Machine Learning Models for Vessel Traffic Flow Forecasting: An Experimental Comparison

Petros Mandalis
Dept. of Statistics and Ins. Sci.
University of Piraeus
Piraeus, Greece
pmandalis@unipi.gr

Eva Chondrodima
Dept. of Informatics
University of Piraeus
Piraeus, Greece
evachon@unipi.gr

Yannis Kontoulis
Dept. of Informatics
University of Piraeus
Piraeus, Greece
ikontoulis@unipi.gr

Nikos Pelekis
Dept. of Statistics and Ins. Sci.
University of Piraeus
Piraeus, Greece
npelekis@unipi.gr

Yannis Theodoridis
Dept. of Informatics
University of Piraeus
Piraeus, Greece
ytheod@unipi.gr

***Abstract*—Within the last years the shipping industry investments continue to grow to improve maritime transport systems. A vital part of the maritime transport systems is the accurate Vessel Traffic Flow Forecasting (VTFF). In this paper, we approach the VTFF problem from two different perspectives: a) indirect - as a vessel route forecasting application via employing predicted vessels locations in the future, and b) direct - as a flow sequence forecasting problem. In both strategies, machine learning methods are employed because they can leverage from the massive vessel surveillance information to enable deeper digitalization in the shipping industry. This work performs an experimental comparative study between the two approaches over a real dataset from the maritime domain.**

***Index Terms*—Machine Learning, Maritime data, Vessel Traffic Flow Forecasting, Route Forecasting**

I. Introduction

In the maritime domain, it is of great importance to ensure the safe and efficient sailing of vessels. Thus, the shipping industry investments continue to grow to improve Maritime Transport Systems (MTS) and better monitor and understand maritime transport and vessel traffic. Vessel Traffic Flow Forecasting (VTFF) is a crucial requisite for maritime navigation. More specifically, forecasting vessel flows is vital for maritime authorities to alleviate congestion, enhance maritime safety, and assist individual vessel users in planning their routes.

In the literature, the methods used in traffic flow prediction can be classified into two major categories, either parametric or non-parametric [1], [2]. The former category includes time series-based models, which examine periodical patterns and mostly rely on a priori distribution [2], such as Autoregressive Integrated Moving Average (ARIMA) [3]. On the other hand, the non-parametric methods are based on knowledge derived from massive historical data [4], such as Neural Network (NN) models [5]. The most promising approaches are based on data-driven techniques, which combine data analysis and non-parametric methods. These techniques are free of assumptions related to the underlying model formulation and the uncertainty involved in estimating the model parameters [4].

Furthermore, attempts on forecasting traffic flow, such as [2], [6], [7], mostly use grid-based representation analysis [8], which can be broadly classified into two categories: a) indirect VTFF methods, and b) direct VTFF approaches. Regarding the former category, one way to tackle the VTFF problem is by applying grid-based analysis [6] over future vessel locations predicted by using Vessel Route Forecasting (VRF) techniques [9]–[11]. Another possible way to tackle the VTFF problem is by approaching it directly, i.e. using sequence analysis along with grid-based processing [4]. In particular, the number of vessels at a specific cell at the grid can formulate the sequence of the vessels' flow, which can be predicted by using Machine Learning (ML) methods.

In this paper, we approach the VTFF problem by examining both the above strategies using ML methods. We also benefit from the available massive amounts of Automatic Identification System (AIS) data [12] that facilitate us train ML models efficient. Currently in the literature, there do exist works that tackle the VTFF problem by using ML methods [2], [4], [6], [7], [13]–[16]. However, these studies mainly focus on the vessels' behaviour in specific places of interest, such as rivers, ports, bridges, etc. On the other hand, our approach aims to be of general purpose, covering open sea as well.

In summary, the main contribution of our study is to examine different perspectives of addressing the VTFF problem and using the most popular ML methods to provide comparison results based on real AIS data. This work could be used as a reference for MTS to comprehend an effective VTFF strategy and ease the realization of the ML methods inside operational MTS.

The rest of this paper is organized as follows: Section II discusses related work; Section III formulates the problem at hand and presents the proposed methodology; Section IV describes the available AIS data, presents the experimental setup and the respective findings, and compares the performance of different solutions; Section V concludes the paper and discusses future extensions.

II. Related Work

The current status of state-of-the-art methods concerning VTFF is presented in [17]. In the following paragraphs, we briefly present the state of the art of the ML methods related to the VTFF problem.

In [4], a VTFF algorithm was proposed for multi-bridge water areas based on regression analysis and Kalman filtering. The model performance was evaluated by taking into account AIS data derived from four specific sections in the Yangtze River, China.

Wang et.al. [2] proposed a multiple hexagon-based convolutional NN model that enables zonal traffic flows prediction for areas that are highly desirable. This method was evaluated by using data derived from the South Atlantic States region.

In [14], an improved particle swarm optimization back propagation prediction model was established to predict the total vessel traffic flow in a designated port area of Los Angeles, USA, by using AIS data obtained from MarineCadastre.gov.

In [15], a NN-based model was proposed for predicting traffic of a caution area. The model receives as input vessel movement and vessel attribute information and produces as output the predicted number of ships in the caution area in future time points. The model was applied to a real AIS sensor dataset from the port of Yeosu, South Korea.

Zhang et al. [16] proposed an integrated Support Vector Machine (SVM) and Genetic Algorithm (GA) model to predict traffic for narrow water passage. The model receives as input basic trajectory-related information (position, speed and course) and was validated over real data from Ningbo Port, China.

In [18] a NN-based model was introduced for traffic flow prediction in inland waterway of Wuhan Yangtze River, China. The traffic flow sequences were preprocessed to remove trend and seasonality and a time-window method was applied to provide multi-step lag observations as features to increase temporal correlation.

The abovementioned research works have shown that valuable knowledge from vessels' behaviour can be extracted through the analysis of historical data. As already mentioned, related work has focused on specific places of maritime interest, while our work proposed a more general approach to the VTFF problem.

In the literature, there are also grid-based VTTF methods. In [6], deep learning-based methods were proposed to forecast the inflow and outflow of vessels within the Singapore area, whereas [7] employed a similarity analysis-based traffic prediction model, and the Sorenson similarity index to measure its accuracy, applied to a real AIS dataset in the Strait of Georgia, USA.

III. Problem Formulation & Proposed Methodology

Consider a maritime dataset D composed of vessel trajectories and each trajectory is a sequence of timestamped locations (t_i, p_i), which consists of timestamp t and location p composed of two coordinate values, x and y, in a Cartesian projection system.

The VTFF problem addressed in this paper is formulated as follows:

- **Given**:
  - a set of vessel trajectories D spanning in Ds (minimum bounding box of locations) in space and D_T in time,
  - a time duration (prediction horizon) Δt,
  - a number of temporal transitions r
  - a spatiotemporal (i.e., 3D) grid that splits a) Ds into grid cells of resolution $G \times G$, and b) $D_T \cup \Delta t$ into r time frames
- **Predict**: the expected number of vessels in each grid cell related to Δt.

As an example, Fig. 1 illustrates a spatiotemporal grid of 4×4 space and 5 frames in time. The four illustrated vessel trajectories evolve over time in a window of (let us suppose) 3 time frames, and the goal of the VTFF problem is to predict the expected number of vessels in each cell of the grid during the 2 future (upper in the figure) time frames. In the following paragraphs we present the proposed VTFF approaches.

A. VRF-based VTFF

In this approach, it is obvious that the VTFF accuracy depends on the prediction performance of the underlying VRF method. In this study, in order to predict future vessels' locations we employ the algorithm presented in [9], where the trained ML model is executed r times to provide predictions for the r transitions that formulate the predicted trajectory.

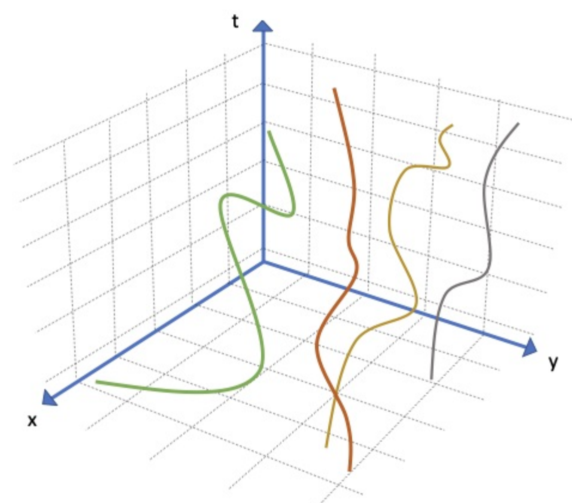


Fig. 1. Overview of a spatiotemporal grid of 4×4 space and 5 frames in time. Coloured lines represent four different trajectories.

In [9], some of the most popular ML methods are being investigated to address the VRF problem, including linear regression, tree-based methods, Support Vector Machine for regression (SVMr) [19], and static and dynamic NNs. Also, the experimental study in [9] proved the dominance of the LSTMs against their rivals. Thus, in this work, we use LSTMs to predict the future vessels' locations (up to Δt) given vessel trajectory coordinates.

In order to employ the algorithm in [9], a few proprocessing steps are necessary: i) stationery simplification (remove records corresponding to speed less than one knot) and insignificant trajectory elimination (remove trajectories composed of very low number of points; less than ten points), and ii) trajectory segmentation into a number of sub-trajectories when the time interval between two consecutive points of the same trajectory exceeds a specific time threshold, equal to 30 min.

The predicted locations resulted by VRF are allocated into the spatiotemporal grid described earlier and the number of vessels located within each cell of the grid is calculated. The resulted numbers represent the volume of vessels, i.e., the traffic flow in the given area and time window.

Regarding the method evaluation and model parameter selection:

1) Method evaluation: Following [9], the processed trajectories are arranged into input and output data and are randomly assigned to three sets according to [20], namely training, validation, and testing, using a 50%–25%–25% percent ratio, respectively. The training set is used to define the parameters for the NN models, while the validation set is used to choose the model. The performance of the selected model is evaluated on the testing set, which is not used during the training and model selection phases and thus can assess generalization capabilities. Experimental results were evaluated by using the Symmetric Mean Absolute Percentage Error (SMAPE), described by Eq. 1, which is a scale-independent accuracy measure, where the lower the SMAPE value, the better the model's accuracy. Also, the Jaccard similarity coefficient, described by Eq.2 , was employed, which in this work, measures the similarity between the set of the actual vessels and the set of the predicted vessels in a specific grid cell.

$$SMAPE = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{F}\sum_{t=1}^{F}2\frac{|y_{b,t}-\hat{y}_{b,t}|}{|y_{b,t}|+|\hat{y}_{b,t}|} \tag{1}$$

$$Jaccard = \frac{1}{B}\sum_{b=1}^{B}\frac{1}{F}\sum_{t=1}^{F}\frac{|Y_{b,t}\cap\hat{Y}_{b,t}|}{|Y_{b,t}\cup\hat{Y}_{b,t}|} \tag{2}$$

where B and F are the total number of cells and time frames in the grid, respectively; y and $\hat{y}$ represent the actual and the predicted number of vessels, respectively; Y and $\hat{Y}$ are the sets including the actual and the predicted vessels, respectively.

2) Model parameter selection: As far as the NN parameter selection is concerned, both the theoretical and experimental perspectives are considered. Also, an early stopping procedure [21] is applied to the validation set to prevent the NN model from overfitting, i.e. the inability of NN to predict correctly based on the data used during the training stage.

B. Flow sequence-based VTFF

In this approach, the vessels' trajectories are allocated to the spatiotemporal grid to formulate the traffic flow sequence in each cell. Then we feed an ML model with these flow sequences.

The trained model can be used to predict the number of vessels in a specific cell in future time frame Δt. The information relevant for making predictions needs to be within a window of a number of past observations [22]. Particularly, the model input N is composed of a number of vessels n in each time frame t, in the b-th box within the grid of l total boxes can be described by the following equation:

$$\mathbf{N}^b = [\ n^b_{t-l}, ..., n^b_{t-1}, n^b_t\ ] \tag{3}$$

The corresponding model output (predicted number of vessels) in the b-th box grid in the future time frame $t+1$ is: $\hat{n^b}_{t+1}$.

Different ML techniques enable different data representations to be learned [23]. In this study we employ XgBoost [24] and ARIMA models. XgBoost [24] is a sparsity-aware algorithm for sparse data and weighted quantile sketch for approximate tree learning, which allows the handling of large datasets with a scalable tree boosting system. Moreover, ARIMA models are able to capture a suite of different standard temporal structures in time series data. The aforementioned algorithms can be applied immediately on data because they simply map input to output.

Regarding the method evaluation and model parameter selection:

1) Method evaluation: The traffic flow sequences for each grid cell are arranged into input and output data. For each grid cell, the initial 75% of the traffic flow sequence is used for the training purpose, whereas the remaining 25% of the traffic flow sequence, except the last three observations, is organized in the validation set. The last three observations are used for the testing purposes. Experimental results were evaluated by using SMAPE, described by Eq.1. Due to the proposed algorithm's nature, the avaluation is performed only in the busy grid cells, i.e. regions of high traffic areas where regular navigation activities occur and are associated with a high risk of accidents [2].

2) Model parameter selection: As far as the model parameterization is concerned, several aspects of each model type were taken into consideration and adjusted with intermediate experiments before the final performance assessment. Particularly, the XgBoost considers three different types of parameters: general tree parameters, booster parameters, and learning task/miscellaneous parameters. We optimize XgBoost models according to: a) learning rate b) the minimum leaf size for pruning, c) the number of features on a node, and d) the number of regression trees. Also, the ARIMA model considers three different types of parameters: the lag order, the degree of differencing, and the order of the moving average. We optimize ARIMA models according to the above

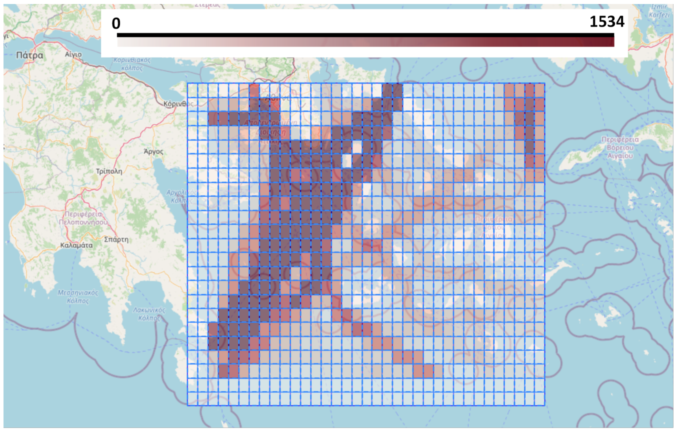


Fig. 2. Overview of the traffic flow for the whole period of November, 2018 for the Aegean-Cyclades dataset, within a spatial grid of $G = 10$km. Darker color indicates higher traffic flow.

three parameters by evaluating the Partial Autocorrelation plot (PACF), the Augmented Dickey Fuller (ADF) test and the Autocorrelation plots (ACF), respectively.

IV. Experimental Study

This section presents the experimental setup, as well as preliminary experimental findings on the performance of the proposed approaches.

A. Experimental setup, dataset description and preparation

The two VTFF methods were implemented in Python3. The machine we used is a workstation with 64 GB RAM, an Intel Core i9-9900KX CPU and a GeForce RTX2080Ti GPU with 11GB graphics memory.

For the purpose of our experimental study, we used a real-world AIS dataset, namely the Aegean-Cyclades dataset provided by MarineTraffic.com. The dataset covers an area in the Aegean Sea defined by latitude in [36...38] and longitude in [23...26], and corresponds to 1,757,440 AIS records broadcasted by 2344 different vessels of various types in November, 2018. Given the fact that the transmission rate of AIS signals varies depending on several parameters, such as the vessel's speed and the type of AIS transponder, the sampling rate (before preprocessing) in this dataset ranges from less than 1 second to several days, with a median value of 2.5 min. Furthermore, the number of AIS messages broadcasted by a vessel varies from 1 to 12801, with a median value of 354. Considering the prerequisite preprocessing phase that is crucial among the core machine learning approaches, the AIS data preprocessing procedure followed in this study includes record deduplication (by removing records at timestamps differing less than one sec.) and noise elimination (by removing records corresponding to speed higher than 50 knots). Fig. 2 presents the traffic flow for the entire period of November, 2018, within a spatial grid of $G = 10$km.

B. Results

As already discussed in Section III, the two alternative VTFF strategies handle differently the AIS data. Although the method evaluation for the two strategies is different, we can still compare their prediction capability in busy regions for certain time horizon.

Table I depicts the results of both VTFF strategies in terms of SMAPE for vessels' future flow forecasting in the 20 most busy grid cells up to $\Delta t = 15$ min. The spatiotemporal grid that was employed separates the maritime region into grid cells of $G = 10$ km and the time window into time frames of 5 min. Regarding the flow sequence-based VTFF strategy prediction results, the XgBoost outperforms the ARIMA. More specifically, the XgBoost model predicts the future traffic flow for the next 5 min about 2.5 times better than the ARIMA model. As far as the VRF-based VTFF strategy is concerned, it predicts the traffic flow up to 10 min 1.5 to 2.5 times better than VTFF with XgBoost. However, the XgBoost model performs better in the 3rd five-minute time frame.

Table II focuses on the VRF-based VTFF strategy and presents results in terms of SMAPE and average Jaccard similarity in all grid cells up to $\Delta t = 15$ min. We employed three different grids of G equal to 5 km, 10 km and 15 km, and $r = 3$ transitions of 5 min each. It is obvious that this method is affected by the granularity of the grid; it predicts better in larger grid cels. As far as the results in terms of the Jaccard similarity are concerned, they range from 0.98 down to 0.78.

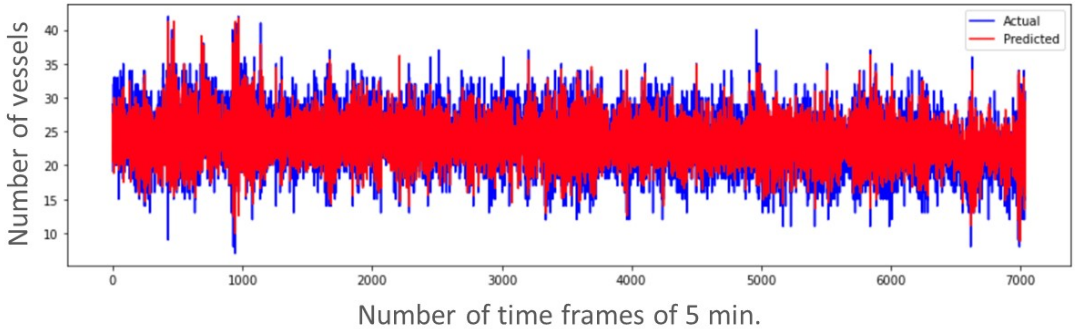
Finally, Fig. 3 presents the actual and the predicted vessel traffic flow per time frame of 5 min., in the training, validation and testing sets, regarding the busiest grid cell. The traffic flow is predicted by the XgBoost model. As expected, the model performs better in the training set than in the validation set. However, it is obvious that the performance in both sets are comparable and there is no sign of overffiting. Finally, in the zoomed area the last three observations used for testing purposes are presented.

C. Discussion

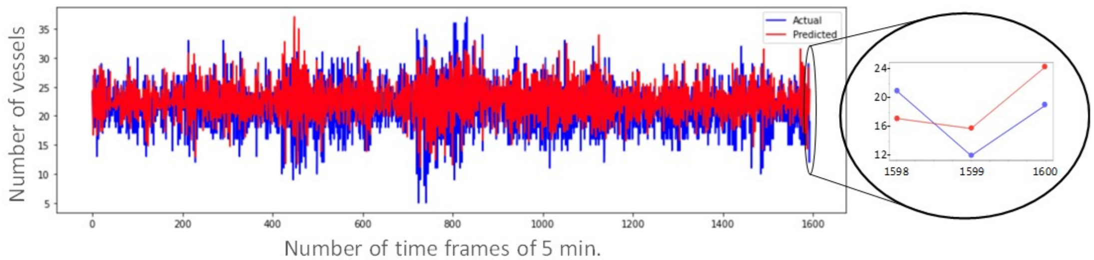
The two alternative VTTF strategies presented in the paper handle the AIS data differently, which affects the forecasted vessel traffic flow. More specifically, the VRF-based VTFF strategy does not take into account vessel records corresponding to low speed or vessel trajectories composed of a very low

TABLE I.
Prediction results (SMAPE) in the testing set (20 busiest grid cells), $G = 10$km .

VTFF strategy	Method	Time prediction horizon (min)		
		5	10	15
Flow sequence-based	XgBoost	17.72	30.41	27.43
	ARIMA	46.94	37.75	48.73
VRF-based	LSTM	6.35	16.76	28.71



(a)



(b)

Fig. 3. Actual vs. predicted vessel traffic flow produced by the VTFF with XgBoost in the busiest grid cell for the: a) training set, and b) validation set and testing set (zoomed graph).

TABLE II.
PREDICTION RESULTS (SMAPE, JACCARD) FOR THE VRF-BASED VTFF STRATEGY IN THE TESTING SET (ALL GRID CELLS) .

Grid cell (km)	Time frame (min)	SMAPE	Jaccard
5	5	9.57	0.95
	10	26.20	0.87
	15	44.00	0.78
10	5	4.97	0.97
	10	14.23	0.93
	15	24.90	0.87
15	5	3.52	0.98
	10	10.08	0.95
	15	18.04	0.91

number of points. As a result, the future vessel flow counts are produced mostly by vessels with significant movement, i.e., it reflects the variable traffic flow anywhere in the grid.

On the other hand, the flow sequence-based VTFF strategy is able to take into account all the available vessel records, even vessel trajectories composed of only one or two records. However, due to its nature, it is able to forecast the traffic volume only within busy regions.

It should be noted that the prediction results for the case of $G = 10$ km grid cells produced by the VRF-based VTFF strategy are different between Table I and Table II. In the former table the SMAPE is calculated by taking into account only the 20 busiest grid cells and in the latter table the SMAPE includes all the available grid cells. It is obvious that the prediction capability of the VRF-based VTFF strategy is affected by the number of non-busy regions; it predicts better when there are also non-busy regions.

V. CONCLUSION

An effective VTFF method is substantial for improving MTS services' quality and reliability. Taking advantage of the wealth of vessel positioning data, this work approaches the VTFF problem from two different perspectives: a) indirect - as a VRF application via employing forecasted vessels locations, and b) direct - as a flow sequence forecasting problem. Through an experimental study of a real AIS dataset, insightful findings are provided, while it is clear that both strategies can efficiently forecast the traffic flow in the maritime area in short times horizon up to $\Delta t = 15$min. In particular for the VRF-based approach, it is able to achieve up to 0.98 Jaccard similarity between the actual and predicted fleet.

Future work includes comparative study with related work, also investigating weather information and further traffic parameters impact on the VTFF problem. We also plan to study the accuracy - and the necessary adjustments - of the proposed methods when applied to a higher prediction time horizon and/or smaller grid cells. Last but not least, and taking into account the large population of today's AIS monitored fleet in conjunction with the inherent locality of the VTFF problem, we aim to research on solutions based on distributed learning.

ACKNOWLEDGMENT

This work was supported in part by EU Horizon 2020 R&I Programme under Grant Agreement No 957237 (project VesselAI, https://vessel-ai.eu), and EU Horizon 2020 R&I Programme under Marie Sklodowska-Curie Grant Agreement No 777695 (project MASTER, http://www.master-project-h2020.eu).

REFERENCES

[1] C. Tang, J. Sun, Y. Sun, M. Peng, and N. Gan, "A general traffic flow prediction approach based on spatial-temporal graph attention," *IEEE Access*, vol. 8, pp. 153 731–153 741, 2020.

[2] X. Wang, J. Li, and T. Zhang, "A machine-learning model for zonal ship flow prediction using ais data: A case study in the south atlantic states region," *J. Mar. Sci. Eng.*, vol. 7, p. 463, 2019.

[3] P. Doganis, E. Aggelogiannaki, and H. Sarimveis, "A combined model predictive control and time series forecasting framework for production-inventory systems," *Int. J. Prod. Res.*, vol. 46, no. 24, pp. 6841–6853, 2008.

[4] W. He, C. Zhong, M. A. Sotelo, X. Chu, X. Liu, and Z. Li, "Short-term vessel traffic flow forecasting by using an improved kalman model," *Cluster Comput.*, vol. 22, no. 4, pp. 7907–7916, 2019.

[5] S. Haykin, *Neural Networks: A Comprehensive Foundation*, 2nd ed. USA: Prentice Hall PTR, 1998.

[6] X. Zhou, Z. Liu, F. Wang, Y. Xie, and X. Zhang, "Using deep learning to forecast maritime vessel flows," *Sensors*, vol. 20, no. 6, p. 1761, 2020.

[7] D. Alizadeh, A. A. Alesheikh, and M. Sharif, "Prediction of vessels locations and maritime traffic using similarity measurement of trajectory," *Annals of GIS*, pp. 1–12, 2020.

[8] D. Nguyen, C. L. Van, and M. I. Ali, "Vessel trajectory prediction using sequence-to-sequence models over spatial grid," in *12th ACM Int. Conf. DEBS*. ACM, 2018, pp. 258–261.

[9] E. Chondrodima, P. Mandalis, N. Pelekis, and Y. Theodoridis, "Machine learning models for vessel route forecasting: An experimental comparison," in *23rd IEEE Int. Conf. MDM*, 2022.

[10] P. Tampakis, E. Chondrodima, A. Tritsarolis, A. Pikrakis, Y. Theodoridis, K. Pristouris, H. Nakos, P. Kalampokis, and T. Dalamagas, "i4sea: a big data platform for sea area monitoring and analysis of fishing vessels activity," *Geo Spat. Inf. Sci.*, 2021.

[11] N. Forti, L. M. Millefiori, P. Braca, and P. Willett, "Prediction oof vessel trajectories from ais data via sequence-to-sequence recurrent neural networks," in *IEEE Int. Conf. ICASSP 2020*, 2020, pp. 8936–8940.

[12] A. Artikis and D. Zissis, *Guide to Maritime Informatics*. Springer, 2021.

[13] Z. Xiao, L. Ponnambalam, X. Fu, and W. Zhang, "Maritime traffic probabilistic forecasting based on vessels' waterway patterns and motion behaviors," *IEEE Trans. Intell. Transp. Syst.*, vol. 18, no. 11, pp. 3122–3134, 2017.

[14] Z.-g. Zhang, J.-c. Yin, N.-n. Wang, and Z.-g. Hui, "Vessel traffic flow analysis and prediction by an improved pso-bp mechanism based on ais data," *Evolving Systems*, vol. 10, no. 3, pp. 397–407, 2019.

[15] K.-I. Kim and K. M. Lee, "Deep learning-based caution area traffic prediction with automatic identification system sensor data," *Sensors*, vol. 18, no. 9, p. 3172, 2018.

[16] H. Zhang, Y. Xiao, X. Bai, L. Chen *et al.*, "Ga-support vector regression based ship traffic flow prediction," *Int. J. Control. Autom.*, vol. 9, no. 2, pp. 219–228, 2016.

[17] Z. Xiao, X. Fu, L. Zhang, and R. S. M. Goh, "Traffic pattern mining and forecasting technologies in maritime traffic service networks: A comprehensive survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 5, pp. 1796–1825, 2020.

[18] Z. Xie and Q. Liu, "Lstm networks for vessel traffic flow prediction in inland waterway," in *IEEE BigComp 2018*, 2018, pp. 418–425.

[19] V. Vapnik, *The nature of statistical learning theory*. Springer science & business media, 2013.

[20] A. Alexandridis, E. Chondrodima, and H. Sarimveis, "Radial basis function network training using a nonsymmetric partition of the input space and particle swarm optimization," *IEEE Trans. Neural Netw. Learn. Syst*, vol. 24, no. 2, pp. 219–230, 2013.

[21] L. Prechelt, "Early stopping - but when?" in *Neural Networks: Tricks of the Trade*, G. B. Orr and K.-R. Müller, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 1998, pp. 55–69.

[22] K. Udeh, D. W. Wanik, N. Bassill, and E. Anagnostou, "Time series modeling of storm outages with weather mesonet data for emergency preparedness and response," in *10th IEEE Conf UEMCON*, 2019, pp. 0499–0505.

[23] M. Veres and M. Moussa, "Deep learning for intelligent transportation systems: A survey of emerging trends," *IEEE Trans. Intell. Transp. Syst.*, vol. 21, no. 8, pp. 3152–3168, 2019.

[24] T. Chen and C. Guestrin, "Xgboost: A scalable tree boosting system," in *22nd ACM SIGKDD*, 2016, pp. 785–794.