Proceedings of the 33rd European Safety and Reliability Conference (ESREL 2023)
Edited by Mário P. Brito, Terje Aven, Piero Baraldi, Marko Čepin and Enrico Zio
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-18-8071-1_P403-cd

Autoencoder-Based Anomaly Detection for Safe Autonomous Ship Operations

Brian Murray
SINTEF Ocean, Norway. E-mail: brian.murray@sintef.no

Pauline Røstum Bellingmo
SINTEF Ocean, Norway. E-mail: pauline.bellingmo@sintef.no

Thorbjørn Tønnesen Lied
Kongsberg Norcontrol, Norway. E-mail: thorbjorn.tonnesen.lied@knc.kongsberg.com

Marianne Hagaseth
SINTEF Ocean, Norway. E-mail: marianne.hagaseth@sintef.no

The development of autonomous ships is advancing, but ensuring their safe operation remains a challenge. To aid safe operations, autonomous ships are expected to be monitored by humans in a remote operation center. A key challenge is ensuring that human operators remain alert and ready to take control of the system when necessary. Maritime traffic poses a potential hazard to autonomous vessels, and systems to aid the operator in identifying abnormal ship behavior in time should be in place. This study develops deep learning models that automatically detect anomalous ship behavior to aid human operators. A case study related to the remote operation center in Horten, Norway is conducted, where four various autoencoder architectures have been trained on historical Automatic Identification System data to detect maritime traffic anomalies in the Oslo fjord. The models are trained in an unsupervised manner, such that they are able to automatically identify anomalies, without the need for manual labelling. The results indicate that a recurrent autoencoder is the most promising architecture for decision support of remote operators, as it is able to identify a variety of anomalies, with fewer false positives.

Keywords: Autonomous ships, Remote operation centre, Machine learning, AI, Anomaly detection, Autoencoder, Maritime traffic, Maritime safety, Decision support

1. Introduction

Significant progress has been made towards realising maritime autonomous surface ships in recent years, with many full-scale projects currently underway (e.g. ASKO and Seafar vessels). Ensuring that the level of risk associated with their operation is at a level equal to or less than that of conventional vessels, however, remains a challenge. Autonomous ship systems will likely be equipped with advanced situational awareness systems that leverage Artificial Intelligence (AI) based technologies to facilitate object detection and tracking. Such systems are envisioned to handle collision avoidance situations in accordance with relevant regulations. As current regulations stand, however, it is unlikely that the automation system can handle all possible ship encounter situations, especially in multi-vessel encounters or if a ship behaves erratically. Current autonomous ship systems are, therefore, designed to incorporate human operators in Remote Operation Centers (ROC) that can intervene in such situations. However, it is challenging for human operators to always stay sufficiently alert such that they can determine when they need to take control of the system. This is assumed to be the cause of a fatal car accident in 2018 where the driver did not pay enough attention when using automated steering functions (National Transportation Safety Board, 2020). As such, the automation system must be capable of determining when its limits are approaching to facilitate a safe and timely handover. Many situations that will require human intervention relate to surrounding maritime traffic. If the system is capable of identifying future situations that the automation system may not be

able to handle, it can alert human operators far in advance. Operators will then have more time to gain adequate situation awareness to handle the situation.

1.1. *Anomaly Detection for Remote Operation Centers*

Anomaly detection refers to the process of identifying patterns that deviate from the expected or normal behavior of a system. In the context of maritime traffic, there exists many different types of anomalies, e.g. route deviation or unexpected activities. Anomalous e.g., erratic, ship behavior is inherently unpredictable, and likely outside the operational envelope of the automation system responsible for collision avoidance. Identifying and alerting an ROC operator to such situations is, therefore, critical for the safety of an autonomous ship. However, automatically identifying anomalous behavior is not straight forward. Several studies have looked into different methods for detecting anomalies in maritime traffic. Wolsing et al. (2022) provides a review of anomaly detection approaches and divides them in the following categories; geometric, stochastic, and machine learning based. This study will use a machine learning (ML) based anomaly detection method, more specifically an autoencoder. ML-based anomaly detection can help to overcome some of the limitations of traditional methods, such as rule-based systems, which can struggle to adapt to changing conditions or detect subtle anomalies. ML algorithms can learn from vast amounts of historical data to identify patterns and anomalies that might be missed by human operators.

Some studies have investigated the application of autoencoders for anomaly detection in maritime traffic. Iltanen (2020) detects anomalies in AIS data by combining the reconstruction error from a recurrent autoencoder (RAE) with an outlier score produced by clustering the encodings of the autoencoder. However, the results showed that the outlier score produced by the clustering on the encodings did not improve the anomaly detection performance compared with simply using the RAE in most cases. Son et al. (2020) uses a convolutional autoencoder on images of vessel trajectories based on AIS to detect anomalies. However, the results showed a low identification rate for the anomaly detection method, partially due to the model not taking into account the speed, ship type, or origin and destination.

Hu et al. (2023) uses a variational recurrent autoencoder (VRAE) to find connections between each dimension of the trajectories and a graph variational autoencoder (GVAE) to find spatial similarities between trajectories. These two reconstruction probabilities (from VRAE and GVAE) are combined using a reinforcement learning method to create an anomaly detection method. Manually labelled data are used to train the anomaly detection algorithm, i.e. in a supervised manner. However, for an unlabelled data set, an unsupervised method is more suitable, as manually labelling large amount of data is time consuming, and not feasible in many cases. Furthermore, the approach does not generalize to other regions where labelled data are unavailable.

1.2. *Contribution*

This study aims to develop a method to discover anomalous ship behavior from AIS data in an unsupervised manner, i.e., without labelled data. In this manner, the approach can be applied to any given region. Deep learning based approaches, as mentioned in this section, show promise for facilitating anomaly detection functions. More specifically, autoencoder-based approaches provide a generic solution to discover the normalcy of data. As opposed to parametric approaches that are constrained to pre-defined distributions, deep autoencoders can learn the distribution of the data. This study will compare the performance of four different autoencoders used for anomaly detection, i.e., a simple Autoencoder (AE), a Variational Autoencoder (VAE), a Recurrent Autoencoder (RAE), and a Variational Recurrent Autoencoder (VRAE). Anomalies will be detected by detecting tracks with a high reconstruction error.

2. Methodology

In this study, deep learning is leveraged to discover anomalous ship behavior. An autoencoder-based approach is utilized, where the normalcy of

the traffic is learned by the model. This section outlines the steps to conduct data processing, as well as details on the investigated autoencoder architectures. Typical anomalies include positional, speed, and course anomalies, which will be investigated in this study.

2.1. *Data Pre-Processing*

To identify the normalcy of ship traffic in a given region, historical ship behavior can be evaluated via historical AIS data. For live traffic monitoring, live AIS streams can be input to the anomaly detection models for inference. AIS data include relevant kinematic information yielding insight into the behavior of a specific vessel. However, these data must be pre-processed prior to model training. In this study, the data are initially filtered to remove unrealistic speed values, as well as data points that intersect land. Subsequently, the data per vessel are aggregated to generate routes between two locations. Each route is further interpolated at one minute intervals.

2.1.1. *Route Clustering*

Once routes have been generated, they are clustered to represent unique origin-destination pairs. Unique origin and destination locations are discovered by applying the DBSCAN (Ester et al., 1996) clustering algorithm to all starting and stopping locations for each route. A cluster is then defined as a unique combination of origin and destination locations. By grouping historical routes in this manner, specific behavior can be discovered, as most vessels travelling between two locations will have similar trajectories. Applying an anomaly detection model to the specific behavior of such a cluster may yield enhanced performance, as it is likely able to identify specific anomalies to the cluster compared to training on all data for the region.

2.1.2. *Trajectory Windowing*

During inference, live AIS data will be monitored. It is, therefore, of interest to determine the duration during which an anomaly should be identified. For instance, a 5, 10, 20 or 30 minute trajectory segment could be considered anomalous. In this study, the past 10 minutes of ship behavior is monitored. As such, the data in each historical route are split using a sliding window technique with a window size of 10. By stepping one minute into the future along a route from start to end, 10 minute trajectory segments are generated for all routes in the relevant data set.

2.1.3. *Feature Scaling*

The following features from the historical AIS data are chosen for training of the anomaly detection models: Latitude, Longitude, Speed over Ground (SOG) and Course over Ground (COG). To make the input more conducive to deep learning models, the features are scaled between 0 and 1. Furthermore, the COG values are decomposed via their Sine and Cosine values, due to their circular nature.

2.2. *Autoencoder-Based Anomaly Detection*

Autoencoders are neural networks that aim to reconstruct their input. They are generally considered to be comprised of two parts; an encoder and a decoder. The encoder is responsible for generating a representation of the input data, and the decoder responsible for reconstructing the input from this representation. In general, there are two types of autoencoders; undercomplete and overcomplete autoencoders. An undercomplete autoencoder compresses the data to a dimensionality less than that of the input feature space. As such, the network learns to preserve as much mutual information between the input and latent (i.e. compressed) representation as possible. Overcomplete autoencoders have the opposite functionality, where the latent space has a dimensionality greater than the input space. Such architectures are often used for de-noising applications.

In this study, we will investigate undercomplete autoencoders. By forcing the network to compress information from the input, anomalous data often disappear, as they are not within the distribution of learned normal data. As such, when reconstructing the input from the encoded representation via the decoder, the error between the input and the reconstruction should be greater for anomalous data

than for normal data. Four autoencoder architectures are investigated; a standard AE, a VAE, an RAE and a VRAE. By evaluating the distribution of the reconstruction loss for the training data set, a threshold can be determined where any losses over the threshold are classified as anomalies.

2.2.1. *Standard Autoencoder*

The simplest deep autoencoder architecture is comprised of a Multi-Layer Perceptron (MLP) (Bourlard and Kamp, 1988), where the encoder and decoder are integrated in the same network. For an undercomplete autoencoder, the number of neurons per layer should decrease until the latent layer. They should then increase until the output layer which should have as many neurons as the input. For time series data, there will be multiple features per time step. The data set will, therefore, be a 3-dimensional tensor, with dimensions batch size, time series length and feature size. For a standard autoencoder, the data must be in the form of a 2-dimensional tensor. As such, the data are flattened along the time and feature dimensions. The network is trained using the reconstruction loss, i.e. the error between the input and output.

2.2.2. *Variational Autoencoder*

A variational autoencoder (Kingma and Welling, 2014) learns a latent distribution of the data via a probabilistic encoder $p_\theta(\mathbf{z}|\mathbf{x})$. The architecture is, therefore, generally better suited for data generation, where a latent variable $\mathbf{z}$ can be sampled from a prior distribution $p_\theta(\mathbf{z})$ and decoded to generate a new data point $\mathbf{x}^i$ via a conditional distribution $p_\theta(\mathbf{x}|\mathbf{z})$. Due to the intractability of the encoder, it must be approximated as $q_\phi(\mathbf{z}|\mathbf{x})$ which is further assumed to be normally distributed as $\mathcal{N}(\boldsymbol{\mu}_z, \boldsymbol{\sigma}_z^2\mathbf{I})$ with diagonal covariance. This approximation is facilitated via neural networks that estimate $\boldsymbol{\mu}_z$ and $\boldsymbol{\sigma}_z$. The network is then trained by maximising a variational lower bound that in effect minimizes the Kullback–Leibler divergence as well as minimize the reconstruction loss. For further details see Kingma and Welling (2014).

2.2.3. *Recurrent Autoencoder*

Recurrent autoencoders (Srivastava et al., 2015) leverage the power of recurrent neural networks (RNN) to process time series data. In this case, both the encoder and decoder networks are comprised of RNNs. Typically used architectures are the Long Short-Term Memory (LSTM) and Gated Recurrent Unit (GRU) architectures, as they are able to handle challenges related to vanishing gradients. In this study, the GRU is investigated as it requires fewer trainable parameters compared to the LSTM.

The encoder network processes a sequence, and compresses it into a single vector, i.e. the final hidden state. The decoder network then takes this hidden state as input, and reconstructs the input sequence solely based on the information stored in the vector. The hidden state therefore acts as a bottleneck for the network, compressing the input to a lower dimensionality.

2.2.4. *Variational Recurrent Autoencoder*

The variational recurrent autoencoder was introduced by Fabius and van Amersfoort (2015). The architecture integrates a recurrent autoencoder into the variational structure outlined in Sec. 2.2.2. A recurrent encoder network will output a hidden state, compressing the input. From this hidden state, $\boldsymbol{\mu}_z$ and $\boldsymbol{\sigma}_z$ will be estimated in the same manner as Sec. 2.2.2, with a recurrent decoder reconstructing the input from the sampled latent state, $\boldsymbol{z}$. The architecture should, therefore, be more conducive with generating new time series by sampling from the latent space.

3. Results and Discussion

3.1. *Case Study - Horten ROC*

This study is based on ASKO's autonomous barges transporting cargo from Moss to Horten, that is, across the highly trafficked Oslo fjord. The introduction of autonomous vessels will require an ROC where human operators monitor the autonomous vessels and the nearby traffic to ensure the safety and efficiency of the autonomous vessels. The main focus of the personnel in the ROC will be the autonomous vessels. Hence, it is

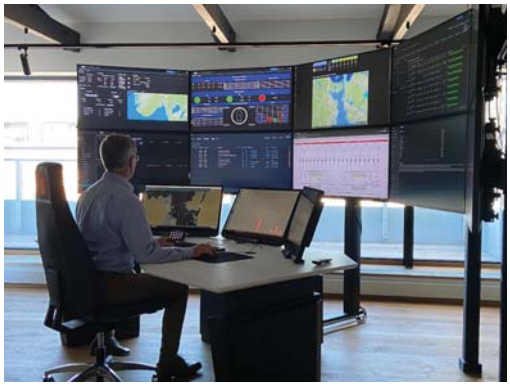


Fig. 1. Massterly Remote Operation Center in Horten. Courtesy: Kongsberg Norcontrol.

vital that surrounding traffic is automatically monitored, and that this monitoring can raise alerts of danger e.g. anomalous behavior. To develop and test the autoencoders for anomaly detection, AIS data from the Oslo fjord for one year (2019) provided by the Norwegian Coastal Administration have been used.

3.2. *Autoencoder Architecture*

Each autoencoder was developed using PyTorch, with scaling functions provided by scikit-learn. Each network was optimized using the Adam optimizer. The sizes of the networks were identified via experimentation of the latent/hidden size of the network. After various configuration trials, a hidden size of 10 was set for all autoencoders for comparative purposes. The learning rate for each network was optimized automatically via PyTorch.

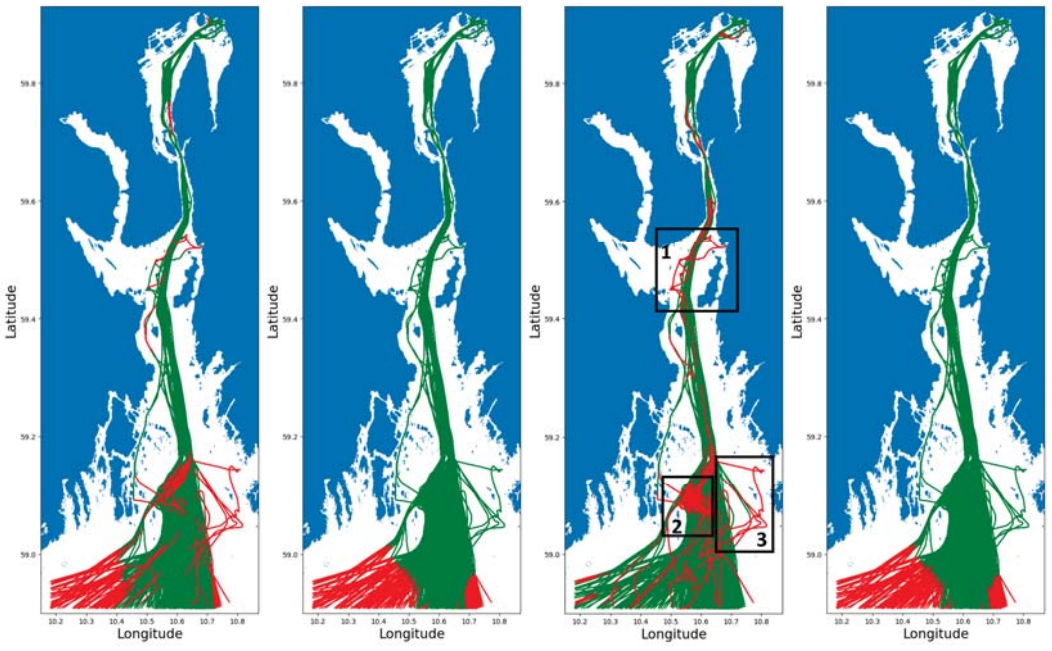
3.3. *Comparing Autoencoders*

The four different autoencoders AE, VAE, RAE, and VRAE have been tested for anomaly detection on one of the clusters in the data set, more specifically all routes from Oslo and out of the Oslo fjord. The original trajectories (green lines) and anomalies (red lines) detected by the four models are shown in Figure 2. The threshold for detecting anomalies has been set to 99% of the reconstruction loss for the training set, further discussed in Sec. 3.5. The models used to achieve these results have only been trained on positional data, i.e., latitude and longitude. As true anomalies are unknown in this data set, the performance of the models is evaluated based on a manual inspection.

In Figure 2(c) some areas of interest have been highlighted with black boxes. In box 1 and 3 there are some trajectories that have large deviation from the others and are, therefore, considered true anomalies. Unlike the VAE and VRAE, the standard AE and RAE are able to detect many of these anomalies. One can see from the results that the RAE is able to detect more anomalies in box 1, whilst the standard AE is able to detect some more anomalies in box 3.

The varitaional autoencoders, however, appear to have poor performance. The majority of discovered anomalies are centered on the southern outskirts of the routes, and appear to be mostly normal behavior. This may be due to the variational models learning more compact representations for the highest density data, and data further from this distribution (i.e. the outskirts) are penalized more than abnormal sub-trajectories in regions of higher density data. Non-variational autoencoders are able to utilize the latent space to a greater extent, and are generally better at reconstruction tasks, but have poorer performance for data generation when sampling form the latent space. The standard AE, however, also seems to focus on regions of data on the outskirts, whilst the RAE is able to capture more true anomalies.

Looking at box 2 in Figure 2(c), a lot of anomalies are detected by the RAE. However, this is an area where ships pick up pilots, which are required for many ships sailing in the Oslo fjord. Thus, the detected anomalies in box 2 are not real anomalies. However, information relating to pilot operations is to be reported. As such, anomalies in this region, can be disregarded for vessels picking up or dropping off pilots. Overall, of the four tested autonencoders, the RAE is considered to have the best performance, likely due to the recurrent architecture's ability to capture dependencies in time series data to a greater extent.



(a) Standard AE. (b) Variational AE. (c) Recurrent AE. (d) Variational recurrent AE.

Fig. 2. Different autoencoders used for anomaly detection on selected cluster trained on position data.

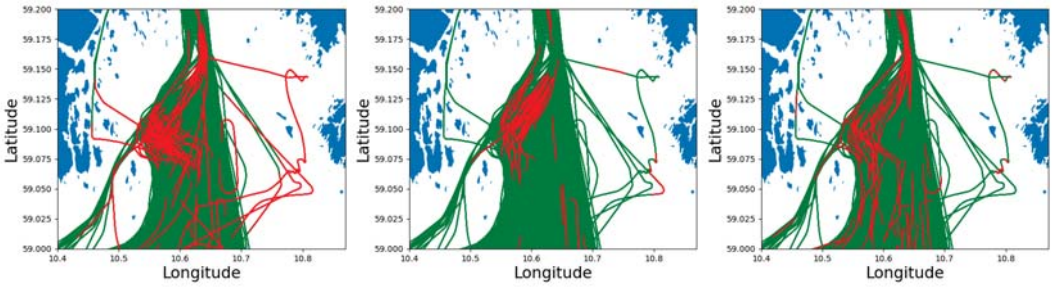
3.4. *Feature Selection*

The RAE has also been tested with additional features using the selected cluster. Figure 3(a) shows the results from the RAE trained on position (zoom of Figure 2(c)), Figure 3(b) shows the results from the RAE trained on position and SOG, and Figure 3(c) shows the RAE trained on position and COG. The anomaly detection threshold is set to 99% of the total reconstruction loss. One can see that many of the deviating trajectories on the East side of the figures are detected by the RAE trained on position alone, whilst the two RAE models trained with the additional features SOG and COG only detect parts of the anomalous trajectories.

The results indicate that the RAE trained solely on position data is the most reliable anomaly detection model. This may be due to adding redundant information. Speed for instance, is implicitly included through the time dimension, as a large deviation in position between subsequent time steps indicate a high speed and vice versa. In the following model evaluations, only positional data are used is as input to the model.

3.5. *Anomaly Detection Threshold*

For detecting anomalous trajectories in the selected cluster, i.e., in and out of Oslo, a detection threshold of 99% of the reconstruction loss was deemed suitable. However, when the RAE is trained on different clusters, this detection threshold is not always suitable. Each cluster will have varying degrees of anomalous behavior. In instances where the distribution is dominated by normal data, setting a threshold of 99% will classify normal behavior as anomalous. Setting a detection threshold in an unsupervised manner is, therefore, challenging. An alternative to using a percentage of the reconstruction loss is to threshold based on a fitted distribution of the reconstruction loss. However, using statistical measures require an assumption of the distribution of the data in the cluster, which might not be the same for all the clusters. The log-normal distribution was for instance tested, but deemed to have poorer



(a) Position. (b) Position and SOG. (c) Position and COG.

Fig. 3. Anomaly detection using a recurrent autoencoder on the selected cluster with different features.

performance. The presented models have been trained and tested on the same cluster, meaning that one model is required for each cluster.

An alternative is to train the anomaly detection model on all the clusters at once, and then apply this global model for each cluster. This method would make it easier to set a detection threshold, as one only needs to set one global threshold for all clusters. When comparing such a globally trained RAE with the RAE trained on the data within the selected cluster, it was found that the global model had degraded performance, detecting fewer anomalies than the local model. This is likely due to the global model being trained on a more diverse data set.

3.6. *Application in ROCs*

As previously introduced, the Horten ROC will be responsible for monitoring the autonomous ASKO vessels crossing the Oslo fjord. In the future, the number of vessels the operators handle may increase further. Being able to monitor multiple autonomous ships simultaneously requires that operators are notified of anomalous behavior via relevant alarms, but also that there are as few false alarms as possible.

Currently, anomalous behavior is detected through statistical evaluation of a vessel's current position, COG, and SOG with respect to historical observations in the same area. Due to among others high variation in traffic density in different areas, it has proven difficult to come up with rigid rules for when a given position should be deemed anomalous. Hence, by looking at a sliding window of the most recent observations for a given vessel, it may be possible to make more targeted detections.

The results of this study indicate the potential of an RAE to identify multiple modalities of anomalous behavior. Both routes that deviate from the primary route are identified, in addition to highly irregular behavior with multiple course corrections and alternating directions. However, as mentioned in Sec. 3.3, some anomalies are in fact normal behavior. Given that the approach is entirely unsupervised, false alarms will, therefore, occur in such cases. Systems with a high frequency of false alarms have negative effects on human operators, as they can lead to the operators ignoring many of the true alarms (Huegli et al., 2020).

Utilizing automation for decision support has also been shown to be problematic in various industries (Endsley, 2017). Some examples include decision biasing, as well as challenges related to increased system complexity. When integrating AI-based systems, the complexity of the system will inherently increase. Work is, therefore, being conducted to identify how to improve human-AI interaction (National Academies of Sciences Engineering and Medicine, 2022). These challenges will need to be addressed when implementing such an AI-based anomaly detection system as outlined in this study.

Furthermore, due to the necessity of monitoring multiple vessels live, the anomaly detection function should be able to respond within seconds. On average the RAE was able to conduct a single

anomaly classification in $1.07 * 10^{-5}$ seconds on a standard laptop. Due to its ability to run multiple predictions in parallel, the model scales well with respect to performance, and deemed to be within acceptable parameters for use in live traffic monitoring.

4. Conclusion and Further Work

To ensure the safety of autonomous ships navigating through busy waterways, it is crucial to identify any abnormal behavior in nearby vessels. This study has, therefore, developed an automatic anomaly detection system using unsupervised machine learning methods trained on historical AIS data. The system employs four deep autoencoders (AE, VAE, RAE, and VRAE) to identify anomalous ship behavior. The RAE proved to be the most effective in detecting anomalies while minimizing the number of false positives. The model is also capable of responding quickly in adherence with requirements for live traffic monitoring. Integrating this system into an ROC can help operators identify potentially hazardous situations for autonomous ships. However, it is important to note that the detection of false positives may hinder human operators' attention and decision-making abilities. Proper integration of such AI-based systems is essential to ensure the situation awareness of the operators.

Further work will focus on identifying anomaly detection thresholds automatically to reduce the number of false positives. Furthermore, deep learning approaches, e.g. autoencoders, require sufficient data volumes to ensure performance, which may not always be available for a given region. As such, future work will investigate utilizing transfer learning to facilitate anomaly detection. Many of the investigated clusters contain routes of vessels with various characteristics (e.g. ship type, length) that impact their maneuverability. As such, future work will investigate incorporating such static information into the models to improve performance.

Acknowledgement

This work has received funding from the Horizon 2020 Framework Programme of the European Union under grant agreement No 957237 (VesselAI).

References

Bourlard, H. and Y. Kamp (1988, 9). Auto-association by multilayer perceptrons and singular value decomposition. *Biological Cybernetics 59*(4-5), 291–294.

Endsley, M. R. (2017, 2). From Here to Autonomy: Lessons Learned from Human-Automation Research. *Human Factors 59*(1), 5–27.

Ester, M., H.-P. Kriegel, J. Sander, and X. Xu (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. In *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*.

Fabius, O. and J. R. van Amersfoort (2015). Variational recurrent auto-encoders. *3rd International Conference on Learning Representations, ICLR 2015 - Workshop Track Proceedings*.

Hu, J., K. Kaur, H. Lin, X. Wang, M. M. Hassan, I. Razzak, and M. Hammoudeh (2023). Intelligent Anomaly Detection of Trajectories for IoT Empowered Maritime Transportation Systems. *IEEE Transactions on Intelligent Transportation Systems 24*(2), 2382–2391.

Huegli, D., S. Merks, and A. Schwaninger (2020, 7). Automation reliability, human–machine system performance, and operator compliance: A study with airport security screeners supported by automated explosives detection systems for cabin baggage screening. *Applied Ergonomics 86*, 103094.

Iltanen, H. (2020, 10). Maritime Anomaly Detection using Autoencoders and OPTICS-OF. Technical report, Helsinki.

Kingma, D. P. and M. Welling (2014). Auto-encoding variational bayes. *2nd International Conference on Learning Representations, ICLR 2014 - Conference Track Proceedings*.

National Academies of Sciences Engineering and Medicine (2022, 1). *Human-AI Teaming: STATE-OF-THE-ART AND RESEARCH NEEDS (2022)*. National Academies Press.

National Transportation Safety Board (2020). Collision Between a Sport Utility Vehicle Operating With Partial Driving Automation and a Crash Attenuator.

Son, J.-H., J.-G. Jang, B. Choi, and K. Kim (2020). Detection of Abnormal Vessel Trajectories with Convolutional Autoencoder. *Journal of the Society of Korea Industrial and Systems Engineering 43*(4), 190–197.

Srivastava, N., E. Mansimov, and R. Salakhudinov (2015). Unsupervised learning of video representations using lstms. In *Proceedings of the 32nd International Conference on Machine Learning, PMLR*, pp. 843–852.

Wolsing, K., L. Roepert, J. Bauer, and K. Wehrle (2022). Anomaly Detection in Maritime AIS Tracks: A Review of Recent Approaches. *Journal of Marine Science and Engineering 10*(1).