The Open Maritime Traffic Analysis Dataset

Martin Masek [a], Chiou Peng Lam[a], Travis Rybicki[a], Jacob Snell[a], Daniel Wheat[a], Luke Kelly[a], Damion Glassborow[b], Cheryl Smith-Gander[b]

[a] School of Science, Edith Cowan University, Joondalup, Western Australia, [b]Defence Science and Technology Group, Department of Defence, Australia
Email: m.masek@ecu.edu.au

Abstract: Ships traverse the world's oceans for a diverse range of reasons, including the bulk transportation of goods and resources, carriage of people, exploration and fishing. The size of the oceans and the fact that they connect a multitude of different countries provide challenges in ensuring the safety of vessels at sea and the prevention of illegal activities. To assist with the tracking of ships at sea, the International Maritime Organisation stipulates the use of the Automatic Identification System (AIS) on board ships. The AIS system periodically broadcasts details of a ship's position, speed and heading, along with other parameters corresponding to the ship's type, size and set destination.

The availability of AIS data has led to a large effort to develop automated systems which could identify and be used to prevent undesirable incidents at sea. For example, detecting when ships are in danger of colliding, running aground, engaged in illegal activity, traveling at unsafe speeds, or otherwise attempting manoeuvres that exceed their physical capabilities. Despite this interest, there is a lack of a publicly available 'standard' dataset that can be used to benchmark different approaches. As such, each new approach to automated maritime activity modelling is tested using a different dataset to previous work, making the comparison of technique efficacy problematic.

In this paper a new public dataset of shipping tracks is introduced, containing data for four vessel types: cargo, tanker, fishing and passenger. Each track corresponds to a leg of a vessel's journey within an area of interest located around the west coast of Australia. The tracks in the dataset have been validated according to a set of rules, consisting of journeys at minimum 10 hours long, with no missing data. The tracks cover a three-year period (2018 to 2020) and are further categorised by month, allowing for the analysis of seasonal variations in shipping. The intention of releasing this dataset is to allow researchers developing methods for maritime behaviour analysis and classification to compare their techniques on a standard set of data.

As an example of how this dataset can be used, we use it to build a model of 'expected' behaviour trained on data for three vessel categories: cargo, tanker, and passenger vessels, using a convolutional autoencoder architecture. We then demonstrate how this model of ship behaviour can be used to test new data that was not used to build the model to determine whether a track fits the model or is an anomaly. Specifically, we verify that the behaviour of fishing vessels, whose movement patterns are quite different to those of the other three vessel types, is classified as an anomaly when presented to the trained model.

Keywords: *Maritime Track Dataset, Automatic Identification System (AIS), anomaly detection, machine learning*

1. INTRODUCTION

Many aspects of society are dependent on the movement of ships across the world's waterways, where a disruption caused by accident or incident can have large consequences in terms of loss of life and economic impact. Analysing maritime shipping movements has many applications in preventing accidents, identifying vessels in distress, uncovering illicit activities and in general optimising the flow of traffic between ports.

Data on the worldwide movement of ships is widely available, through the Automatic Identification System (AIS) (IMO, 1998). AIS is a radio frequency-based system where a ship broadcasts its position, among other static and kinematic parameters, at regular intervals. As stipulated by the International Maritime Organisation (IMO), all passenger ships, and all other ships over a defined gross tonnage are required to carry an AIS transceiver and operate it when safe to do so (IMO, 2015). This produces data that can be used for navigation and tracking purposes, enabling a ship to be aware of traffic around it and to make other traffic aware of itself. Besides real-time use, a number of commercial and government organisations collect and curate AIS data. The historical data can be analysed to determine past event sequences and occurrences.

Automated analysis of maritime traffic is an active field of research, somewhat mirroring the automated traffic analysis revolution that is occurring for cars. Consequently, a wide spectrum of modelling techniques has been employed: statistical models, machine learning models and rule-based models. Riveiro et al. (2018) provide a review of techniques that have been used. More recent techniques include the use of recurrent neural networks (Yang et al. 2020) and the use of autoencoder neural networks (Iltanen, 2020). An important aspect of work in this area is being able to test and validate new techniques. However, researchers typically employ a process where they gather their own sub-set of raw AIS data, employ their own custom procedure for cleaning it, and (as it is often obtained under commercial conditions) have no means to distribute it. Although some attempts have been made at producing a standard database of clean maritime traffic data suited to machine analysis (Mao et al. 2018), after an extensive search, we have been unable to find an available data set.

In this paper, the Open Maritime Traffic Analysis Dataset (OMTAD) is introduced. This is a dataset of cleaned and processed maritime tracks that has been produced from AIS-based data curated by the Australian Maritime Safety Authority (AMSA) for shipping within Australia's maritime search and rescue region (AMSA, n.d.). As AMSA provides their dataset in a form where the vessel identification is anonymised and under an open-source license, OMTAD can be used by the research community with minimal hurdles to explore a wide range of research questions. To demonstrate the use of this dataset, several experiments are presented. Using a sub-set of the dataset a model of 'expected' behaviour is constructed. It is then shown that this model can be used to find 'unexpected' vessel behaviour as deviations from the model.

The rest of the paper is organised as follows: In Section 2, details of the OMTAD dataset are provided, including the workflow used in its construction, structure, and details of the data contained. Section 3 details a set of experiments, building a model based on the dataset and presenting two ways of using the model for anomaly detection. Section 4 includes a discussion, possible future work, and a conclusion.

2. THE DATASET

The OMTAD dataset has been constructed from a sub-set of publicly available data, distributed by AMSA (AMSA, n.d.) under the terms of the Creative Commons Attribution-Non-commercial 3.0 Australia Licence. Since 2012, AMSA has been providing a monthly data record of vessel traffic within Australia's search and rescue region. The data released by AMSA, which has been sourced from AIS messages, is provided in a processed format. The processing includes a thinning of the data so that points from any vessel are no less than sixty minutes apart. Data is anonymised by removing vessel identifying information (such as vessel name, Maritime Mobile Service Identity number etc.) and replacing it with a unique ID number.

2.1. Dataset Construction Approach

Dataset construction consisted of choosing a region of interest and cleaning the data within that region to remove tracks that did not meet length and destination criteria, and to interpolate gaps. A 'clean' track is defined as a significant part of a vessel's journey that begins and ends either out of the region of interest or at a port and contains no gaps larger than one hour. For a month's worth of data for a particular vessel type, the following cleaning procedure was undertaken:

1. Each vessel's journey over the month was divided into individual candidate tracks.
2. Inclusion/Exclusion criteria applied to each track.
3. Interpolation of missing points.
4. Manual Verification.

The specific destination can be either a port or coastal anchorage, or for fishing vessels: a fishing area. Generating an initial list of candidate tracks (step 1) involved splitting the data for a vessel's journey at points where a speed of zero is encountered four consecutive times or where there is a gap between data points greater than 3.5 hours. The resulting journey segments were considered candidate tracks if they contained a minimum of 10 data points. Segments shorter than this (i.e. shorter than 10 hours) were discarded. Each candidate track was then examined (step 2), and if its initial and final point corresponded to either a port or the bounded area of interest, it was brought forward to step 3. In step 3, interpolation is performed, where gaps between points that are greater than 1.5 hours are filled. For gaps between 1.5 and 2.5 hours one new point is generated halfway between the points on either side of the gap, using linear interpolation to generate attributes for the new point. For gaps between 2.5 and 3.5 hours, two new points are generated, equally spaced in the gap, also using linear interpolation to generate their attributes. In the final step (step 4), each track was manually inspected as a final check of meeting the inclusion criteria before being added to the dataset. This verification step was performed as a form of quality control, since the previous parts of the process are mostly automated using rule-based scripts, and thus susceptible to inconsistent data. In turn, issues identified in the manual verification step were used to improve the automated scripts, making manual verification less necessary in future expansions of the dataset.

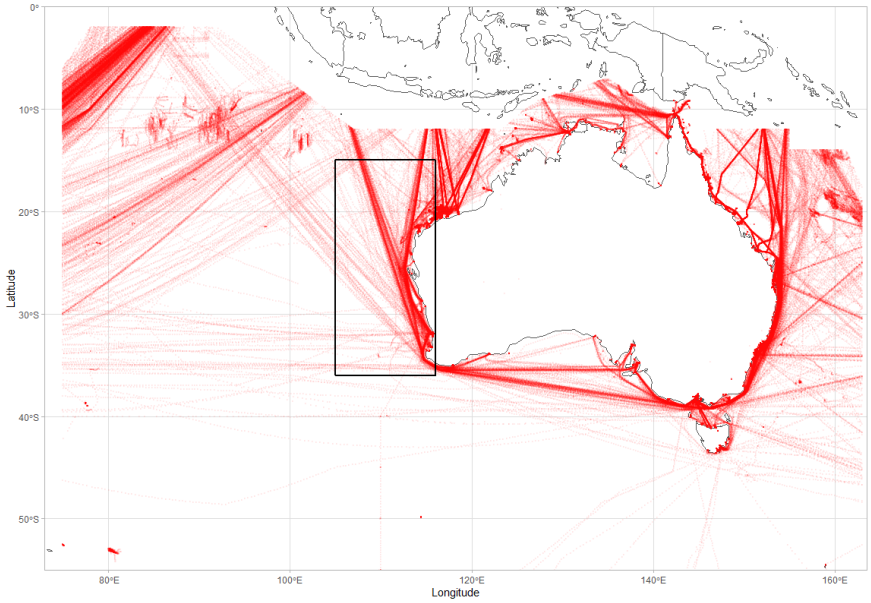


Figure 1. West coast region of interest, indicated (black rectangle) along with one month's worth of AMSA data (from January 2019).

2.2. Dataset Contents

The region of interest chosen consisted of the west coast of Australia, between 105 and 116 degrees in longitude and between -36 and -15 degrees in latitude, referred to here as the Western region of interest. To give a visual indication, Figure 1 shows a map displaying all the points provided by AMSA for January 2019, with the Western region of interest indicated using a rectangle around its bounds. The traffic patterns seen in the January data are consistent throughout the year, apart from fishing vessel tracks which vary seasonally.

The Western region was chosen, rather than the entire AMSA search and rescue region due to time constraints in producing the dataset, as it contains less traffic than the eastern side of Australia, whilst still containing a mix of shipping lanes with various vessel traffic. Also due to the time constraint, the dataset is focused on four vessel types: cargo, tanker, fishing and passenger, over three years: 2018, 2019 and 2020. The Western region of interest dataset contains 19,124 total vessel tracks. Table 1 provides a breakdown of the number of tracks for each vessel type for each year. The dataset is available from: https://github.com/EdithCowan/OMTAD.

The directory structure of the dataset contains a sub-directory for each year. Each year directory contains a sub-directory for the vessel type (*cargo*, *fishing*, *passenger* and *tanker*). Each vessel type directory contains 12 sub-directories, one for each month of the year. Each of the month directories contains a Master Processed File (MPF) containing a set of processed track data for that vessel type as comma separated values (csv). The MPF datafile uses the following naming convention: 'MPF_[month]_[year]_Grid_[TYPE].csv'. The directory

structure is provided for convenience, but users of the dataset are free to copy all MPF files into a single directory, as the naming convention includes the same information as provided by the directory structure.

Table 1. Breakdown of vessel tracks in the dataset by year and vessel type.

Year	Cargo Tracks	Tanker Tracks	Fishing Tracks	Passenger Tracks
2018	4,974	1,343	75	101
2019	4,943	1,373	155	65
2020	4,467	1,304	236	88
Total	**14,384**	**4,020**	**466**	**254**

Each MPF file is organised such that the first row is a header with the names for each column. Following the header is the data, with each row containing the field values for a single point on the track. Tracks are separated by a row that consists of the word "END". The field names, along with the corresponding type and description used in the MPF files are given in Table 2.

Table 2. Fields in the dataset Master Processed Files.

Field Name	Type	Description
CRAFT_ID	Text	Unique identifier for each vessel – matches AMSA CRAFT_ID
LON	Double	Longitude in decimal degrees
LAT	Double	Latitude in decimal degrees
COURSE	Double	Course over ground in decimal degrees
SPEED	Double	Speed over ground in knots
TIMESTAMP	Text	Vessel position report UTC timestamp in dd/mm/yyyy hh:mm:ss, 24 hour format
Track_ID	Text	Unique track identifier in the format yyyy[Type][Month][sequence number] (eg. "2020CargoJan1" for the first track in the 2020, January cargo vessel MPF file).

3. EXPERIMENTS

In this section some guidance to those wishing to build models of maritime activity using the OMTAD dataset is provided. The typical workflow starts at selecting a model appropriate to the task, tuning the model parameters, and then building and using/evaluating the model. The chosen example application is to model the 'expected' kinematic behaviour for a particular behaviour type and then test the model using vessels of a type that typically exhibits different behaviour, treating them as a previously 'unseen' test set. Examining the behaviour of vessel types visually, the behaviour of cargo, tanker and passenger vessels is very similar (in terms of the routes travelled, moving in straight line segments between ports), whereas fishing vessel behaviour is noticeably different. Thus, for a model trained on the behaviour evident in cargo, tanker and passenger vessels, it would be expected for the fishing vessel track set to be identified as anomalies, whilst new, unseen tracks of cargo, tanker and passenger vessel types should be classified as fitting the model. The model architecture, tuning procedure and experiments will now be described.

3.1. Model Architecture

The model is based on the convolutional autoencoder neural network, which uses convolutional layers within the autoencoder architecture first proposed by Rumelhart et al. (1985). Autoencoders can learn to encode a compressed representation of their input, accomplished by training the network to reproduce its input as its output, whilst constricting the number of neurons in its middle layers, forming a bottleneck. This is a form of unsupervised learning, as it does not need the training data to be labelled (the expected output should simply be equal to the input) and is thus ideal for use in anomaly detection, where an anomaly is defined as 'different' from the expected, rather than having a specific label in the dataset. As such convolutional autoencoders have been used for anomaly detection in a variety of fields, including detection of network anomalies (Chen et al. 2018) and anomalies in video (Ribeiro et al. 2018).

The input for the model is a set of sequential data points, each comprising a small 'window' of an entire track (the optimal window size is the subject of the tuning step). The architecture of the autoencoder is shown in

Figure 2 for a window size of 16. This means that the input layer takes in a sequence of 16 datapoints from a track, where in each data point three values are included: latitude, longitude and speed of the ship at that point (48 values altogether). Similarly, the output also consists of 48 values. The middle, bottleneck, layer however only accommodates 16 single values, thus compressing the 48 input values into 16 values at that point in the network.

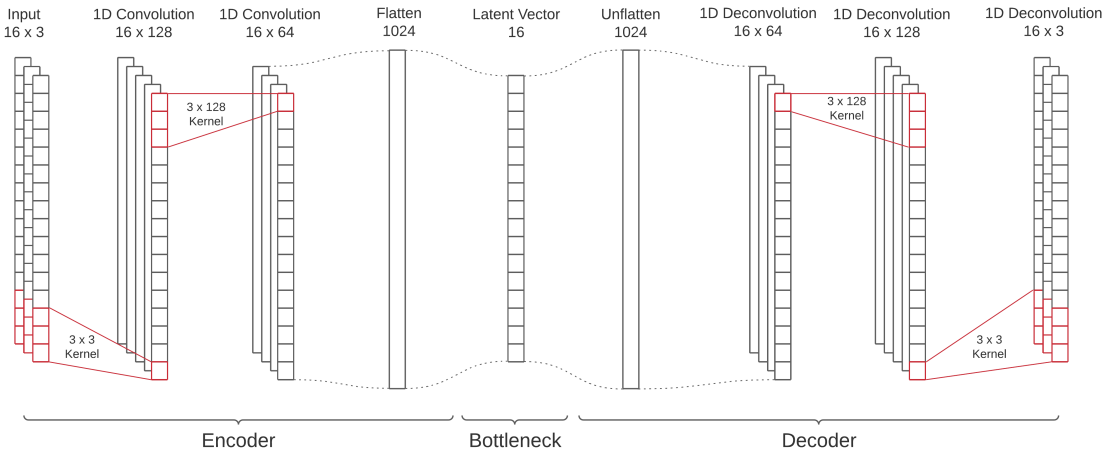


Figure 2. Architecture of the autoencoder neural network, here for a window size of 16, where the input is 16 consecutive data points, each made up of a latitude, longitude and speed attribute.

The model is trained using a training set of track windows, with the aim of minimising the reconstruction error. The reconstruction error is calculated as the mean square error between the input and output vectors. After training, the model can be presented with a new 'unknown' window of data, with the response of the network being used to determine whether that data fits the model or is an anomaly. For this anomaly detection, experiments focused on two approaches of using network response to flag anomalies.

In the first anomaly detection approach, the reconstruction error was used as an indication of how well a new window of data fits the model. A static threshold was defined for the reconstruction error, using a percentage of the maximum reconstruction error when the model is tested on the training data.

In the second anomaly detection approach, the trained autoencoder was taken, and the output of its 'bottleneck' layer used as input to a one-class support vector machine (OC-SVM) (Schölkopf et al. 2001). The aim of this OC-SVM is to predict if a vessel track window is an outlier. This approach was used by Wang et al. (2021) to detect structural damage anomalies, where the autoencoder and OC-SVM were trained using sensor data from undamaged structures. The OC-SVM is trained using the same training data as used for the training of the autoencoder.

3.2. Tuning Approach

Each modelling technique will include some parameters that need to be set. Typically, the optimal value of these settings is application-dependent, and thus some kind of tuning needs to be performed. In our case, the main data-related parameter is the window size. Four different sliding window sizes were tested: 3, 6, 10, and 16. In the absence of normal/abnormal labels for each track, tracks of different vessel types were used to tune the window size. In this approach, the models were trained on tracks of specific vessel types such as cargo (using 80% of the dataset for training and 20% for validation). The models were then tested on all four vessel types in the dataset. An 'optimal' window size was chosen based on the one that produced a model that classified as anomalies a low percentage of the vessel type it was trained on and a high percentage of the vessel type it was not trained on. Of the window sizes tested in this way, a size of 16 was found to have the best performance.

3.3. Results

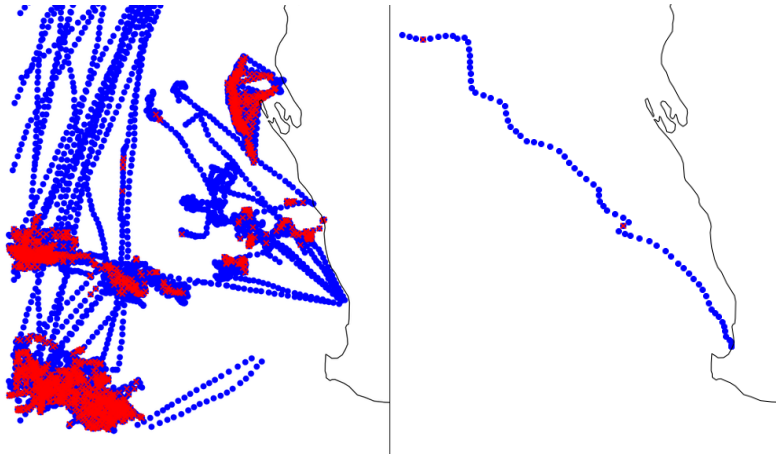
Here results are presented from the two convolutional autoencoder models trained on a combined dataset of cargo, tanker and passenger vessel tracks from 2019. As discussed, these three vessel types exhibit similar behaviour of traversing port-to-port, typically using shortest path routing. Results are presented in terms of how well the trained models can flag tracks belonging to the fishing vessel types (also from 2019) as anomalies. The training data was split with 80% of the combined cargo, tanker and passenger tracks (5,105 tracks), with

the remaining 20% (1,275 tracks) used as a test set, to verify the resulting model on the behaviour type it was trained with. All tracks belonging to the fishing vessel type (155 tracks) were used as an unseen test set of anomalies. A sliding window size of 16 was used based on the tuning experiments. The threshold-based autoencoder had its static threshold set at 45% of the maximum reconstruction error on the training dataset. This was chosen so that nearly 5% of the cargo/tanker/passenger testing tracks were predicted as outliers for analysis purposes. The OC-SVM based autoencoder also had its OC-SVM hyperparameters adjusted so that nearly 5% of the cargo/tanker/passenger test tracks were predicted as outliers. The training and evaluation process for both models was repeated 10 times on randomly shuffled tracks, with Table 3 showing the average percentage of tracks predicted as outliers across both the cargo/tanker/passenger test set and the fishing vessel track data.

Table 3. Average percentage of tracks for each vessel type classified as anomalies by models trained on the combination of cargo, tanker and passenger vessels in the dataset.

Model \ Vessel Type	Cargo, Tanker & Passenger	Fishing
Threshold Autoencoder	5.66%	68.06%
OC-SVM Autoencoder	5.75%	34.78%

These results validate the visual findings that 'normally' cargo vessels, tankers and passenger ships display the same behaviour, traversing identical shipping channels (typically travelling in straight lines between ports and around coastlines, at consistent speeds). Fishing vessels, however, often move very slowly, and stay within a small region for large lengths of time. The performance of the threshold-based method on fishing vessel data can be seen visually in Figure 3(a), where the data points that were classified as anomalous are shown in red. This figure illustrates that points along the vessel's journey to a fishing area are considered normal (as this part of the track is a straight line) but points corresponding to fishing activity, consisting of manoeuvring back and forth within a local area of the ocean, are classified as anomalies. When examining the tracks in the cargo/tanker/passenger vessel test set that were also classified as anomalies, such deviations from the usual straight-line travel behaviour are also seen. An example for a cargo vessel track flagged as an anomaly, displaying a movement pattern that is not straight, is shown in Figure 3(b).



(a) (b)

Figure 3. (a) The locations (shown in red) where threshold-based model, trained on a combination of cargo, tanker and passenger vessels, predicted anomalies in the Fishing vessel tracks, and (b) an example of a Cargo track that was flagged as an anomaly by the same cargo/tanker/passenger-trained model.

4. DISCUSSION AND CONCLUSION

In this paper the OMTAD dataset of maritime traffic has been introduced, documenting the process of its construction, the data it holds, and example usage for the application of anomaly detection. Though raw AIS maritime traffic data is widely available for both live and historical contexts, to the best of our knowledge, ours is the only publicly available dataset of cleaned tracks that have been validated against our defined criteria. As such, it is suitable for use as a 'ground truth' dataset of normal vessel behaviour for the vessel types that have been included, in the region of interest.

Some examples were shown of how the dataset can be used to build models of maritime behaviour and to determine if an 'unknown' vessel fits that behavioural model. This has applications in anomaly detection, where a vessel is claiming to be particular type but exhibiting different behaviour, and also vessel identification, where an unknown vessel could be classified based how close its behaviour is to one of the known vessel types. These are limited to vessel behaviour that is present in the dataset. In future work we are aiming to expand the types of models used, with a more detailed investigation in terms of effects of various parameters, with a view of automating the optimisation process. We are also aiming to expand the region of interest to cover a larger proportion of the total Australian search and rescue zone, and to capture data on more vessel types to enable modelling of their behaviour. The hope is that the OMTAD dataset can become a standard benchmark for testing and comparing methods of maritime behaviour analysis.

ACKNOWLEDGEMENTS

This research is supported by the Commonwealth of Australia as represented by the Defence Science and Technology Group of the Department of Defence.

REFERENCES

[dataset] AMSA., n.d. Vessel Tracking Data. Australian Maritime Safety Authority, https://www.operations.amsa.gov.au/Spatial/DataServices/DigitalData.

Chen, Z., Yeo, C. K., Lee, B. S., & Lau, C. T., 2018. Autoencoder-based network anomaly detection. Wireless Telecommunications Symposium, 2018-April, 1–5. https://doi.org/10.1109/WTS.2018.8363930

Davenport, M., 2008, Kinematic Behaviour Anomaly Detection (KBAD)-Final Report. DRDC CORA report KBAD-RP-52-6615.

Iltanen, H., 2020. Maritime Anomaly Detection using Autoencoders and OPTICS-OF. Masters Thesis, University of Helsinki.

IMO., 2015. Revised Guidelines for the onboard operational use of shipborne automatic identification systems (AIS), Resolution A. 1106 (29).

IMO., 1998. Recommendation on Performance Standards for a ship-borne Automatic Identification System (AIS). MSC 74(69) Annex 3.

Mao, S., Tu, E., Zhang, G., Rachmawati, L., Rajabally, E., & Huang, G. B., 2018. An automatic identification system (AIS) database for maritime trajectory prediction and data mining. In Proceedings of ELM-2016 (pp. 241-257). Springer, Cham.

Ribeiro, M., Lazzaretti, A. E., & Lopes, H. S., 2018. A study of deep convolutional auto-encoders for anomaly detection in videos. Pattern Recognition Letters, 105, 13-22.

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

Rumelhart, D. E., Hinton, G. E., & Williams, R. J., 1985. Learning Internal Representations by Error Propagation. Readings in Cognitive Science: A Perspective from Psychology and Artificial Intelligence, V, 399–421. https://doi.org/10.1016/B978-1-4832-1446-7.50035-2

Schölkopf, B., Platt, J. C., Shawe-Taylor, J., Smola, A. J., & Williamson, R. C., 2001. Estimating the support of a high-dimensional distribution. Neural computation, 13(7), 1443-1471.

Wang, Z., & Cha, Y. J. (2021). Unsupervised deep learning approach using a deep auto-encoder with a one-class support vector machine to detect damage. Structural Health Monitoring, 20(1), 406–425. https://doi.org/10.1177/1475921720934051

Yang, S., Xinya, P., Zexuan, D. and Jiansen, Z., 2020. An Approach to Ship Behavior Prediction Based on AIS and RNN Optimization Model. International Journal of Transportation Engineering and Technology, 6(1), p.16.