.1145/304181.304187. URL https://doi-org.proxy.findit.cvt.dk/10.1145/304181.304187.

M. Anneken, Y. Fischer, and J. Beyerer. Evaluation and comparison of anomaly detection algorithms in annotated datasets from the maritime domain. In *IntelliSys 2015 - Proceedings of 2015 SAI Intelligent Systems Conference*, pages 169–178. Institute of Electrical and Electronics Engineers Inc., 12 2015. ISBN 9781467376068. doi: 10.1109/IntelliSys.2015.7361141.

V. F. Arguedas, G. Pallotta, and M. Vespe. Automatic generation of geographical networks for maritime traffic surveillance. In *17th International Conference on Information Fusion (FUSION)*. IEEE, 2014.

V. F. Arguedas, F. Mazzarella, and M. Vespe. Spatio-temporal data mining for maritime situational awareness. In *MTS/IEEE OCEANS 2015 - Genova: Discovering Sustainable Ocean Energy for a New World*. Institute of Electrical

and Electronics Engineers Inc., 9 2015. ISBN 9781479987368. doi: 10.1109/OCEANS-Genova.2015.7271544.

J. Bai, W. Wang, and C. Gomes. Contrastively Disentangled Sequential Variational Autoencoder. *Advances in Neural Information Processing Systems*, 34, 12 2021.

Bloomberg News. Ukraine's Grain Corridors Still Need Ships to Ease Food Crisis, 8 2022. URL https://www.bloomberg.com/news/articles/2022-08-06/ukraine-s-grain-corridors-still-need-ships-to-ease-food-crisis.

G. Bombara, C. I. Vasile, F. Penedo, H. Yasuoka, and C. Belta. A decision tree approach to data classification using signal temporal logic. In *HSCC 2016 - Proceedings of the 19th International Conference on Hybrid Systems: Computation and Control*, pages 1–10. Association for Computing Machinery, Inc, 4 2016. ISBN 9781450339551. doi: 10.1145/2883817.2883843.

A. Boubekki, M. Kampffmeyer, U. Brefeld, and R. Jenssen. Joint optimization of an autoencoder for clustering and embedding. *Machine Learning*, 110(7):1901–1937, 7 2021. ISSN 15730565. doi: 10.1007/S10994-021-06015-5/TABLES/15. URL https://link.springer.com/article/10.1007/s10994-021-06015-5.

S. R. Bowman, L. Vilnis, O. Vinyals, A. M. Dai, R. Jozefowicz, and S. Bengio. Generating Sentences from a Continuous Space. *CoNLL 2016 - 20th SIGNLL Conference on Computational Natural Language Learning, Proceedings*, pages 10–21, 11 2015. doi: 10.18653/v1/k16-1002. URL https://arxiv.org/abs/1511.06349v4.

P. Braca, E. d'Afflisio, L. M. Millefiori, and P. Willett. Detecting Anomalous Deviations from Standard Maritime Routes Using the Ornstein-Uhlenbeck Process. *IEEE Transactions on Signal Processing*, 66(24):6474 – 6487, 2018. doi: 10.1109/TSP.2018.2875887.

M. M. Breunig, H.-P. Kriegel, R. T. Ng, and J. Sander. LOF: Identifying Density-Based Local Outliers. In *Proceedings of the 2000 ACM SIGMOD International Conference on Management of Data*, SIGMOD '00, pages 93–104, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132174. doi: 10.1145/342009.335388. URL https://doi-org.proxy.findit.cvt.dk/10.1145/342009.335388.

B. Brito, H. Zhu, W. Pan, and J. Alonso-Mora. Social-VRNN: One-Shot Multimodal Trajectory Prediction for Interacting Pedestrians. *CoRR*, 2020. URL https://github.com/tud-amr/social_vrnn.git.

S. Capobianco, N. Forti, L. M. Millefiori, P. Braca, and P. Willett. Uncertainty-Aware Recurrent Encoder-Decoder Networks for Vessel Trajectory Prediction. *Proceedings of 2021 IEEE 24th International Conference on Information Fusion, FUSION 2021*, 2021a. doi: 10.23919/FUSION49465.2021.9626839.

S. Capobianco, L. M. Millefiori, N. Forti, P. Braca, and P. Willett. Deep Learning Methods for Vessel Trajectory Prediction Based on Recurrent Neural Networks. *IEEE Transactions on Aerospace and Electronic Systems*, 57(6):4329–4346, 12 2021b. ISSN 15579603. doi: 10.1109/TAES.2021.3096873.

S. Capobianco, N. Forti, L. M. Millefiori, P. Braca, S. Member, and P. Willett. Recurrent Encoder-Decoder Networks for Vessel Trajectory Prediction with Uncertainty Estimation. 5 2022. doi: 10.48550/arxiv.2205.05404. URL `https://arxiv.org/abs/2205.05404v1`.

Y. L. Chang, A. Anagaw, L. Chang, Y. C. Wang, C. Y. Hsiao, and W. H. Lee. Ship Detection Based on YOLOv2 for SAR Imagery. *Remote Sensing 2019, Vol. 11, Page 786*, 11(7):786, 4 2019. ISSN 2072-4292. doi: 10.3390/RS11070786. URL `https://www.mdpi.com/2072-4292/11/7/786/htmhttps://www.mdpi.com/2072-4292/11/7/786`.

X. Chen, Y. Liu, K. Achuthan, and X. Zhang. A ship movement classification based on Automatic Identification System (AIS) data using Convolutional Neural Network. *Ocean Engineering*, 218:108182, 12 2020. ISSN 0029-8018. doi: 10.1016/J.OCEANENG.2020.108182.

E. Chondrodima, P. Mandalis, N. Pelekis, and Y. Theodoridis. Machine Learning Models for Vessel Route Forecasting: An Experimental Comparison. pages 262–269, 8 2022. doi: 10.1109/MDM55031.2022.00056.

J. Chung, C. Gulcehre, K. Cho, and Y. Bengio. Empirical Evaluation of Gated Recurrent Neural Networks on Sequence Modeling. 12 2014. URL `https://arxiv.org/abs/1412.3555v1`.

J. Chung, K. Kastner, L. Dinh, K. Goel, A. C. Courville, and Y. Bengio. A Recurrent Latent Variable Model for Sequential Data. In *Advances in Neural Information Processing Systems*, volume 28, 2015. URL `http://www.github.com/jych/nips2015_vrnn`.

C. Costello, L. Cao, S. Gelcich, M. Cisneros-Mata, C. M. Free, H. E. Froehlich, C. D. Golden, G. Ishimura, J. Maier, I. Macadam-Somer, T. Mangin, M. C. Melnychuk, M. Miyahara, C. L. de Moor, R. Naylor, L. Nøstbakken, E. Ojea, E. O'Reilly, A. M. Parma, A. J. Plantinga, S. H. Thilsted, and J. Lubchenco. The future of food from the sea. *Nature*, 588(7836):95–100, 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2616-y. URL `https://doi.org/10.1038/s41586-020-2616-y`.

E. D'Afflisio, P. Braca, L. M. Millefiori, and P. Willett. Maritime Anomaly Detection Based on Mean-Reverting Stochastic Processes Applied to a Real-World Scenario. In *2018 21st International Conference on Information Fusion, FUSION 2018*, pages 1171–1177. Institute of Electrical and Electronics Engineers Inc., 9 2018. ISBN 9780996452762. doi: 10.23919/ICIF.2018.8455854.

S. Dan Vukša, P. Vidan, M. Bukljaš, and S. P. Pavi´c. Research on Ship Collision Probability Model Based on Monte Carlo Simulation and Bi-LSTM. *Journal of Marine Science and Engineering 2022, Vol. 10, Page 1124*, 10(8):1124, 8 2022. ISSN 2077-1312. doi: 10.3390/JMSE10081124. URL https://www.mdpi.com/2077-1312/10/8/1124/htmhttps://www.mdpi.com/2077-1312/10/8/1124.

P. Dijt and P. Mettes. Trajectory Prediction Network for Future Anticipation of Ships. 2020. doi: 10.1145/3372278.3390676. URL https://doi.org/10.1145/3372278.3390676.

M. Ding, W. Su, Y. Liu, J. Zhang, J. Li, and J. Wu. A Novel Approach on Vessel Trajectory Prediction Based on Variational LSTM. In *Proceedings of 2020 IEEE International Conference on Artificial Intelligence and Computer Applications, ICAICA 2020*, pages 206–211. Institute of Electrical and Electronics Engineers Inc., 6 2020. ISBN 9781728170046. doi: 10.1109/ICAICA50127.2020.9182537.

D. H. Douglas and T. K. Peucker. Algorithms for the Reduction of the Number of Points Required to Represent a Digitized Line or its Caricature. *Classics in Cartography: Reflections on Influential Articles from Cartographica*, pages 15–28, 1 2011. doi: 10.1002/9780470669488.CH2.

H. Duan, F. Ma, L. Miao, and C. Zhang. A semi-supervised deep learning approach for vessel trajectory classification based on AIS data. *Ocean & Coastal Management*, 218:106015, 2022. ISSN 0964-5691. doi: https://doi.org/10.1016/j.ocecoaman.2021.106015. URL https://www.sciencedirect.com/science/article/pii/S096456912100497X.

L. Eljabu, M. Etemad, and S. Matwin. Anomaly Detection in Maritime Domain based on Spatio-Temporal Analysis of AIS Data Using Graph Neural Networks. *Proceedings - 2021 5th International Conference on Vision, Image and Signal Processing, ICVISP 2021*, pages 142–147, 2021. doi: 10.1109/ICVISP54630.2021.00033.

M. R. Endsley. Toward a theory of situation awareness in dynamic systems. *Human Factors*, 37(1):32–64, 1995. ISSN 00187208. doi: 10.1518/001872095779049543. URL https://www.researchgate.net/publication/210198492_Endsley_MR_Toward_a_Theory_of_Situation_Awareness_in_Dynamic_Systems_Human_Factors_Journal_371_32-64.

M. R. Endsley. From Here to Autonomy: Lessons Learned From Human–Automation Research. *Human Factors*, 59(1):5–27, 2017. doi: 10.1177/0018720816681350. URL https://doi.org/10.1177/0018720816681350.

M. Ester, H.-p. Kriegel, J. Sander, and X. Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. *Proceedings of 4th Int. Conf. on Knowledge Discovery and Data Mining*, 1996.

M. D. Ferreira, G. Spadon, A. Soares, and S. Matwin. A Semi-Supervised Methodology for Fishing Activity Detection Using the Geometry behind the Trajectory of Multiple Vessels. *Sensors*, 22(16), 2022. ISSN 1424-8220. doi: 10.3390/s22166063. URL https://www.mdpi.com/1424-8220/22/16/6063.

J. A. Figueroa and A. R. Rivera. Learning to Cluster with Auxiliary Tasks: A Semi-Supervised Approach. *Proceedings - 30th Conference on Graphics, Patterns and Images, SIBGRAPI 2017*, pages 141–148, 11 2017. doi: 10.1109/SIBGRAPI.2017.25.

N. Forti, L. M. Millefiori, and P. Braca. Hybrid Bernoulli Filtering for Detection and Tracking of Anomalous Path Deviations. In *2018 21st International Conference on Information Fusion, FUSION 2018*, pages 1178–1184. Institute of Electrical and Electronics Engineers Inc., 9 2018. ISBN 9780996452762. doi: 10.23919/ICIF.2018.8455567.

N. Forti, L. M. Millefiori, P. Braca, and P. Willett. Anomaly Detection and Tracking Based on Mean-Reverting Processes with Unknown Parameters. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2019-May, pages 8449–8453. Institute of Electrical and Electronics Engineers Inc., 5 2019. ISBN 9781479981311. doi: 10.1109/ICASSP.2019.8683428.

N. Forti, L. M. Millefiori, P. Braca, and P. Willett. Prediction oof Vessel Trajectories from AIS Data Via Sequence-To-Sequence Recurrent Neural Networks. *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, 2020-May:8936–8940, 5 2020. ISSN 15206149. doi: 10.1109/ICASSP40776.2020.9054421.

S. Freitas, H. Silva, J. M. Almeida, and E. Silva. Convolutional neural network target detection in hyperspectral imaging for maritime surveillance. *International Journal of Advanced Robotic Systems*, 16(3), 4 2019. ISSN 1729-8814. doi: 10.1177/1729881419842991. URL https://doaj.org/article/4e9e6d8cc8bf490a988e9bb0bdbad946.

S. Gan, S. Liang, K. Li, J. Deng, and T. Cheng. Ship trajectory prediction for intelligent traffic management using clustering and ANN. *2016 UKACC International Conference on Control, UKACC Control 2016*, 11 2016. doi: 10.1109/CONTROL.2016.7737569.

S. Gan, S. Liang, K. Li, J. Deng, and T. Cheng. Trajectory Length Prediction for Intelligent Traffic Signaling: A Data-Driven Approach. *IEEE Transactions on Intelligent Transportation Systems*, 19(2):426–435, 2 2018. ISSN 15249050. doi: 10.1109/TITS.2017.2700209.

D. w. Gao, Y. s. Zhu, J. f. Zhang, Y. k. He, K. Yan, and B. r. Yan. A novel MP-LSTM method for ship trajectory prediction based on AIS data. *Ocean Engineering*, 228, 5 2021. ISSN 00298018. doi: 10.1016/j.oceaneng.2021.108956.

M. Gao, G. Shi, and S. Li. Online prediction of ship behavior with automatic identification system sensor data using bidirectional long short-term memory recurrent neural network. *Sensors (Switzerland)*, 18(12), 2018. ISSN 14248220. doi: 10.3390/s18124211.

A. Goyal, A. Sordoni, M. Maluuba, M.-A. Côté, N. Rosemary, K. Mila, P. Montréal, and Y. Bengio. Z-Forcing: Training Stochastic Recurrent Networks. *Neural information processing systems foundation*, 2017(31st Annual Conference on Neural Information Processing Systems):6714–6724, 2017. ISSN 10495258.

T. Grant and B. Kooter. Comparing OODA & other models as Operational View C2 Architecture. 12 2005.

A. Graves. Generating Sequences With Recurrent Neural Networks. *arXiv preprint arXiv:1308.0850*, 8 2013. URL `https://arxiv.org/abs/1308.0850v5`.

D. O. D. Handayani, W. Sediono, and A. Shah. Anomaly detection in vessel tracking using support vector machines (SVMs). In *Proceedings - 2013 International Conference on Advanced Computer Science Applications and Technologies, ACSAT 2013*, pages 213–217. IEEE Computer Society, 2013. ISBN 9781479927586. doi: 10.1109/ACSAT.2013.49.

J. A. Hanley and B. J. McNeil. The meaning and use of the area under a receiver operating characteristic (ROC) curve. *Radiology*, 143(1):29–36, 5 1982. doi: https://doi.org/10.1148/radiology.143.1.7063747.

S. Hochreiter and J. Schmidhuber. Long Short-Term Memory. *Neural Computation*, 9(8):1735–1780, 11 1997. ISSN 08997667. doi: 10.1162/NECO.1997.9.8.1735.

L. Hovgaard. GPS-ekspert om udfald over Kattegat: 'Sandsynligvis kraftig jamming' - In Danish, 2022. URL `https://ing.dk/artikel/gps-ekspert-udfald-kattegat-sandsynligvis-kraftig-jamming-262079`.

J. Hu, K. Kaur, H. Lin, X. Wang, M. M. Hassan, I. Razzak, and M. Hammoudeh. Intelligent Anomaly Detection of Trajectories for IoT Empowered Maritime Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems*, 2022. ISSN 15580016. doi: 10.1109/TITS.2022.3162491.

G. Huang, S. Lai, C. Ye, and H. Zhou. Ship trajectory anomaly detection based on multi-feature fusion. In *2021 IEEE International Conference on Smart Data Services (SMDS)*, pages 72–81, 2021. doi: 10.1109/SMDS53860.2021.00020.

IMO. About IMO, 2021. URL `http://www.imo.org/en/About/Pages/Default.aspx`.

G. Jin, H. Sha, Y. Feng, Q. Cheng, and J. Huang. Modeling Spatiotemporal Geographic-Semantic Dynamics for Urban Hotspots Prediction. 2020. doi: 10.1145/1122445.1122456. URL `https://doi.org/10.1145/1122445.1122456`.

A.-L. Jousselme and G. Pallotta. Dissecting uncertainty-based fusion techniques for maritime anomaly detection. In *2015 18th International Conference on Information Fusion (Fusion)*. IEEE, 2015.

J. S. Kim, J. S. Lee, and K. I. Kim. Anomalous vessel behavior detection based on SVR seaway model. *International Journal of Fuzzy Logic and Intelligent Systems*, 19(1):18–27, 2019. ISSN 2093744X. doi: 10.5391/IJFIS.2019.19.1.18.

D. P. Kingma and M. Welling. Auto-Encoding Variational Bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*, 12 2013. URL `https://arxiv.org/abs/1312.6114v10`.

G. Klaas, D. De Vries, and M. Van Someren. Machine learning for vessel trajectories using compression, alignments and domain knowledge. doi: 10.1016/j.eswa.2012.05.060. URL `http://dx.doi.org/10.1016/j.eswa.2012.05.060`.

K. Kowalska and L. Peel. Maritime anomaly detection using Gaussian Process active learning. In *15th International Conference on Information Fusion, Fusion 2012*, page 6289940. IEEE Computer Society, 2012. ISBN 9781467304177.

D. Kroodsma, T. Hochberg, F. Paolo, and P. Davis. Dark Vessels - Revealing all vessel traffic at sea. URL `https://globalfishingwatch.org/research-project-dark-vessels/`.

Lauro Snidaro, Ingrid Visentini, Karna Bryan, and Gian Luca Foresti. Markov Logic Networks for context integration and situation assessment in maritime domain. In *2012 15th International Conference on Information Fusion*. IEEE, 2012.

R. Laxhammar and G. Falkman. Sequential Conformal Anomaly Detection in trajectories based on Hausdorff distance. In *14th International Conference on Information Fusion, Fusion 2011*. IEEE, 2011.

R. Laxhammar and G. Falkman. Online Detection of Anomalous Sub-trajectories: A Sliding Window Approach Based on Conformal Anomaly Detection and Local Outlier Factor. In *8th International Workshop on Artificial Intelligence Applications and Innovations, AIAI 2012*. Springer New York LLC, 2012.

R. Laxhammar and G. Falkman. Inductive conformal anomaly detection for sequential detection of anomalous sub-trajectories. *Annals of Mathematics and Artificial Intelligence*, 74(1-2):67–94, 6 2015. ISSN 10122443. doi: 10.1007/s10472-013-9381-7.

P. R. Lei. A framework for anomaly detection in maritime trajectory behavior. *Knowledge and Information Systems*, 47(1):189–214, 4 2016. ISSN 02193116. doi: 10.1007/s10115-015-0845-4.

M. Liang, Y. Zhan, and R. W. Liu. MVFFNet: Multi-view feature fusion network for imbalanced ship classification. *Pattern Recognition Letters*, 151:26–32, 2021. ISSN 0167-8655. doi: https://doi.org/10.1016/j.patrec.2021.07.024. URL `https://www.sciencedirect.com/science/article/pii/S0167865521002737`.

B. Liu, E. N. de Souza, C. Hilliard, and S. Matwin. Ship movement anomaly detection using specialized distance measures. In *2015 18th International Conference on Information Fusion (Fusion)*. IEEE, 2015a.

B. Liu, E. N. De Souza, S. Matwin, and M. Sydow. Knowledge-based clustering of ship trajectories using density-based approach. *Proceedings - 2014 IEEE International Conference on Big Data, IEEE Big Data 2014*, pages 603–608, 1 2015b. doi: 10.1109/BIGDATA.2014.7004281.

H. . Liu, Y. . Liu, B. . Li, Z. Qi, J. Rizvi, H. Liu, Y. Liu, B. Li, and Z. Qi. Ship Abnormal Behavior Detection Method Based on Optimized GRU Network. *Journal of Marine Science and Engineering 2022, Vol. 10, Page 249*, 10(2):249, 2 2022a. ISSN 2077-1312. doi: 10.3390/JMSE10020249. URL `https://www.mdpi.com/2077-1312/10/2/249/htmhttps://www.mdpi.com/2077-1312/10/2/249`.

H. Liu, Y. Liu, and Z. Zong. Research on Ship Abnormal Behavior Detection Method Based on Graph Neural Network. *2022 IEEE International Conference on Mechatronics and Automation, ICMA 2022*, pages 834–838, 2022b. doi: 10.1109/ICMA54519.2022.9856198.

R. W. Liu, M. Liang, J. Nie, X. Deng, Z. Xiong, J. Kang, H. Yang, and Y. Zhang. Intelligent Data-Driven Vessel Trajectory Prediction in Marine Transportation Cyber-Physical System. In *2021 IEEE International Conferences on Internet of Things (iThings) and IEEE Green Computing & Communications (GreenCom) and IEEE Cyber, Physical & Social Computing (CPSCom) and IEEE Smart Data (SmartData) and IEEE Congress on Cybermatics (Cybermatics)*, pages 314–321, 2021. doi: 10.1109/iThings-GreenCom-CPSCom-SmartData-Cybermatics53846.2021.00058.

R. W. Liu, M. Liang, J. Nie, W. Y. B. Lim, Y. Zhang, and M. Guizani. Deep Learning-Powered Vessel Trajectory Prediction for Improving Smart Traffic Services in Maritime Internet of Things. *IEEE Transactions on Network Science and Engineering*, 2022c. ISSN 23274697. doi: 10.1109/TNSE.2022.3140529.

R. W. Liu, M. Liang, J. Nie, Y. Yuan, Z. Xiong, H. Yu, and N. Guizani. STMGCN: Mobile Edge Computing-Empowered Vessel Trajectory Prediction Using Spatio-Temporal Multi-Graph Convolutional Network. *IEEE Transactions on Industrial Informatics*, 2022d. ISSN 19410050. doi: 10.1109/TII.2022.3165886.

M. Ljungqvist. Confirmed sabotage at Nord Stream - In Swedish. URL `https://www.aklagare.se/nyheter-press/pressmeddelanden/2022/november/bekraftat-sabotage-vid-nord-stream/`.

B. Lu, R. Lin, and H. Zou. A Novel CNN-LSTM Method for Ship Trajectory Prediction. In *2021 IEEE 23rd Int Conf on High Performance Computing & Communications; 7th Int Conf on Data Science & Systems; 19th Int Conf on Smart City; 7th Int Conf on Dependability in Sensor, Cloud & Big Data Systems & Application (HPCC/DSS/SmartCity/DependSys)*, pages 2431–2436, 2021. doi: 10.1109/HPCC-DSS-SmartCity-DependSys53884.2021.00366.

P. Mandalis, E. Chondrodima, Y. Kontoulis, N. Pelekis, and Y. Theodoridis. Machine Learning Models for Vessel Traffic Flow Forecasting: An Experimental Comparison. pages 431–436, 8 2022. ISSN 15516245. doi: 10.1109/MDM55031.2022.00094.

T. Mantecon, D. Casals, J. J. Navarro-Corcuera, C. R. Del-Blanco, and F. Jaureguizar. Deep learning to enhance maritime situation awareness. In *Proceedings International Radar Symposium*, volume 2019-June, 2019. ISBN 9783736998605. doi: 10.23919/IRS.2019.8768142.

MarineTraffic. MarineTraffic - A day in numbers, 2016. URL `https://www.marinetraffic.com/blog/a-day-in-numbers/`.

S. Mascaro, A. Nicholson, and K. Korb. Anomaly detection in vessel tracks using Bayesian networks. *International Journal of Approximate Reasoning*, 55(1):55, 2014. doi: 10.1016/j.ijar.2013.03.012.

E. Mathieu, T. Rainforth, N. Siddharth, and Y. W. Teh. Disentangling Disentanglement in Variational Autoencoders. *36th International Conference on Machine Learning, ICML 2019*, 2019-June:7744–7754, 12 2018. URL https://arxiv.org/abs/1812.02833v3.

A. K. Mazur, A. K. Wåhlin, and A. Krężel. An object-based SAR image iceberg detection algorithm applied to the Amundsen Sea. *Remote Sensing of Environment*, 189:67–83, 2 2017. ISSN 0034-4257. doi: 10.1016/J.RSE.2016.11.013.

MI News Network. 'Dark Ships' Spotted In The Nord Stream Mystery, 11 2022. URL https://www.marineinsight.com/shipping-news/dark-ships-spotted-in-the-nord-stream-mystery/.

L. M. Millefiori, P. Braca, K. Bryan, and P. Willett. Modeling vessel kinematics using a stochastic mean-reverting process for long-term prediction. *IEEE Transactions on Aerospace and Electronic Systems*, 52(5):2313–2330, 10 2016. ISSN 00189251. doi: 10.1109/TAES.2016.150596.

B. Murray and L. P. Perera. A dual linear autoencoder approach for vessel trajectory prediction using historical AIS data. *Ocean Engineering*, 209, 8 2020. ISSN 00298018. doi: 10.1016/j.oceaneng.2020.107478.

B. Murray and L. P. Perera. An AIS-based deep learning framework for regional ship behavior prediction. *Reliability Engineering & System Safety*, 215:107819, 11 2021. ISSN 0951-8320. doi: 10.1016/J.RESS.2021.107819.

H. Namgung and J. S. Kim. Regional Collision Risk Prediction System at a Collision Area Considering Spatial Pattern. *Journal of Marine Science and Engineering 2021, Vol. 9, Page 1365*, 9(12):1365, 12 2021. ISSN 2077-1312. doi: 10.3390/JMSE9121365. URL https://www.mdpi.com/2077-1312/9/12/1365/htmhttps://www.mdpi.com/2077-1312/9/12/1365.

M. Nanni and D. Pedreschi. Time-focused clustering of trajectories of moving objects. *J Intell Inf Syst*, 27:267–289, 2006. doi: 10.1007/s10844-006-9953-7. URL http://www-kdd.isti.cnr.it.

J. Neves, R. Maia, V. Conceicao, and M. M. Marques. Behaviour Analysis and Anomaly Detection Algorithms for the Maritime Integrated Surveillance Awareness. In *2019 IEEE International Underwater Technology Symposium, UT 2019 - Proceedings*. Institute of Electrical and Electronics Engineers Inc., 4 2019. ISBN 9781538641880. doi: 10.1109/UT.2019.8734404.

D. Nguyen and R. Fablet. TrAISformer-A generative transformer for AIS trajectory prediction.

D. Nguyen, R. Vadaine, G. Hajduch, R. Garello, and R. Fablet. A multi-task deep learning architecture for maritime surveillance using AIS data streams. In *Proceedings - 2018 IEEE 5th International Conference on Data Science and Advanced Analytics, DSAA 2018*, 2019. ISBN 9781538650905. doi: 10.1109/DSAA.2018.00044.

D. Nguyen, M. Simonin, G. Hajduch, R. Vadaine, C. Tedeschi, and R. Fablet. Detection of Abnormal Vessel Behaviours from AIS data using GeoTrackNet: from the Laboratory to the Ocean. *Proceedings - IEEE International Conference on Mobile Data Management*, 2020-June:264–268, 8 2020. doi: 10.1109/MDM48529.2020.00061. URL `http://arxiv.org/abs/2008.05443`.

D. Nguyen, R. Vadaine, G. Hajduch, R. Garello, and R. Fablet. GeoTrackNet-A Maritime Anomaly Detector using Probabilistic Neural Network Representation of AIS Tracks and A Contrario Detection. *IEEE Transactions on Intelligent Transportation Systems*, 2021. doi: 10.1109/TITS.2021.3055614.

K. V. Olesen, A. N. Christensen, and L. K. H. Clemmensen. Towards latent representation interpretability for maritime anomaly detections. *Preprint*, 2021.

K. V. Olesen, A. Boubekki, M. Kampffmeyer, R. Jenssen, A. N. Christensen, S. Hørlück, and L. K. H. Clemmensen. A two-step clustering method for maritime behaviour identification. *Manuscript submitted for publication*, 2022a.

K. V. Olesen, A. N. Christensen, S. Hørlück, and L. K. H. Clemmensen. AIS Trajectories from Danish Waters for Abnormal Behavior Detection, 2022b. URL `https://figshare.com/s/0012239be1c55e988a32`.

K. V. Olesen, A. N. Christensen, S. Hørlück, and L. K. H. Clemmensen. A Review of Current Deep Learning Techniques for Maritime Abnormality Detection and Directions for Future Progress. *Manuscript submitted for publication*, 2022c.

K. V. Olesen, A. N. Christensen, S. Hørlück, and L. K. H. Clemmensen. Detecting Abnormal Maritime Trajectories using Ensembles and Transfer Learning. *Preprint*, 2022d.

E. Osekowska and B. Carlsson. Learning Maritime Traffic Rules Using Potential Fields. doi: 10.1007/978-3-319-24264-4. URL `http://www.bth.se/com/`.

E. Osekowska, H. Johnson, and B. Carlsson. Grid size optimization for potential field based maritime anomaly detection. In *Transportation Research Procedia*, volume 3, pages 720–729. Elsevier, 2014. doi: 10.1016/j.trpro.2014.10.051.

E. Osekowska, H. Johnson, and B. Carlsson. Maritime vessel traffic modeling in the context of concept drift. In *Transportation Research Procedia*, volume 25, pages 1457–1476. Elsevier B.V., 2017. doi: 10.1016/j.trpro.2017.05.173.

G. Pallotta and A.-L. Jousselme. Data-driven detection and context-based classification of maritime anomalies. In *2015 18th International Conference on Information Fusion (Fusion)*. IEEE, 2015.

G. Pallotta, M. Vespe, and K. Bryan. Traffic knowledge discovery from AIS data. In *Proceedings of the 16th International Conference on Information Fusion*. IEEE, 2013a.

G. Pallotta, M. Vespe, and K. Bryan. Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction. *Entropy*, 15(12):2218–2245, 6 2013b. doi: 10.3390/e15062218.

G. Pallotta, S. Horn, P. Braca, and K. Bryan. Context-enhanced vessel prediction based on Ornstein-Uhlenbeck processes using historical AIS traffic patterns: Real-world experimental results. In *17th International Conference on Information Fusion (FUSION)*. IEEE, 2014.

G. Pang, C. Shen, L. Cao, and A. V. D. Hengel. Deep Learning for Anomaly Detection: A Review. *ACM Comput. Surv.*, 54(2), 3 2021. ISSN 0360-0300. doi: 10.1145/3439950. URL https://doi.org/10.1145/3439950.

J. Park, J. S. Jeong, and Y. S. Park. Ship Trajectory Prediction Based on Bi-LSTM Using Spectral-Clustered AIS Data. *Journal of Marine Science and Engineering*, 9(1037):1037, 9 2021. ISSN 2077-1312. doi: 10.3390/JMSE9091037. URL https://doaj.org/article/9ecd952a7bcb41bc8f94851c9d20e065.

E. Protopapadakis, A. Voulodimos, A. Doulamis, N. Doulamis, D. Dres, and M. Bimpas. Stacked Autoencoders for Outlier Detection in Over-the-Horizon Radar Signals. *Computational Intelligence and Neuroscience*, 2017, 2017. ISSN 16875273. doi: 10.1155/2017/5891417.

L. Qian, Y. Zheng, L. Li, Y. Ma, C. Zhou, and D. Zhang. A New Method of Inland Water Ship Trajectory Prediction Based on Long Short-Term Memory Network Optimized by Genetic Algorithm. *Applied Sciences 2022, Vol. 12, Page 4073*, 12(8):4073, 4 2022. ISSN 2076-3417. doi: 10.3390/APP12084073. URL https://www.mdpi.com/2076-3417/12/8/4073/htmhttps://www.mdpi.com/2076-3417/12/8/4073.

M. Quadrana, A. Karatzoglou, B. Hidasi, and P. Cremonesi. Personalizing Session-Based Recommendations with Hierarchical Recurrent Neural Networks. In *Proceedings of the Eleventh ACM Conference on Recommender Systems*, RecSys '17, pages 130–137, New York, NY, USA, 2017. Association for Computing Machinery. ISBN 9781450346528. doi: 10.1145/3109859.3109896. URL https://doi-org.proxy.findit.cvt.dk/10.1145/3109859.3109896.

A. N. Radon, K. Wang, U. Glasser, H. Wehn, and A. Westwell-Roper. Contextual verification for false alarm reduction in maritime anomaly detection. In *Proceedings - 2015 IEEE International Conference on Big Data, IEEE Big Data 2015*, pages 1123–1133. Institute of Electrical and Electronics Engineers Inc., 12 2015. ISBN 9781479999255. doi: 10.1109/BigData.2015.7363866.

D. J. Rezende, S. Mohamed, and D. Wierstra. Stochastic Backpropagation and Approximate Inference in Deep Generative Models. *31st International Conference on Machine Learning, ICML 2014*, 4:3057–3070, 1 2014. URL `https://arxiv.org/abs/1401.4082v3`.

M. Riveiro, G. Falkman, and T. Ziemke. Visual analytics for the detection of anomalous maritime behavior. *Proceedings of the International Conference on Information Visualisation*, pages 273–279, 2008. ISSN 10939547. doi: 10.1109/IV.2008.25.

M. Riveiro, G. Pallotta, and M. Vespe. Maritime anomaly detection: A review. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, 8 (5):e1266, 9 2018. doi: 10.1002/widm.1266.

C. Rizogiannis and S. C. A. Thomopoulos. A fuzzy inference system for ship-ship collision alert generation. In *Proceedings of Spie - the International Society for Optical Engineering*, volume 11018, page 10. SPIE-Intl Soc Optical Eng, 5 2019. ISBN 9781510627017. doi: 10.1117/12.2519475.

P. J. Rousseeuw. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20:53–65, 1987. ISSN 0377-0427. doi: https://doi.org/10.1016/0377-0427(87)90125-7. URL `https://www.sciencedirect.com/science/article/pii/0377042787901257`.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. Learning representations by back-propagating errors. *Nature 1986 323:6088*, 323(6088):533–536, 1986. ISSN 1476-4687. doi: 10.1038/323533A0. URL `https://www-nature-com.proxy.findit.dtu.dk/articles/323533a0`.

J. Salerno, M. Hinman, and D. Boulware. Building a framework for situation awareness. 10 2022.

W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K.-R. Muller. *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*. Springer Publishing Company, Incorporated, 1st edition, 2019. ISBN 978-3-030-28954-6. doi: https://doi-org.proxy.findit.cvt.dk/10.1007/978-3-030-28954-6.

V. Satopaa, J. Albrecht, D. Irwin, and B. Raghavan. Finding a "Kneedle" in a Haystack: Detecting Knee Points in System Behavior. In *31st International Conference on Distributed Computing Systems Workshops*, 2011.

S. Semeniuta, A. Severyn, and E. Barth. A Hybrid Convolutional Variational Autoencoder for Text Generation. *EMNLP 2017 - Conference on Empirical Methods in Natural Language Processing, Proceedings*, pages 627–637, 2 2017. doi: 10.18653/v1/d17-1066. URL `https://arxiv.org/abs/1702.02390v1`.

A. Sfyridis, T. Cheng, and M. Vespe. Detecting vessels carrying migrants using machine learning. In *ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences*, volume 4, pages 53–60. Copernicus GmbH, 10 2013. doi: 10.5194/isprs-annals-IV-4-W2-53-2017.

M. Shamos and F. Preparata. Computational Geometry An Introduction. In F. Schneider and D. Gries, editors, *Computational Geometry An Introduction*, chapter 5, page 223. Springer, New York, 1 edition, 1985.

A. Sidibé and G. Shu. Study of automatic anomalous behaviour detection techniques for maritime vessels. *Journal of Navigation*, 70(4):847–858, 7 2017. doi: 10.1017/S0373463317000066.

S. K. Singh and F. Heymann. Machine Learning-Assisted Anomaly Detection in Maritime Navigation Using AIS Data. 2020a.

S. K. Singh and F. Heymann. On the effectiveness of AI-assisted anomaly detection methods in maritime navigation. *Proceedings of 2020 23rd International Conference on Information Fusion, FUSION 2020*, 7 2020b. doi: 10.23919/FUSION45008.2020.9190533.

S. K. Singh, J. S. Fowdur, J. Gawlikowski, and D. Medina. Leveraging Graph and Deep Learning Uncertainties to Detect Anomalous Trajectories. *IEEE Transactions on Intelligent Transportation Systems*, 7 2021. ISSN 15580016. doi: 10.48550/arxiv.2107.01557. URL `https://arxiv.org/abs/2107.01557v2`.

J. Smith, I. Nouretdinov, R. Craddock, C. Offer, and A. Gammerman. Conformal Anomaly Detection of Trajectories with a Multi-class Hierarchy. doi: 10.1007/978-3-319-17091-6.

J. Smith, I. Nouretdinov, R. Craddock, C. Offer, and A. Gammerman. Anomaly Detection of Trajectories with Kernel Density Estimation by Conformal Prediction. *Ifip Advances in Information and Communication Technology*, 437, 2014a. doi: 10.1007/978-3-662-44722-2{_}29.

M. Smith, S. Reece, S. Roberts, and I. Rezek. Online maritime abnormality detection using Gaussian Processes and extreme value theory. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 645–654, 2012. ISBN 9780769549057. doi: 10.1109/ICDM.2012.137.

M. Smith, S. Reece, and S. Roberts. Maritime abnormality detection using Gaussian processes. *Knowl Inf Syst*, 38:717–741, 2014b. doi: 10.1007/s10115-013-0685-z.

L. Snidaro, I. Visentini, and K. Bryan. Fusing uncertain knowledge and evidence for maritime situational awareness via Markov Logic Networks. *Information Fusion*, 21(1):159–172, 2015. ISSN 15662535. doi: 10.1016/j.inffus.2013.03.004.

C. K. Sønderby, T. Raiko, L. Maaløe, S. K. Sønderby, and O. Winther. Ladder Variational Autoencoders. *Advances in Neural Information Processing Systems*, 0:3745–3753, 2 2016. ISSN 10495258. URL `https://arxiv.org/abs/1602.02282v3`.

A. Sordoni, Y. Bengio, H. Vahabi, C. Lioma, J. Grue Simonsen, and J.-Y. Nie. A Hierarchical Recurrent Encoder-Decoder for Generative Context-Aware Query Suggestion. In *Proceedings of the 24th ACM International on Conference on Information and Knowledge Management*, CIKM '15, pages 553–562, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450337946. doi: 10.1145/2806416.2806493. URL `https://doi-org.proxy.findit.cvt.dk/10.1145/2806416.2806493`.

K. A. Sørensen, P. Heiselberg, and H. Heiselberg. Probabilistic Maritime Trajectory Prediction in Complex Scenarios Using Deep Learning. *Sensors 2022, Vol. 22, Page 2058*, 22(5):2058, 3 2022. ISSN 1424-8220. doi: 10.3390/S22052058. URL `https://www.mdpi.com/1424-8220/22/5/2058/htmhttps://www.mdpi.com/1424-8220/22/5/2058`.

G. Spadon, M. D. Ferreira, A. Soares, and S. Matwin. Unfolding AIS transmission behavior for vessel movement modeling on noisy data leveraging machine learning. *IEEE Access*, 2022. ISSN 21693536. doi: 10.1109/ACCESS.2022.3197215.

N. Srivastava, E. Mansimov, and R. Salakhutdinov. Unsupervised Learning of Video Representations using LSTMs. *32nd International Conference on Machine Learning, ICML 2015*, 1:843–852, 2 2015. doi: 10.48550/arxiv.1502.04681. URL `https://arxiv.org/abs/1502.04681v3`.

Statista Research Department. Total damages caused by recreational boating accidents in the U.S. from 2001 to 2020, 8 2021. URL `https://www.statista.com/statistics/240641/recreational-boating-accidents-in-the-us-total-damages/`.

F. Sun, Y. Deng, F. Deng, Q. Zhu, and H. Chu. Unsupervised maritime traffic pattern extraction from spatio-temporal data. In *Proceedings - International Conference on Natural Computation*, volume 2016-January, pages 1218–1223.

IEEE Computer Society, 1 2016. ISBN 9781467376792. doi: 10.1109/ICNC.2015.7378165.

Y. Suo, W. Chen, C. Claramunt, and S. Yang. A ship trajectory prediction framework based on a recurrent neural network. *Sensors (Switzerland)*, 20 (18):1–21, 9 2020. ISSN 14248220. doi: 10.3390/s20185133.

I. Sutskever Google, O. Vinyals Google, and Q. V. Le Google. Sequence to Sequence Learning with Neural Networks. *Advances in Neural Information Processing Systems*, 27, 2014.

S. C. Thomopoulos, C. Rizogannis, K. G. Thanos, K. Dimitros, K. Panou, and D. Zacharakis. OCULUS Sea™ Forensics: An Anomaly Detection Toolbox for Maritime Surveillance. In *Lecture Notes in Business Information Processing*, volume 373 LNBIP, pages 485–495. Springer, 2019. ISBN 9783030366902. doi: 10.1007/978-3-030-36691-9{_}41.

M. Uney, L. M. Millefiori, and P. Braca. Data Driven Vessel Trajectory Forecasting Using Stochastic Generative Models. In *ICASSP, IEEE International Conference on Acoustics, Speech and Signal Processing - Proceedings*, volume 2019-May, pages 8459–8463. Institute of Electrical and Electronics Engineers Inc., 5 2019. ISBN 9781479981311. doi: 10.1109/ICASSP.2019.8683444.

I. Varlamis, K. Tserpes, M. Etemad, A. Soares, and S. Matwin. A network abstraction of multi-vessel trajectory data for detecting anomalies. In *CEUR Workshop Proceedings*, volume 2322, 2019.

J. Venskus, M. Kurmis, A. Andziulis, Z. Lukosius, M. Voznak, and D. Bykovas. Self-learning adaptive algorithm for maritime traffic abnormal movement detection based on virtual pheromone method. In *Proceedings of the 2015 International Symposium on Performance Evaluation of Computer and Telecommunication Systems, SPECTS 2015 - Part of SummerSim 2015 Multiconference*. Institute of Electrical and Electronics Engineers Inc., 9 2015. ISBN 9781510810600. doi: 10.1109/SPECTS.2015.7285281.

J. Venskus, P. Treigys, J. Bernatavičiene, V. Medvedev, M. Voznak, M. Kurmis, and V. Bulbenkiene. Integration of a Self-Organizing Map and a Virtual Pheromone for Real-Time Abnormal Movement Detection in Marine Traffic. *Informatica (Netherlands)*, 28(2):359–374, 2017. ISSN 08684952. doi: 10.15388/Informatica.2017.133.

J. Venskus, P. Treigys, J. Bernatavičienė, G. Tamulevičius, and V. Medvedev. Real-time maritime traffic anomaly detection based on sensors and history data embedding. *Sensors (Switzerland)*, 19(17), 9 2019. ISSN 14248220. doi: 10.3390/s19173782.

G. Vivone, L. M. Millefiori, P. Braca, and P. Willett. Performance assessment of vessel dynamic models for long-term prediction using heterogeneous data. *IEEE Transactions on Geoscience and Remote Sensing*, 55(11):6533–6546, 8 2017. ISSN 01962892. doi: 10.1109/TGRS.2017.2729622.

C. Wang, H. Ren, and H. Li. Vessel trajectory prediction based on AIS data and bidirectional GRU. In *Proceedings - 2020 International Conference on Computer Vision, Image and Deep Learning, CVIDL 2020*, pages 260–264. Institute of Electrical and Electronics Engineers Inc., 7 2020. ISBN 9781728194813. doi: 10.1109/CVIDL51233.2020.00-89.

L. Wang, P. Chen, L. Chen, and J. Mou. Ship AIS Trajectory Clustering: An HDBSCAN-Based Approach. *Journal of Marine Science and Engineering 2021, Vol. 9, Page 566*, 9(6):566, 5 2021. ISSN 2077-1312. doi: 10.3390/JMSE9060566. URL `https://www.mdpi.com/2077-1312/9/6/566/htmhttps://www.mdpi.com/2077-1312/9/6/566`.

Y. Wang. Application of neural network in abnormal AIS data identification. *Proceedings of 2020 IEEE International Conference on Artificial Intelligence and Computer Applications, ICAICA 2020*, pages 173–179, 6 2020. doi: 10.1109/ICAICA50127.2020.9182703.

Y. Wen, Z. Sui, C. Zhou, C. Xiao, Q. Chen, D. Han, and Y. Zhang. Automatic ship route design between two ports: A data-driven method. *Applied Ocean Research*, 96, 3 2020. ISSN 01411187. doi: 10.1016/j.apor.2019.102049.

World Food Program. War in Ukraine drives global food crisis, 6 2022. URL `https://www.wfp.org/publications/war-ukraine-drives-global-food-crisis`.

Z. Xia and S. Gao. Analysis of Vessel Anomalous Behavior Based on Bayesian Recurrent Neural Network. In *2020 IEEE 5th International Conference on Cloud Computing and Big Data Analytics, ICCCBDA 2020*, pages 393–397. Institute of Electrical and Electronics Engineers Inc., 4 2020. ISBN 9781728160245. doi: 10.1109/ICCCBDA49378.2020.9095567.

T. Xu, X. Liu, and X. Yang. Ship trajectory online prediction based on BP neural network algorithm. *Proceedings - 2011 International Conference of Information Technology, Computer Engineering and Management Sciences, ICM 2011*, 1:103–106, 2011. doi: 10.1109/ICM.2011.288.

R. Yan and S. Wang. Study of Data-Driven Methods for Vessel Anomaly Detection Based on AIS Data. In *Smart Innovation, Systems and Technologies*, volume 149, pages 29–37. Springer Science and Business Media Deutschland GmbH, 2019. ISBN 9789811386824. doi: 10.1007/978-981-13-8683-1{_}4.

C. H. Yang, C. H. Wu, J. C. Shao, Y. C. Wang, and C. M. Hsieh. AIS-Based Intelligent Vessel Trajectory Prediction Using Bi-LSTM. *IEEE Access*, 10: 24302–24315, 2022a. ISSN 21693536. doi: 10.1109/ACCESS.2022.3154812.

J. Yang, Y. Liu, L. Ma, and C. Ji. Maritime traffic flow clustering analysis by density based trajectory clustering with noise. *Ocean Engineering*, 249: 111001, 4 2022b. ISSN 0029-8018. doi: 10.1016/J.OCEANENG.2022.111001.

D. Yao, C. Zhang, Z. Zhu, J. Huang, and J. Bi. Trajectory clustering via deep representation learning. *Proceedings of the International Joint Conference on Neural Networks*, 2017-May:3880–3887, 6 2017. doi: 10.1109/IJCNN.2017.7966345.

H. Yao, F. Wu, J. Ke, X. Tang, Y. Jia, S. Lu, P. Gong, J. Ye, and Z. Li. Deep Multi-View Spatial-Temporal Network for Taxi Demand Prediction. 2 2018. URL `http://arxiv.org/abs/1802.08714`.

J. Y. Yu, M. O. Sghaier, and Z. Grabowiecka. Deep learning approaches for AIS data association in the context of maritime domain awareness. *Proceedings of 2020 23rd International Conference on Information Fusion, FUSION 2020*, 7 2020. doi: 10.23919/FUSION45008.2020.9190283.

H. Yuan, X. Zhu, Z. Hu, and C. Zhang. Deep multi-view residual attention network for crowd flows prediction. *Neurocomputing*, 404:198–212, 9 2020. ISSN 18728286. doi: 10.1016/j.neucom.2020.04.124.

Y. Zang, L. Deng, S. Sun, Y. Ai, D. Geng, and X. Kang. Abnormal Behavior Detection of Vessels Based on Deep Learning Algorithm: Case Study. In *2021 6th International Conference on Transportation Information and Safety (ICTIS)*, pages 205–212, 2021. doi: 10.1109/ICTIS54573.2021.9798408.

B. Zhang, H. Ren, P. Wang, and D. Wang. Research Progress on Ship Anomaly Detection Based on Big Data. *Proceedings of the IEEE International Conference on Software Engineering and Service Sciences, ICSESS*, 2020-October: 316–320, 10 2020a. ISSN 23270594. doi: 10.1109/ICSESS49938.2020.9237642.

J. Zhang, Y. Zheng, J. Sun, and D. Qi. Flow Prediction in Spatio-Temporal Networks Based on Multitask Deep Learning. *IEEE Transactions on Knowledge and Data Engineering*, 32(3):468–478, 3 2020b. ISSN 15582191. doi: 10.1109/TKDE.2019.2891537.

R. Zhang, P. Xie, C. Wang, G. Liu, and S. Wan. Classifying transportation mode and speed from trajectory data via deep multi-Scale learning. *Computer Networks*, 162:106861, 10 2019. ISSN 13891286. doi: 10.1016/j.comnet.2019.106861.

S. Zhang, L. Wang, M. Zhu, S. Chen, H. Zhang, and Z. Zeng. A Bi-directional LSTM Ship Trajectory Prediction Method based on Attention Mechanism. *IEEE Advanced Information Technology, Electronic and Automation Control Conference (IAEAC)*, pages 1987–1993, 2021a. ISSN 26896621. doi: 10.1109/IAEAC50856.2021.9391059.

T. Zhang, S. Zhao, B. Cheng, and J. Chen. ATeDLW: Intelligent Detection of Abnormal Trajectory in Ship Data Service System. In *2021 IEEE International Conference on Services Computing (SCC)*, pages 401–406, 2021b. doi: 10.1109/SCC53864.2021.00057.

Z. Zhang, G. Ni, and Y. Xu. Ship Trajectory Prediction based on LSTM Neural Network. *Proceedings of 2020 IEEE 5th Information Technology and Mechatronics Engineering Conference, ITOEC 2020*, pages 1356–1364, 6 2020c. doi: 10.1109/ITOEC49072.2020.9141702.

B. Zhao, X. Li, and X. Lu. Hierarchical Recurrent Neural Network for Video Summarization. *CoRR*, abs/1904.12251, 2019. URL `http://arxiv.org/abs/1904.12251`.

L. Zhao and G. Shi. A trajectory clustering method based on Douglas-Peucker compression and density for marine traffic pattern recognition. *Ocean Engineering*, 172:456–467, 1 2019a. ISSN 0029-8018. doi: 10.1016/J.OCEANENG.2018.12.019.

L. Zhao and G. Shi. Maritime Anomaly Detection using Density-based Clustering and Recurrent Neural Network. *Journal of Navigation*, 72(4):894–916, 7 2019b. ISSN 14697785. doi: 10.1017/S0373463319000031.

R. Zhen, Y. Jin, Q. Hu, Z. Shao, and N. Nikitakos. Maritime Anomaly Detection within Coastal Waters Based on Vessel Trajectory Clustering and Naïve Bayes Classifier. *Journal of Navigation*, 70(3):648–670, 5 2017. ISSN 14697785. doi: 10.1017/S0373463316000850.

H. Zhou, Y. Chen, and S. Zhang. Ship Trajectory Prediction Based on BP Neural Network. *Article in Journal on Artificial Intelligence*, 1(1):29–36, 2019. doi: 10.32604/jai.2019.05939. URL `www.techscience.com/jai`.

X. Zhou, Z. Liu, F. Wang, Y. Xie, and X. Zhang. Using deep learning to forecast maritime vessel flows. *Sensors (Switzerland)*, 20(6), 3 2020. ISSN 14248220. doi: 10.3390/s20061761.

Y. Zhu, M. R. Min, A. Kadav, and H. P. Graf. S3VAE: Self-Supervised Sequential VAE for Representation Disentanglement and Data Generation. *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 6537–6546, 5 2020. doi: 10.1109/CVPR42600.2020.00657. URL `https://arxiv.org/abs/2005.11437v1`.

D. Zissis, E. K. Xidias, and D. Lekkas. A cloud based architecture capable of perceiving and predicting multiple vessel behaviour. *Applied Soft Computing Journal*, 35:652–661, 7 2015. ISSN 15684946. doi: 10.1016/j.asoc.2015.07.002.

Soefartsstyrelsen. Historical AIS data. URL `https://dma.dk/safety-at-sea/navigational-information/ais-data`.