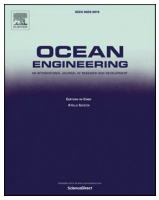
Ocean Engineering 319 (2025) 120043

Contents lists available at ScienceDirect

Ocean Engineering

journal homepage: www.elsevier.com/locate/oceaneng



Research paper

Vessel classification using AIS data

Rory Meyer *, Waldo Kleynhans

University of Pretoria, South Africa

A B S T R A C T

Maritime Domain Awareness (MDA) relies heavily on Automated Identification System (AIS) data for vessel tracking. This research focuses on developing a novel vessel classification framework that uses AIS derived features. The algorithm effectively classifies ocean-going vessels into behavioural categories, providing valuable insights for MDA.

Results: demonstrate the effectiveness of the classification framework in achieving high accuracy (F1 score of 0.88–0.9) in vessel classification. The choice of class labels and data pre-filtering significantly impacts performance. The algorithm's feature importance analysis highlights the relevance of self-reported vessel dimensions, location, and behaviour.

While cargo and tanker vessels exhibit some overlap, fishing vessels are accurately classified. However, recreational and passenger vessels, due to limited samples, require further refinement. Future research could explore time series methods and tailored algorithms for specific vessel classes to enhance classification accuracy. Overall, this study contributes to improving MDA by providing a robust vessel classification tool. Further investigation is needed to address the high proportion of unlabeled vessels classified as fishing vessels.

1. Introduction

Maritime Domain Awareness (MDA) is the effective understanding of the resources, operations, and vessels that could impact the safety, security, and economy of the nation who is responsible for the nearby ocean. Automated Identification System (AIS) data has become the default method for locating and identifying ocean going ships and has become a primary data source for MDA. While it is a legal requirement for most ships to carry AIS devices it is very easy for sailors to disable the transmitter, insert incorrect identifying data fields, or for coastal AIS receivers to lose track of vessels due to VHF propagation limitations. An examination of an AIS dataset will also show that a large portion of ocean going vessels provide non ideal classification information either due to not filling anything into the correct field, and therefore defaulting to "Unknown", or the exact class of the vessel does not fit into the limited set of options described in the AIS protocol and therefore self describing as "Other".

The objective of this research was to develop a novel vessel classification algorithm that could be used in a production environment. This algorithm can be used to quickly classify ocean going vessels into one of several behavioural classes, rather than self reported AIS classes, therefore allowing operators to quickly identify possible classes, requirements, or risks associated with an unknown vessel. Providing decision makers with automated tools that operate on minimally filtered AIS data would reduce the load on operators, allow the quick identification of vessels not behaving as expected and improve overall MDA.

There are some recent published research results for an algorithm that serves a similar purpose (Wang et al., 2021). This work aims to obtain similar results using novel feature engineering and while using more data (from a different source) that covers a larger area, set of activities, and time scale than previously used while also not filtering out vessels near shore. The bounding box of the dataset used in this work covers Latitudes between −70 and 38°, and Longitudes between −60 and 100°. The correct choice of machine learning algorithm would allow operators to interrogate the classification decision to determine which data features resulted in the decision. This removal of the black box nature of some algorithms would allow operators-in-the-loop to detect false positives.

This work details the dataset and methods used to obtained accuracy results similar to other published work.

- Using a widely available, performant, open source algorithm
- Using only a commercially available, near-global dataset consisting of ~1.8 Billion AIS messages
- Without location or behaviour based filtering of vessels

Using an open and easily reproducible algorithm would allow the quick reproduction of this work for researchers and decision makers with access to their own AIS data sources. The LightGBM algorithm is

* Corresponding author.
E-mail addresses: rory@oceanmaps.xyz (R. Meyer), waldo.kleynhans@up.ac.za (W. Kleynhans).

https://doi.org/10.1016/j.oceaneng.2024.120043
Received 3 April 2024; Received in revised form 10 November 2024; Accepted 5 December 2024
Available online 30 December 2024

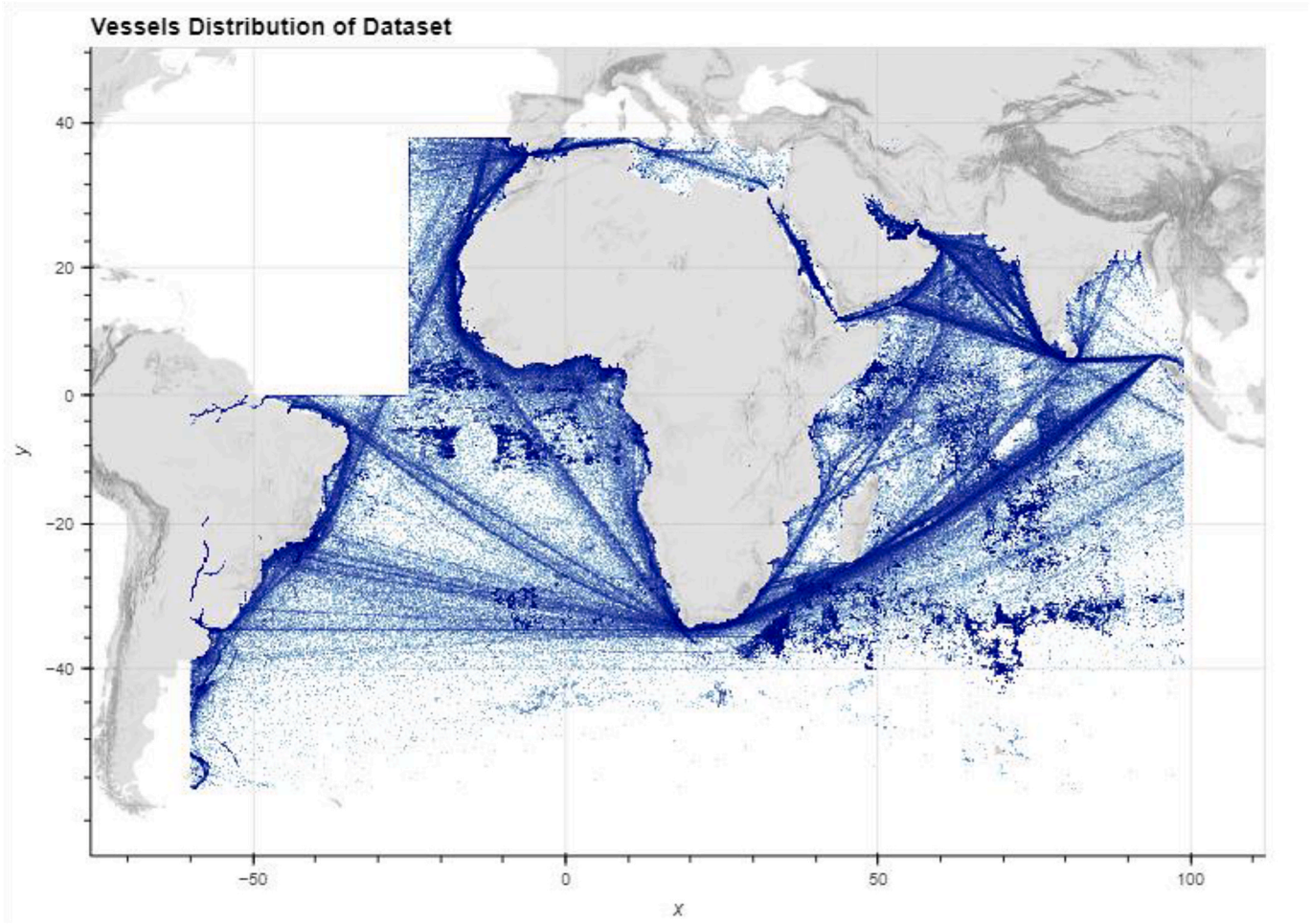


Fig. 1. Spatial distribution of OCIMS dataset.

available through many different common libraries and is a practical alternative to complicated, custom algorithms currently published.

1.1. Background and related work

The Automated Identification System (AIS) was originally developed to allow vessels to avoid collisions at sea and is a legal requirement for all passenger ships, vessels above 300 gross tons involved in international travel, and for cargo ships above 500 gross tons ('International Convention for the Safety, 1974). Other promoted benefits are accident investigation, aids to navigation and search and rescue (SAR) operations. AIS is used to report the location and identity of a vessel to other nearby vessels using VHF radio. The propagation mechanics of VHF radio over oceans, and the location of the transmitting/receiving antenna result in the AIS message being propagated mostly line of sight (Tang et al., 2019). To overcome this limitation satellite based receiver networks are used to receive AIS messages out of sight of the coastline.

The AIS protocol has 27 different message types. The data used in this study was the Position and Voyage Reports for class A and class B receivers; message types 1,2,3,5,18 and 19. The AIS protocol does not include timestamps into the message format but, in general, these are added by the data provider using a provider specific metadata format. The Position Reports are derived from global positioning systems and are, in general, fairly accurate (Jankowski et al., 2021). These reports contain information on the position, speed and direction of travel of the vessel. The Voyage Reports contain information on the size and classification of the vessel as well as information on the destination and ETA but this information is less accurate as it is configured by sailors on board (Harati-Mokhtari et al., 2007). Vessel trajectories, built from aggregations of AIS messages, have been used to study vessel behaviour (Kabir et al., 2022), (Kabir et al., 2024), (Paudel et al., 2024).

Vessel identification and classification information can be unreliable in AIS messages (Balduzzi et al., 2014). Operators responsible for monitoring a large region often have thousands of vessels in their area of interest. Automatically identifying a few vessels that are operating unusually, either due to malicious intent or unusual circumstances, would reduce the load on operators. This can be achieved by a combination of automated vessel classification and anomaly detection (Balci and Pegg, 2006). These tools also need the ability to be quickly retrained on new data and to examine the reasons for any classification decision made in order for operators to quickly evaluate data that caused the classification.

AIS classification, or moving object classification, has been a topic of interest to many researchers who have taken multiple different approaches to the problem.

- Classification of specific vessel classes from movement data (Sánchez et al., 2020)
- Using image classification techniques on generated images of vessel trajectories (Chen et al., 2020), (Rintoul and Wilson, 2015)
- Classification of ships using clustering techniques (Zhou et al., 2019)
- Classification using ensemble algorithms(Wang et al., 2021)

These research projects also tend to work with subsets of data, either due to a specific focus or lack of high quality global data.

- AIS covers single port (Sánchez et al., 2020) (Zhou et al., 2019)
- AIS covers single country/region or several small countries (Chen et al., 2020)
- AIS covers multiple regions/countries/oceans or entire globe (Wang et al., 2021)
- Subsets of data that behaviours or classes removed (Wang et al., 2021)

The results of this work are compared to Wang et al. (2021) since their AIS dataset is similar to the dataset used in this work; both AIS datasets cover large portions of the globe, have been collected from satellite based receivers, and have been collected over multiple months. The footprint in (Wang et al., 2021) covers the entire globe while the dataset used in this work is covered by the bounding box described by Latitudes between −70 and 38°, and Longitudes between −60 and 100°. Some of the preprocessing steps in (Wang et al., 2021) result in filtering out coastal AIS data and removal of vessels at rest.

These classification algorithms make use of a combination of features from Position Reports, trajectories and Voyage Reports. The classification algorithm examined in this paper makes use of AIS messages and features derived from a trajectory built from AIS messages from the previous 12–24 h without removing vessels at rest.

Ensemble machine learning algorithms, particularly gradient boosting algorithms, have shown good results with classification and regression problems as measured by accuracy, training speed, and generalisation capabilities (Bentéjac et al., 2021). LightGBM is a modern implementation of a gradient boosting algorithm that prioritises training speed. Given the large size of the training dataset, the number of features and classes, it was beneficial to have a short training time that would allow researchers to quickly iterate and experiment with ideas and feature engineering. LightGBM is a gradient boosting algorithm that has shown to be performant while requiring reduced compute capabilities when compared to other gradient boosting algorithms (Liao et al., 2022), (Monteiro et al., 2024), (Ke et al. et al., 2017). Another benefit of using ensemble models, like LightGBM, is the ability to explain the output of a model based on the input features (Lundberg et al., 2018) which is not covered in this work, but nonetheless had an impact on the decision to use LightGBM.

The data used in this study was collected through South Africa's National Oceans and Coasts Information Management System project (OCIMS)[1] The AIS was data gathered through a commercial network of satellite based receivers and covers the area between the Mediterranean sea and Antarctica, and South America and Malaysia. This footprint covers approximately 32% of the planet's surface along with multiple ports, fishing grounds and shipping lanes. Fig. 1 shows the footprint of the data collected. The data used stretches between 2020-01-01 and 2020-10-01. The size of this dataset, in spatial coverage and time, makes this study relevant to research covering global spatio-temporal datasets.

The encoded AIS data was read, decoded, filtered and inserted into a spatial database using open source software.[2] Postgresql was used as the database and this allowed the use of geospatial (PostGIS) and time-series functions (TimeScaleDB) to further process the data. The OCIMS AIS dataset, for the period of study, contained approximately 1.8 billion AIS messages. Using the complete dataset for training and testing purposes would be infeasible so the dataset was sampled. The method of sampling used was to first find the position and Voyage Report, for each vessel, closest to 00:00:00 + 0 GMT for each day and then to build a trajectory for the sampled vessels. Additional processing on the AIS sample and trajectories was performed to extract features for the machine learning algorithm. Daily data was sampled for the first 9 months of 2020.

Some fields from the Voyage Reports were joined to the Position Reports to provide the self reported physical dimensions and class of the vessel. Vessels with unknown classes or conflicting classes, due to multiple reports stating different classes, were dropped from the dataset.

To avoid cross contamination between the training and testing datasets the two sets were not split by time but instead by the vessel identifier; MMSI. Experiments that were run with training on Q1 and Q2 datasets and validated with Q3 datasets showed unrealistically high accuracy and lack of generalisation.

The sampled OCIMS datasets are described in Table 1.

Table 1
Dataset descriptors.

Dataset Name	First Date	Last Date	Row Count	Unique Vessels
OCIMS 2020 Q1, Q2 and Q3	2020-01-01	2020-10-01	2,293,675	32,021

1.2. Dataset overview

The dataset covers a large area of space and time and contains all the various vessel classes and activities that take place therein. The number of vessels, and their associated AIS class is shown in Fig. 2. This figure shows that the class distribution is heavily skewed towards Cargo, Tanker, Fishing classes, and there are significant numbers of vessels with "Other" or "Not Available" as class labels. This large class of unknown vessels is due to the adequate class not being selectable from the limited values in the AIS protocol (net markers for example) or not being correctly inserted into the AIS equipment by sailors.

This table also shows that there are significant numbers of vessels self reporting as "Reserved for Future Use" classes that, strictly speaking, is incorrect. Having a large number of classification targets when it is known that the self reported classes are inaccurate could result in model inaccuracies and user confusion. Another way of describing the vessels would be to aggregate classes into groups; for example cargo ships of all types can be labelled as "Cargo" and vessels associated with port activity (tugboats, towing boats, pilot vessels etc) grouped into a new "Port" class.

2. Machine learning methodology

Fig. 3 shows a top level block diagram of the algorithm training and testing steps. The steps can be broadly broken down to the preprocessing steps required to create the machine learning dataset, the steps used to train a vessel classifier, and lastly the steps used to evaluate the classifier's performance. The results from the algorithm evaluation are used to optimise the preprocessing and training steps.

Raw AIS data was decoded and inserted into a geospatial database. This schema of this database forms part of the OpenAIS open source project. Several queries were used to create and join Trajectory Derived features to Position and Voyage Report features. No class stratification was implemented so any implementation of a machine learning algorithm would need to take into account possible unbalanced classes.

LightGBM is an open source, gradient framework based on decision trees with a focus made on performance and scalability (Balduzzi et al., 2014). There is no need to normalise or scale features, due to it being based on decision trees, and imbalanced classes can be handled by weighting classes by automatically adjusting weights inversely proportional to class frequencies in the training data.

Fig. 4 shows the machine learning dataset being created from the OCIMS dataset. The AIS dataset was decoded, and gathered into Voyage Reports and Position Reports based on AIS message type. The Trajectory Features were created from a time-series aggregate of vessel positions.

The training and testing datasets were created by first removing all vessels that did not have a reported class and then split into training and testing datasets by randomly allocating each vessel's MMSI to one of the two datasets. This was done to avoid having the same vessel in both

[1] https://ocims.environment.gov.za/About.html.
[2] https://openais.xyz/.

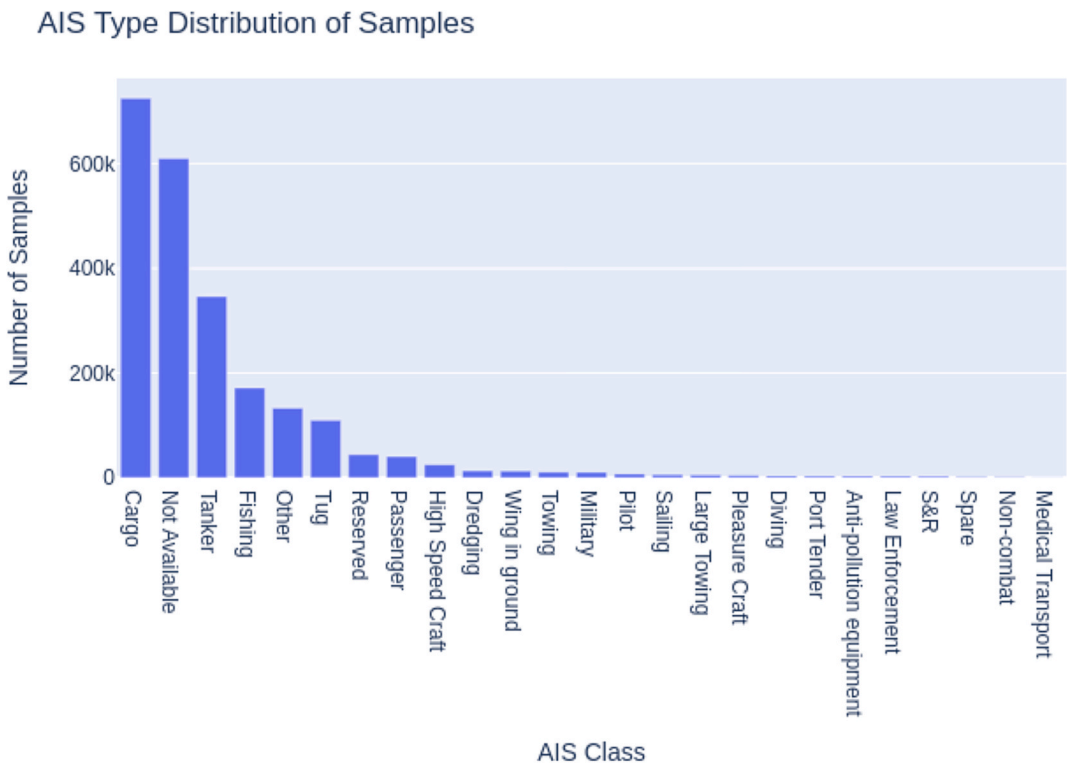


Fig. 2. Distribution of AIS Types from OCIMS dataset.

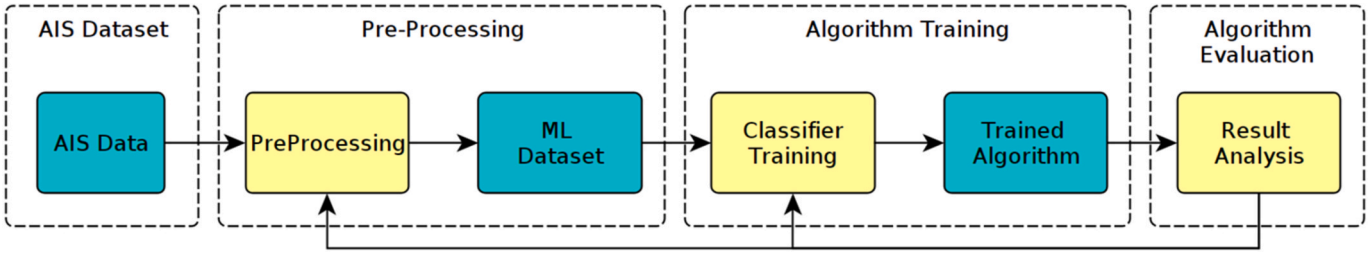


Fig. 3. Block Diagram of Classifier Training and Testing. Yellow blocks represent a processing step while green blocks represent a static object or dataset. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

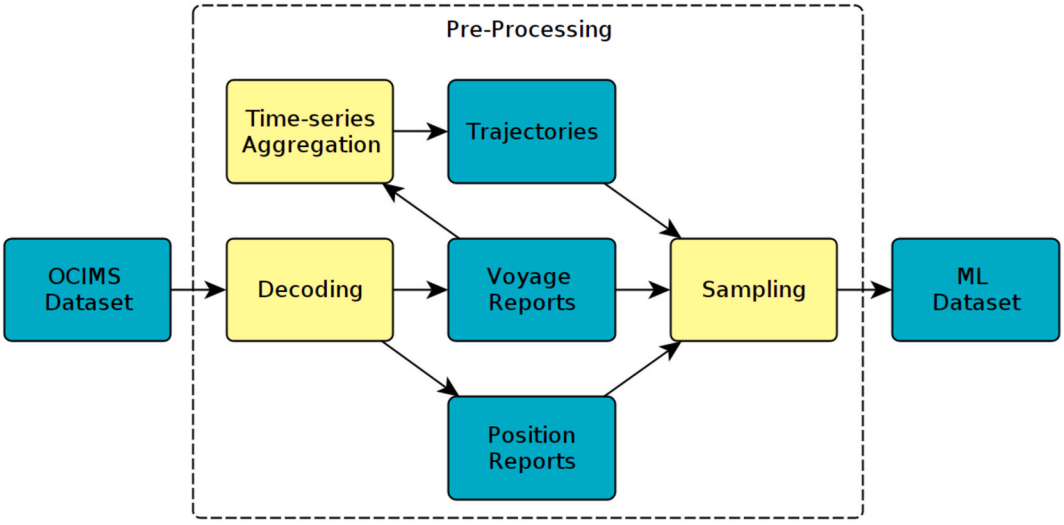


Fig. 4. Block diagram showing the OCIMS dataset being processed and sampled to create the ML dataset.

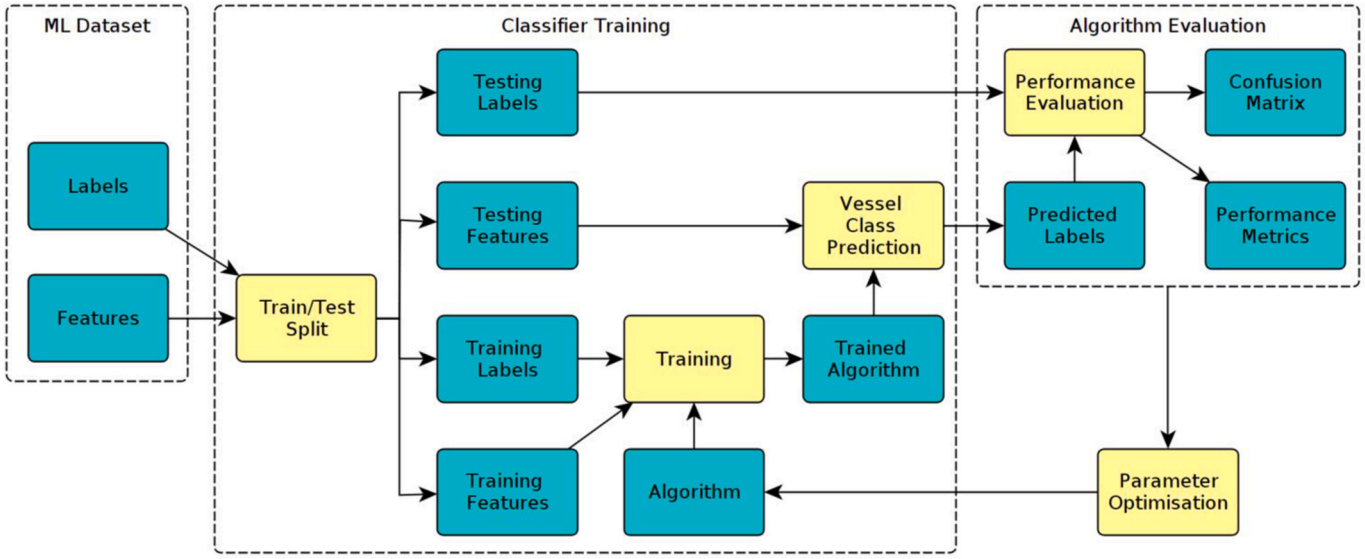


Fig. 5. Block diagram showing Training and Testing data flow.

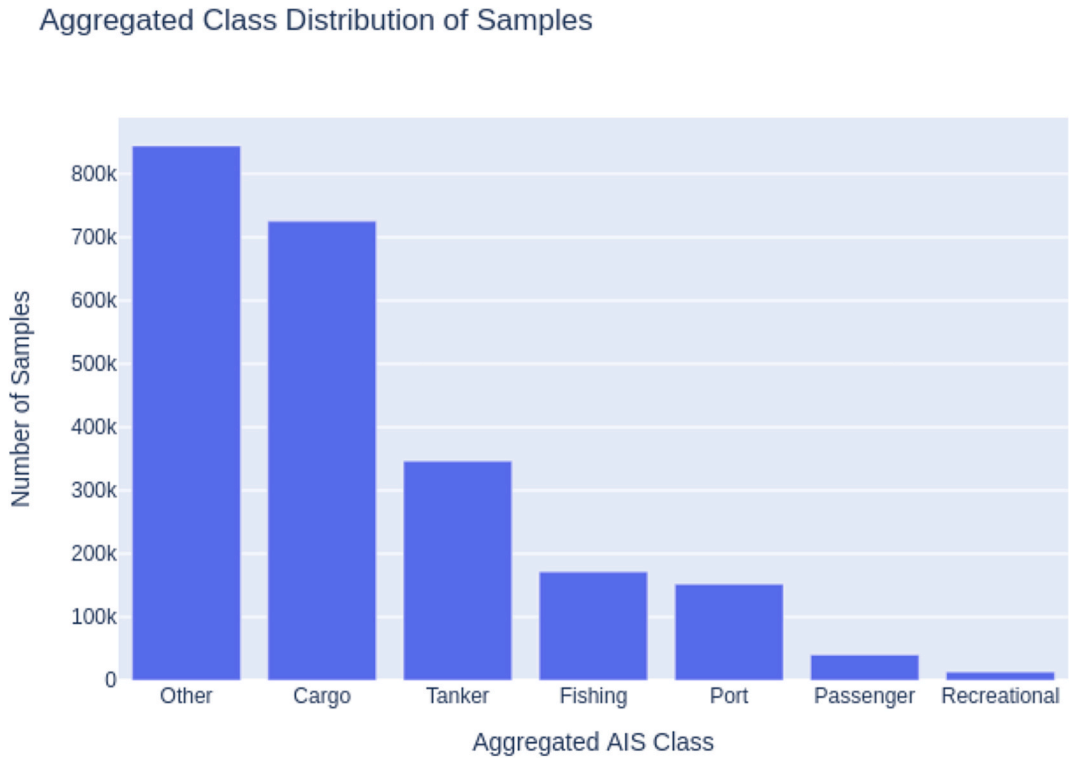


Fig. 6. Aggregated AIS vessel classes.

training and testing sets as experimentation showed that this led to overfitting. The training portion formed 70% of the split while testing formed 30%. The vessels were allocated to a classification target by mapping their AIS Voyage Report class to a Classification Targets label set. A full table of class mapping is shown in the Supporting Information section.

Fig. 5 shows the train/test split, and how the different datasets are used to train, and then test the algorithm. These steps are repeated for different algorithm parameter sets to find those that return the highest performance metrics.

A hyperparameter grid search (Feurer et al., 2019), optimised for F1 accuracy (Grandini et al., 2020), was run for each set of labels. The training dataset was used for each Classification Target to determine the feature importance, confusion matrix and accuracy score tables. These scores describe the performance of the algorithm and were further used to tune any parameters and features. LightGBM allowed a single training and testing run to be completed in under 3 min.

The LightGBM parameters used for each class label, as well as the various scores, are listed in the results section.

3. classification targets

Any classification algorithm requires an accurate label with enough samples to support the training step in order to correctly classify future

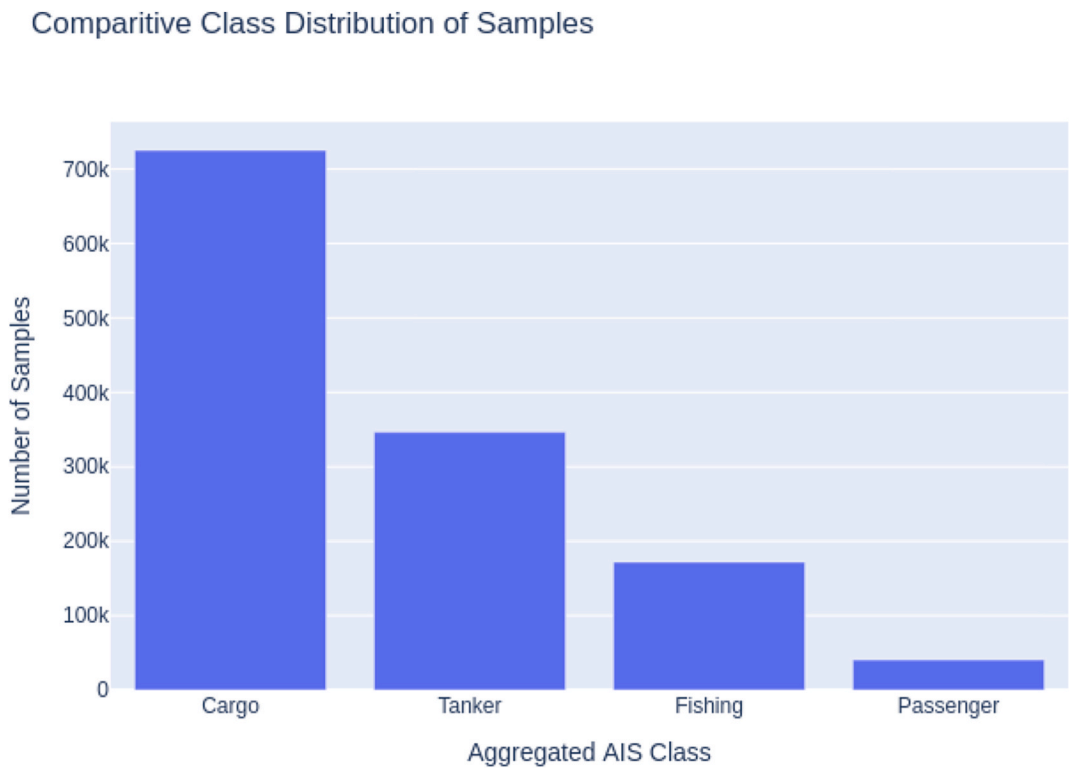


Fig. 7. Comparative class distribution.

samples. The AIS protocol allows for approximately 83 different classes but these follow a long tail distribution that would make the classification of the smaller, less representative, classes difficult. Any training metric or algorithm chosen would also need to be able to handle unequal class distributions. Collapsing these classes into smaller sets would provide additional training samples for each super-class while also still providing human meaningful classification results.

Fig. 6 shows an aggregation of classes where all cargo, tanker, and passenger subclasses were grouped together and the human readable labels applied. Unknown, not reported and incorrect classes were aggregated into an "Other" label.

Fig. 6 shows that there is still a long tail distribution of classes so a further step was taken where classes that typically operate in ports (Tugs, Towing, Pilot, Dredging, Port Tender, Law Enforcement) were grouped together into "Port" classes and classes related to recreational activities (Sailing, Diving, and Pleasure Craft) were grouped together into "Recreational" classes.

In order to compare the results of the algorithm with another paper (Wang et al., 2021) the experiments were rerun with only the Cargo, Tanker, Fishing and Passenger classes in the dataset and the distribution is shown in Fig. 7.

4. Features

The features used as inputs to the classification algorithm are derived from three sources; a single Position Report, a historical trajectory built from several hours of Position Reports, and a Voyage Report all associated with the same vessel by MMSI.

4.1. Position Report features

AIS Position reports are considered to be the AIS messages that include position vector data objects derived from global positioning systems and consist of the AIS message types 1, 2, 3, 18 and 19. A subset of the data contained in these messages was used to create the Position Report feature set and are described in Table 2.

Table 2
Features extracted from Position Reports.

Name	Description	Units
Course Over Ground [COG]	Course Over Ground is the direction of travel of a vessel, between two points, relative to the earth's surface.	Degrees from North
Speed Over Ground [SOG]	Speed over Ground is the vessel's speed relative to the land or any other fixed object such as buoys.	Knots
Longitude and Latitude	Location of ship in WGS 84 format	WGS 84

4.2. Trajectory derived features

Some features were derived from the trajectory built from several hours of AIS data for each ship. The positions reported by the vessel were ordered in time, and grouped by MMSI to create a line representing the vessel's travel over time. This was achieved using PostGIS aggregate and ST_MakeLine functions to create a "linestring" data object, with a start and end timestamp, this was then used to extract features that describe the vessel's trajectory.

These trajectory derived features were chosen to describe components that "represent succinctly the salient information in trajectories" (Rintoul and Wilson, 2015). Two features describing the changes in a vessel's course-over-ground were added, one that only considered the absolute values of course adjustments and another that considered the total sum, where left-hand turns are considered negative. The hope for these features would be to provide a differentiator between "zig-zag" and "circling" behaviours. The time duration of the trajectory cannot be guaranteed during creation due to missing or erroneous data. To avoid misrepresentative data the features are normalised by dividing the feature by the number of hours between the start and end times of the trajectory. Fig. 8 shows an example of a trajectory with variables that are

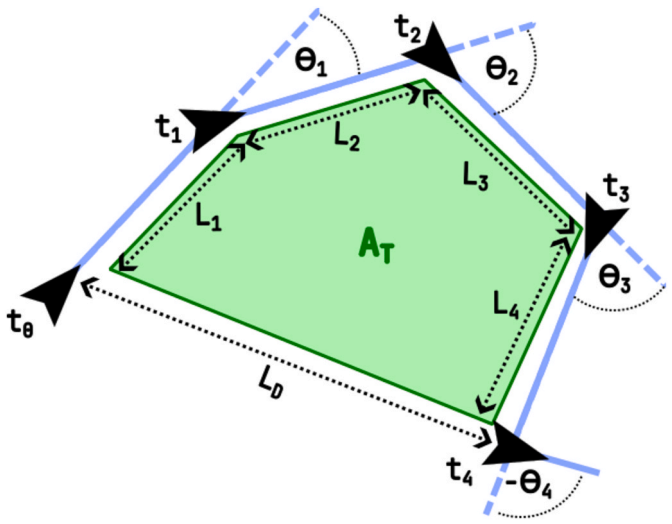


Fig. 8. Example trajectory showing a sequence of Position Reports from a single vessel. Note that left hand turns are considered negative turns.

Table 3
Features derived from trajectory per vessel.

Name	Description	Equation	Units
Normalised Trajectory Length	The length of a trajectory divided by the time duration.	$\overline{L} = \frac{\sum_0^n L_n}{t_n - t_o}$	Degrees/ Hour
Normalised Trajectory Distance	The shortest distance between the start and end points of a trajectory	$\overline{L_D} = \frac{L_D}{t_n - t_0}$	Degrees/ Hour
Normalised Trajectory Area	The area covered by the convex hull of the trajectory	$\overline{A_T} = \frac{A_T}{t_n - t_0}$	Degrees^2/ Hour
Normalised Sum of SOG changes	The sum of the differences in speed between Position Reports.	$\overline{F_{SOG}} = \frac{\sum_0^n (SOG_n - SOG_{n-1})}{t_n - t_0}$	Knots/ Hour
Normalised Sum of COG changes	The sum of the differences in COG between Position Reports.	$\overline{F_{COG}} = \frac{\sum_0^n \theta_n}{t_n - t_0}$	Degrees/ Hour
Normalised Sum of absolute course changes	The sum of the absolute value of differences in course between Position Reports.	$\overline{F_{ACOG}} = \frac{\sum_0^n \lvert\theta_n\rvert}{t_n - t_0}$	Degrees/ Hour
Average COG	COG averaged over the duration of the trajectory	$\overline{COG} = \frac{\sum_0^n COG_n}{n}$	Degrees
Average SOG	SOG averaged over the duration of the trajectory.	$\overline{SOG} = \frac{\sum_0^n SOG_n}{n}$	Knots
Variance in COG	Variance in COG over the duration of the trajectory.	$S^2_{COG} = \frac{\sum_0^n (COG_n - \overline{COG})^2}{n-1}$	Degrees^2
Variance in SOG	Variance in SOG over the duration of the trajectory.	$S^2_{SOG} = \frac{\sum_0^n (SOG_n - \overline{SOG})^2}{n-1}$	Knots^2

associated with the feature equations and Table 3 shows the equations and units that describe the trajectory based features.

4.3. Voyage Report features

AIS Voyage reports are considered to be the AIS messages that include data describing the reporting vessel, the ETA and destination for the current voyage and consists of the AIS message types 5, 19 and 24.

Table 4
Features extracted from Voyage Reports.

Name	Description	Units
Distance to Bow	Distance between GPS antenna and front of ship in metres.	Metres
Distance to Stern	Distance between GPS antenna and rear of ship in metres.	Metres
Distance to Port	Distance between GPS antenna and port side of ship in metres.	Metres
Distance to Starboard	Distance between GPS antenna and starboard side of ship in metres.	Metres

Voyage Report features are not as reliable as those derived from Position Reports due to being configured by the sailors on board. There are several studies that characterise, and correct, the inaccuracies of reported class and physical dimensions (Harati-Mokhtari et al., 2007), (Meyers et al., 2022) but the inclusion of these features into the algorithm, without any corrections applied, has improved accuracy. The four size variables are used to describe the location of the GPS antenna, the source of the vessels' location information, in relation to the outside dimensions of the vessel (IALA Guideline, 2016). These variables are often collapsed into a set of simple "width" and "breadth" variables but this removes some limited information describing the structure of the vessel. There is an incentive to mount the antenna on a high spot near the bridge and so the four size variables can, at some times, describe the location of the bridge relative to the edges of the ship. Table 4 described the Voyage Report features.

5. Results

5.1. Feature importance

The importance of the features was measured using the "split" and "gain" methods. The split method counts the number of times each feature was used in the model while the gain method calculates the improvement in accuracy by including the feature in the model. Fig. 9 shows the feature importance for both measurements and for both classification targets.

Fig. 9 shows that the most important features, for both models, would be the static features described in the Voyage Reports and the vessel's latest position. There is also agreement between the two measurement methods on which features are of the least importance; the instantaneous measurement of SOG and COG and some features derived from the trajectory.

Table 5 shows the accuracy of the algorithm when using all the features and subsets of the features as well as the time to train a single classifier using a modern commercial laptop computer. This time to train would be significantly higher when performing cross validation and hyper parameter optimization but a single training loop was chosen as this was quicker to test for and would be correlated to the time required for a full cross-validation, hyperparameter grid search test.

This table shows that including all the features, even those with a low feature importance, improves the performance of the algorithm. The performance of the LightGBM algorithm results in an insignificant cost to train using extra features. An additional step was taken where the CatBoost algorithm, a gradient boosting algorithm that optimises accuracy over training time (Bentéjac et al., 2021), was used in place of the LightGBM algorithm. The performance of the trained CatBoost was no better than the LightGBM algorithm while the time to train a single classifier was almost 5 times more.

Fig. 10 shows a confusion matrix for an algorithm trained to predict the Comparative Labels from Position Report and Trajectory feature sets. A confusion matrix shows the relationship between predicted and actual classes for algorithm predictions made from the testing dataset. A perfect classifier would have a normalised confusion matrix that was an identity matrix.

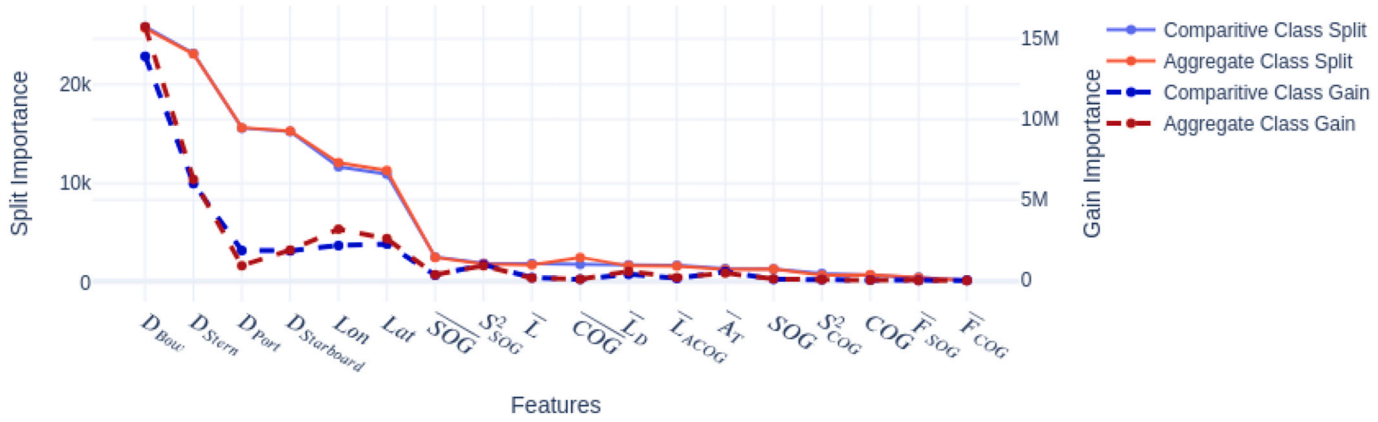


Fig. 9. Feature importance measured by gain and split.

Table 5
Algorithm performance by feature subsets.

Features Included in Model	Precision	Recall	F1 Score	Time To Train
All	0.90	0.90	0.90	30.4 s ± 997 ms
Position Report only	0.71	0.67	0.68	19 s ± 926 ms
Trajectory Derived only	0.66	0.61	0.62	27 s ± 902 ms
Voyage Report only	0.86	0.86	0.86	17.3 s ± 728 ms
Position + Trajectory	0.74	0.71	0.72	27.2 s ± 907 ms
Position + Voyage	0.89	0.89	0.89	20.7 s ± 911 ms
Trajectory + Voyage	0.88	0.88	0.88	23.9 s ± 603 ms
All Features with CatBoost Classifier	0.90	0.90	0.89	2min 25s ± 9.81 s

The drop in performance between the full feature set and subsets seems to be mainly due to the algorithm being unable to separate the two largest classes; Cargo and Tanker vessels. Performance on fishing vessels is still good, although not as good as algorithms trained on the full feature set.

Table 6 shows the best LightGBM parameters for both classification targets found using the parameter grid search method.

5.2. Comparative AIS class aggregation

Table 7 and Fig. 11 show the results when using only the limited classes used in previous research papers (Wang et al., 2021). This Classification Target was chosen to allow a comparison between this work and previous research papers that had a similar dataset. The results are compared by presenting a confusion matrix of results and published results pulled from multiple places in the paper.

Table 6
Parameter grid search results for classification targets.

	Comparison Labels	Aggregated Labels
Number of Leaves	180	200
Maximum Depth	15	15
Class Weight	Balanced	Balanced
Learning Rate	0.05	0.05
Number of Estimators	150	150

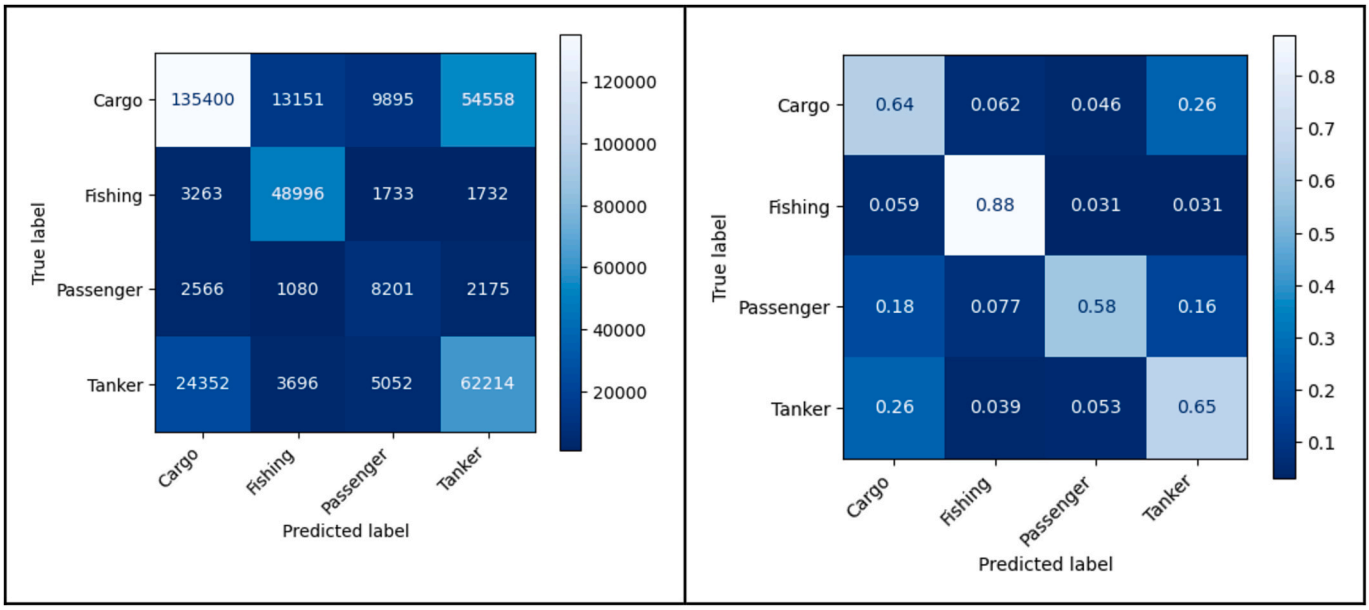


Fig. 10. Comparative Class confusion matrix trained with Position and Trajectory feature sets only.

Table 7
Comparison between published results from (Wang et al., 2021) and this study.

	Literature Results		Comparative Results	
Class	F1 Score	Unique Ships	F1 Score	Unique Ships
Passenger	0.8889	189	0.66	149
Tanker	0.9322	738	0.93	1539
Fishing	0.9355	674	0.86	575
Cargo	0.875	2161	0.92	3283
Weighted Average F1 Score	0.9		0.89	

5.3. AIS class aggregation

The results of the LightGBM when used with aggregated AIS class labels is shown below with confusion matrices, Fig. 12, and scoring metrics in Table 8.

6. Discussion

The results show that a classification algorithm, when trained with position and trajectory based features, is able to perform near, or better, than results from published complex ensemble classification algorithms in overall accuracy.

This work details one method of classifying vessels from AIS data covering a significant portion of the globe. This was compared to a similar published work, Wang et al. (2021), that obtained their vessel classification results by creating an ensemble algorithm combining

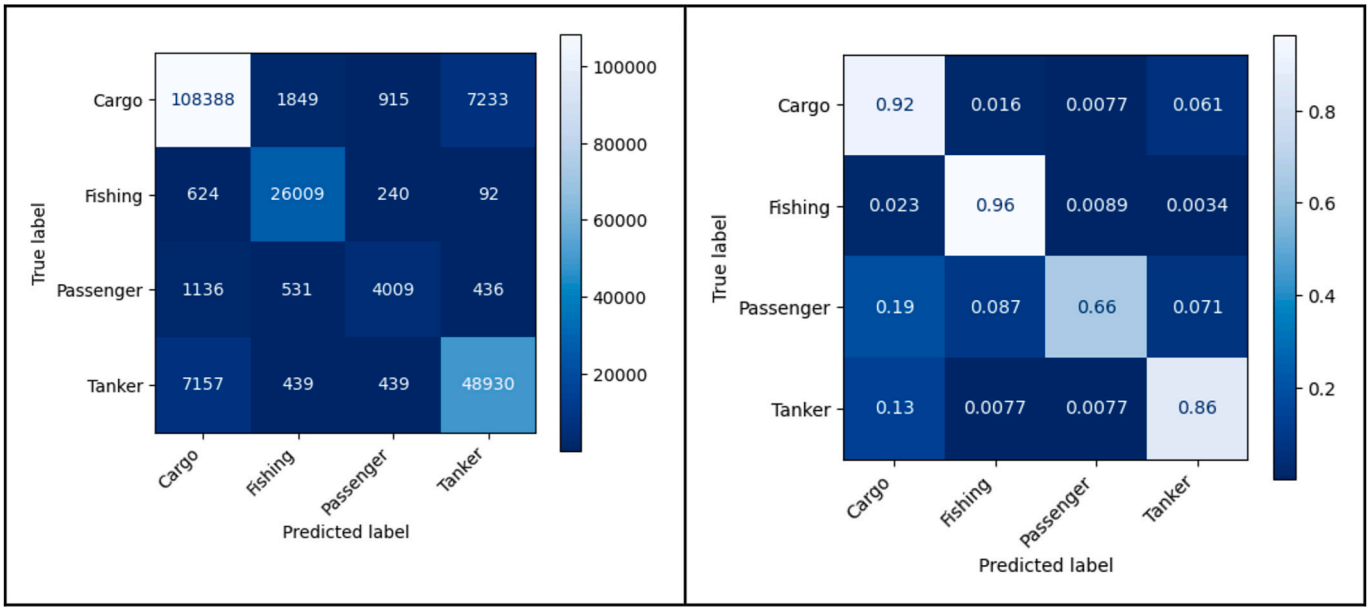


Fig. 11. Comparative Class confusion matrix.

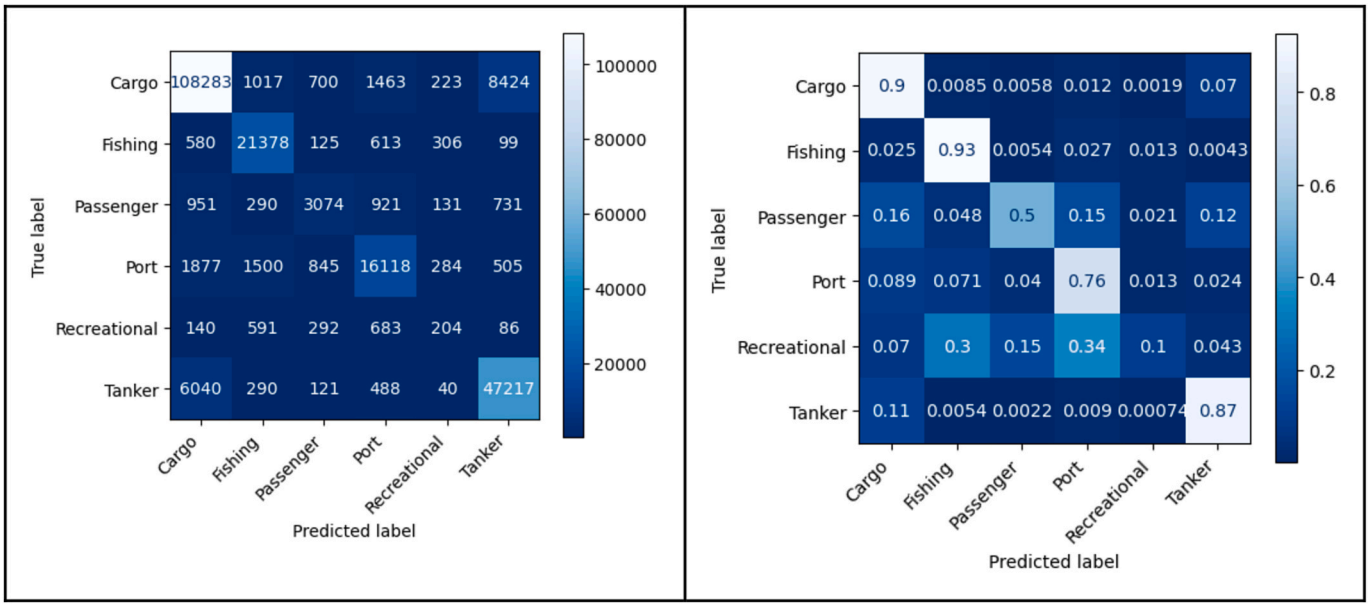


Fig. 12. Aggregate Class confusion matrix.

Table 8
Algorithm performance with aggregate class labels.

	Precision	Recall	F1 Score	Support
Cargo	0.93	0.9	0.91	140703
Fishing	0.9	0.93	0.91	32961
Passenger	0.57	0.55	0.56	7063
Port	0.82	0.84	0.83	27530
Recreational	0.24	0.21	0.23	1755
Tanker	0.85	0.88	0.87	67478
Accuracy			0.88	277490
Macro Avg	0.72	0.72	0.72	277490
Weighted Avg	0.88	0.88	0.88	277490

XGBoost, Random Forest, Convolutional Neural Network, and GRU Recurrent Neural Network classifiers. Table 7 shows this work achieved similar results by using a single classification algorithm combined with more complex feature engineering. Using a single open-source algorithm allows future users to quickly recreate the methods described in this paper and for this work. LightGBM has been shown to be a quick and accurate algorithm for classification tasks from multiple fields (Liao et al., 2022), (Monteiro et al., 2024), (Ke et al. et al., 2017). Table 5 shows that it is quick to iterate through different feature sets to investigate the contributions of specific features. Table 5 also shows that similar accuracy results could be obtained using CatBoost but the training time was approximately 5 times longer. Figs. 10–12 show that the results of the classifier are strongly influenced by the choice of classification target. Given that different MDA users would have different goals and be focused on different vessel types and behaviours, the ability to quickly retrain with different classification targets is vital to building MDA tools.

The differences in class performance is made up by the LightGBM algorithm performing better than the literature when classifying the largest class, Cargo vessels. The training and testing was also used without any spatial filters, such as removing vessels close to the coastline or ports, and without altering the class distribution. Single class performance might be improved by using stratified classes in future work but this might result in worse overall performance.

The feature importance plot shows that the most important variables in the classification process are the vessel's self reported dimensions, location and some variables describing its behaviour over the last 12 h. The Voyage Report variables were used without any filtering, or checking with third party data sources. This seems to indicate that the majority of self reported data is accurate enough to perform classification and that future investigations into vessel classification would benefit from using time series methods.

As expected cargo and tanker vessels have an overlap in physical structure and behaviour that lead to them being misclassified as each other. Fishing vessels also have high accuracy which would be of use to fishery enforcement officials. Recreational and passenger vessels have

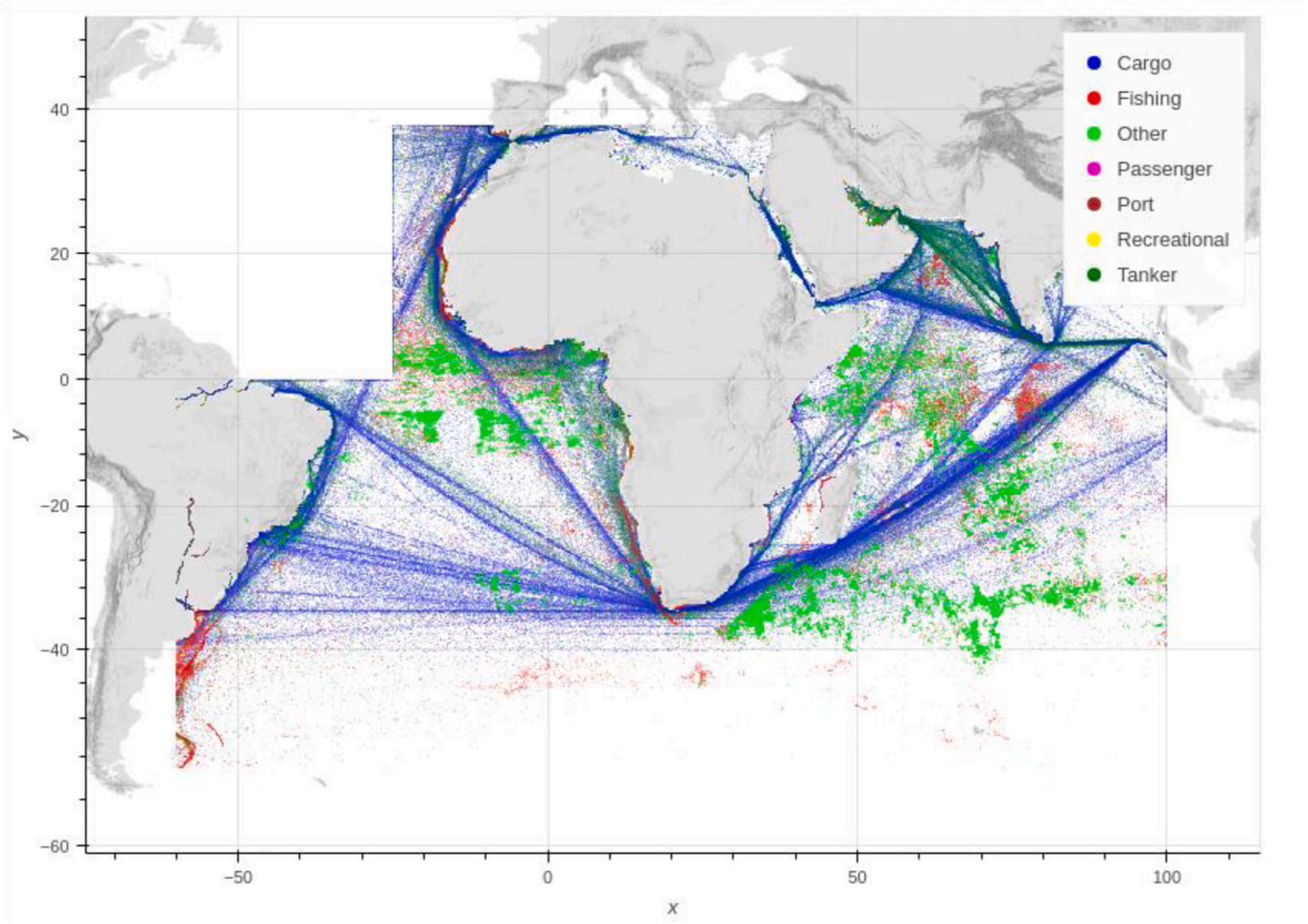


Fig. 13. Distribution of vessels with known labels, coloured by reported class.

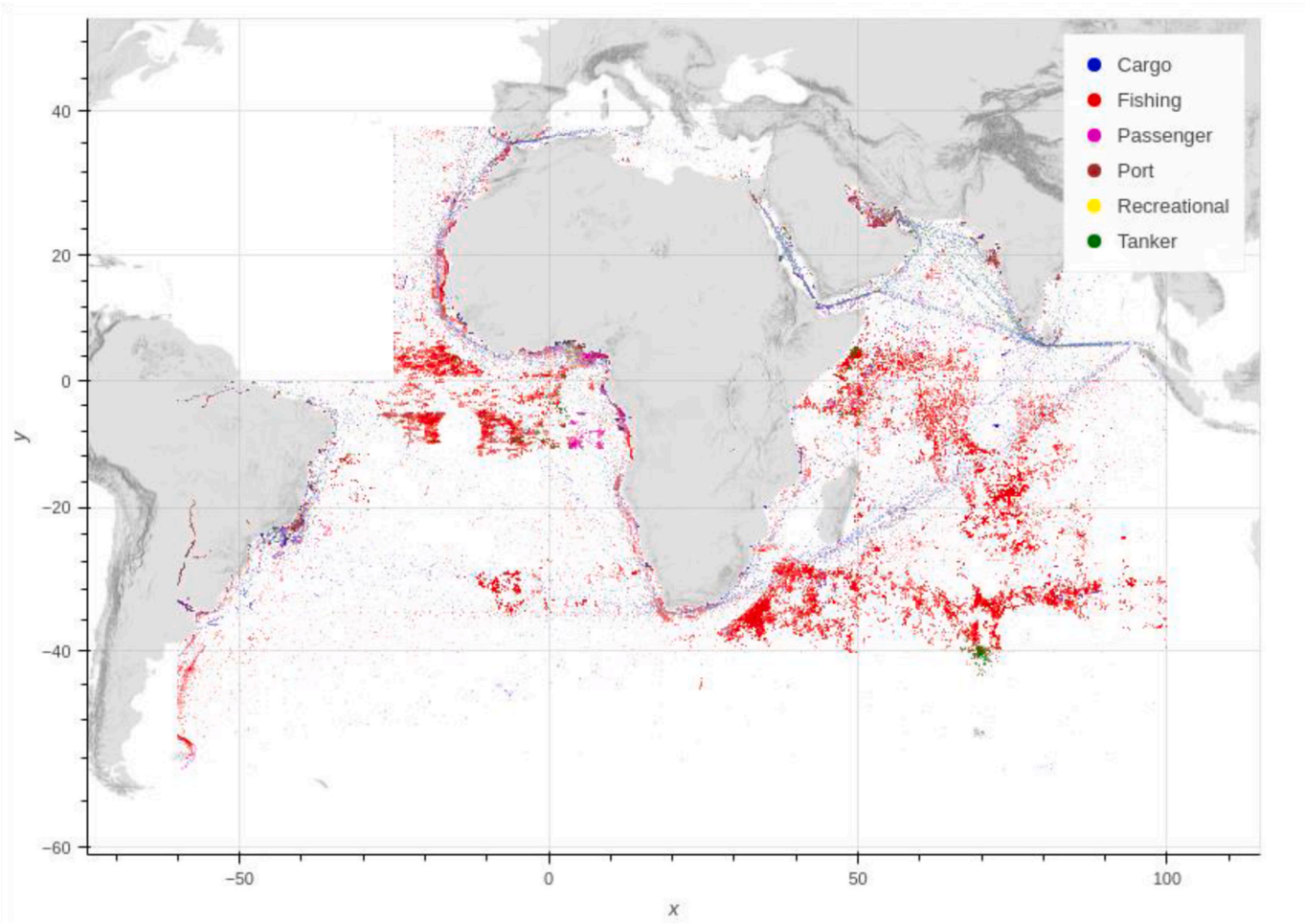


Fig. 14. Distribution of unknown vessels, coloured by algorithm classification results.

the smallest number of samples and lowest accuracy of the classes. It is likely that improvements could be made by developing specific classifiers for smaller classes like recreational and passenger vessels.

There is a clear difference in performance between the algorithm in this study and with (Wang et al., 2021) with regards to recreational and passenger vessels. The "Global Shipping Traffic Density" dataset, published by the World Bank (CerdeiroKomaromiLiuSaeed), shows that the bulk of global recreational and passenger vessels are located in Europe and North America and therefore it is not surprising that the OCIMS dataset would have a reduced representation of this class due to the lack of coverage in those areas. The difference between the proportion of samples, and therefore class weighting and performance, can be attributed to the difference in geographic coverage between the two datasets used in this work and (Wang et al., 2021).

6.1. Unknown classes

Fig. 13 shows the database plotted with classes represented by colour. There are a large number of unknown classes, represented by light green. Running these unknown class messages through the classifier resulted in a large portion of them being labelled as fishing vessels. When plotting these samples onto a map, Fig. 14, they followed a distribution and pattern similar to those of fishing vessels but are much denser. One possibility for this large number of predicted fishing vessels is AIS transmitters being used as net markers. This would explain the shared location and distribution pattern of the unknown vessels but more work is needed to characterise and separate AIS messages of fishing vessels from fishing gear.

Other predicted classes are mostly limited to coastal areas, indicating that near coastal vessel traffic might be more complex than open ocean traffic.

6.2. Feature choice and algorithm performance

Table 5 and Fig. 9 indicate that the features that contribute most to algorithm performance are vessel dimensions, obtained from Voyage Reports, followed by vessel location. Testing showed the trajectory feature set provided some small improvement in accuracy when added to any other feature sets. While this improvement does not seem too significant it should be mentioned that the cost of generating and adding the additional data is low: it can be generated with minimal cost during the database query to fetch all other features and the training cost was insignificant when compared to the cost of using a less performant algorithm.

This work shows that combining more complex feature engineering with a performant algorithm is able to obtain similar performance to more complex existing techniques. In a real time production environment, where retraining might be required sporadically, there would be no benefit to fast training once a specific feature set, algorithm and parameters have been identified. The value in the LightGBM algorithm's performance was to quickly iterate through ideas and processing steps to identify features and parameters that could provide good performance. Researchers were able to train and test different feature engineering ideas, using nearly a year of AIS data covering a significant portion of the globe, using a commercial laptop. This speed of iteration allowed the researchers to avoid the costs involved with using cloud compute or

server infrastructure.

It must be noted that Voyage Reports can also be a significant source of errors (Jankowski et al., 2021). While this study does not address the source of errors, or impact that potential errors could have on the algorithm accuracy it does provide some insight into the value that Voyage Report features could have on classification tasks and provides a starting point for future work into class improving classification based solely on position, trajectory and rotation of detected vessels (Meyer, 2017). To improve overall accuracy without using potentially erroneous feature sets would require additional features that would improve separation between Cargo and Tanker classes.

7. Conclusion

This study has examined using LightGBM as an algorithm for determining a vessel's class from vessel transponder data. This data was collected from a commercial supplier and covered approximately a third of the world's surface, mostly in the Southern Hemisphere. The classification results were compared with a published study that made use of a similar dataset, covering a larger area, and achieved similar results while using a simpler and computationally efficient implementation with more complex feature engineering.

Results have shown that an overall F1 score of 0.88–0.9 with a short training time was achievable. This has allowed various features and parameters to be iterated over or for the algorithm to be retrained using newer data. This score was obtained after performing a grid search parameter optimization with a k-fold cross validation on the dataset after it was split into training and testing datasets based off of a vessel's ID. No vessel appeared in both the training and testing dataset.

The most valuable feature when determining the class of a vessel was shown to be the self-reported physical dimensions of the vessel; distance between the GPS antenna and edges of the ship. This was followed by the position information of the vessel and historical trajectory derived features. This would indicate that to replicate the results using a different dataset would require deriving features that describe the location and dimensions of a vessel.

The algorithm was most accurate when classifying Tanker, Cargo and Fishing vessels while being not as accurate with Passenger and Recreational vessels. This is likely due to the low proportion of these vessels in this dataset, due to the low proportion of these vessels being present in the region covered by the dataset.

The choice of class labels and prefiltering of samples can have a large impact on the score of the algorithm and further work needs to be done to determine accurate labels given that the set of possible labels is limited by the AIS algorithm and may not cover all physical vessel classes or behaviours. More work is required to determine an accurate class label that would be indicative of a vessel's physical construction, behaviour, or legal status.

CRediT authorship contribution statement

Rory Meyer: Writing – original draft, Visualization, Software, Formal analysis, Conceptualization. **Waldo Kleynhans:** Writing – review & editing, Supervision, Resources.

Declaration of competing interest

The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: Rory Meyer reports administrative support was provided by CSIR. If there are other authors, they declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgments

Special thanks to the OCIMS project for the data used in this study and the CSIR for their support in this work.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.oceaneng.2024.120043.

Data availability

The authors do not have permission to share data.

References

Balci, M., Pegg, R., 2006. Towards global maritime Domain awareness ¿ "recent developments and challenges". In: 2006 9th International Conference on Information Fusion. IEEE, Florence, pp. 1–5. https://doi.org/10.1109/ICIF.2006.301702.

Balduzzi, M., Pasta, A., Wilhoit, K., 2014. A security evaluation of AIS automated identification system. In: Proceedings of the 30th Annual Computer Security Applications Conference. New Orleans Louisiana USA: ACM, pp. 436–445. https://doi.org/10.1145/2664243.2664257.

Bentéjac, C., Csörgő, A., Martínez-Muñoz, G., 2021. A comparative analysis of gradient boosting algorithms. Artif. Intell. Rev. 54 (3), 1937–1967. https://doi.org/10.1007/s10462-020-09896-5.

Cerdeiro, Komaromi, Liu, Saeed. Global shipping traffic density. https://datacatalog.worldbank.org/search/dataset/0037580/Global%20Shipping%20Traffic%20Density?version=5,WBDataCatalog. (Accessed 18 January 2023).

Chen, X., Liu, Y., Achuthan, K., Zhang, X., 2020. A ship movement classification based on Automatic Identification System (AIS) data using Convolutional Neural Network. Ocean Eng. 218, 108182. https://doi.org/10.1016/j.oceaneng.2020.108182.

Feurer, M., Hutter, F., 2019. Hyperparameter optimization. In: Hutter, F., Kotthoff, L., Vanschoren, J. (Eds.), Automated Machine Learning. Springer International Publishing, Cham, pp. 3–33. https://doi.org/10.1007/978-3-030-05318-5_1. Eds., in The Springer Series on Challenges in Machine Learning.

Grandini, M., Bagli, E., Visani, G., 2020. Metrics for Multi-Class Classification: an Overview. https://doi.org/10.48550/ARXIV.2008.05756.

Harati-Mokhtari, A., Wall, A., Brooks, P., Wang, J., 2007. Automatic identification system (AIS): data reliability and human error implications. J. Navig. 60 (3), 373–389. https://doi.org/10.1017/S0373463307004298.

Jankowski, D., Lamm, A., Hahn, A., 2021. Determination of AIS position accuracy and evaluation of reconstruction methods for maritime observation data. IFAC-PapersOnLine 54 (16), 97–104. https://doi.org/10.1016/j.ifacol.2021.10.079.

Kabir, M., Kang, M.J., Wu, X., Hamidi, M., 2022. Study on U-turn behavior of vessels in narrow waterways based on AIS data. Ocean Eng. 246, 110608. https://doi.org/10.1016/j.oceaneng.2022.110608.

Kabir, M.M., Toosi, G., Wu, X., Zaloom, V.A., 2024. Study of ship entrance delays to deep draft channels. Ocean Eng. 312, 119104. https://doi.org/10.1016/j.oceaneng.2024.119104.

Ke, G., et al., 2017. LightGBM: a highly efficient gradient boosting decision tree. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc. [Online]. Available: https://proceedings.neurips.cc/paper_files/paper/2017/file/6449f44a102fde848669bdd9eb6b76fa-Paper.pdf

Liao, H., Zhang, X., Zhao, C., Chen, Y., Zeng, X., Li, H., 2022. LightGBM: an efficient and accurate method for predicting pregnancy diseases. J. Obstet. Gynaecol. 42 (4), 620–629. https://doi.org/10.1080/01443615.2021.1945006.

Lundberg, S.M., Erion, G.G., Lee, S.-I., 2018. Consistent Individualized Feature Attribution for Tree Ensembles. https://doi.org/10.48550/ARXIV.1802.03888.

Meyer, R.G.V., 2017. Classification of Ocean Vessels from Low Resolution Satellite SAR Images. Dissertation, University of Pretoria [Online]. Available. https://repository.up.ac.za/handle/2263/66224. (Accessed 10 April 2023).

Meyers, S.D., Yilmaz, Y., Luther, M.E., 2022. Some methods for addressing errors in static AIS data records. Ocean Eng. 264, 112367. https://doi.org/10.1016/j.oceaneng.2022.112367.

Monteiro, P., Lino, J., Araújo, R.E., Costa, L., 2024. Comparison between LightGBM and other ML algorithms in PV fault classification. EAI Endorsed Trans. Energy Web 11 (Jan). https://doi.org/10.4108/ew.4865.

Paudel, S., Toosi, G., Wu, X., Zaloom, V.A., 2024. Study on utilization of Inland deep-draft waterway based on ship trajectories: applied to Sabine-Neches Waterway. Ocean Eng. 298, 117038. https://doi.org/10.1016/j.oceaneng.2024.117038.

Rintoul, M.D., Wilson, A.T., 2015. Trajectory analysis via a geometric feature space approach. Stat. Anal. Data Min. ASA Data Sci. J. 8 (5–6), 287–301. https://doi.org/10.1002/sam.11287.

Sanchez Pedroche, D., Amigo, D., García, J., Molina, J.M., 2020. Architecture for trajectory-based fishing ship classification with AIS data. Sensors 20 (13), 3782. https://doi.org/10.3390/s20133782.

Tang, W., Cha, H., Wei, M., Tian, B., 2019. The effect of atmospheric ducts on the propagation of AIS signals. Aust. J. Electr. Electron. Eng. 16 (2), 111–116. https://doi.org/10.1080/1448837X.2019.1622491.

Wang, Y., Yang, L., Song, X., Chen, Q., Yan, Z., 2021. A multi-feature ensemble learning classification method for ship classification with space-based AIS data. Appl. Sci. 11 (21), 10336. https://doi.org/10.3390/app112110336.

Zhou, Y., Daamen, W., Vellinga, T., Hoogendoorn, S.P., 2019. Ship classification based on ship behavior clustering from AIS data. Ocean Eng. 175, 176–187. https://doi.org/10.1016/j.oceaneng.2019.02.005.

IALA guideline-an overview of AIS, edition 2.0, 2016. https://www.navcen.uscg.gov/pdf/IALA_Guideline_1082_An_Overview_of_AIS.pdf.

International Convention for the Safety of Life at Sea', 1974. IMO [Online]. Available: https://www.refworld.org/docid/46920bf32.html.