A Transformer Network With Sparse Augmented Data Representation and Cross Entropy Loss for AIS-Based Vessel Trajectory Prediction

Duong Nguyen, Ronan Fablet

IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Received 27 November 2023, accepted 18 December 2023, date of publication 4 January 2024, date of current version 14 February 2024.

Digital Object Identifier 10.1109/ACCESS.2024.3349957

RESEARCH ARTICLE

A Transformer Network With Sparse Augmented Data Representation and Cross Entropy Loss for AIS-Based Vessel Trajectory Prediction

DUONG NGUYEN, (Member, IEEE), AND RONAN FABLET, (Senior Member, IEEE)
Lab-STICC, IMT Atlantique, 29238 Brest, France

Corresponding author: Ronan Fablet (ronan.fablet@imt-atlantique.fr)

This work was supported in part by the Publics Funds through Ministère de l'Education Nationale, Ministère de l'Enseignement Supérieur et de la Recherche, Fonds Européen de DEveloppement Régional (FEDER), Région Bretagne, Conseil Général du Finistère, Brest Métropole; in part by High-Performance Computing (HPC) and Graphics Processing Unit (GPU) resources from Azure (Microsoft European Union Ocean awards); in part by Grand Equipement National de Calcul Intensif-Institut du Développement et des Ressources en Informatique Scientifique (GENCI-IDRIS) under Grant 2020-101030; and in part by Agence Nationale de la Recherche (ANR) (France) through ANR Artificial Intelligence (AI) Chair OceaniX (ANR-19-CHIA-0016).

ABSTRACT Vessel trajectory prediction plays a pivotal role in numerous maritime applications and services. While the Automatic Identification System (AIS) offers a rich source of information to address this task, forecasting vessel trajectory using AIS data remains challenging, even for modern machine learning techniques, because of the inherent heterogeneous and multimodal nature of motion data. In this paper, we propose a novel approach to tackle these challenges. We introduce a discrete, high-dimensional representation of AIS data and a new loss function designed to explicitly address heterogeneity and multimodality. The proposed model—referred to as *TrAISformer*—is a modified transformer network that extracts long-term temporal patterns in AIS vessel trajectories in the proposed enriched space to forecast the positions of vessels several hours ahead. We report experimental results on real, publicly available AIS data. *TrAISformer* significantly outperforms state-of-the-art methods, with an average prediction performance below 10 nautical miles up to ∼10 hours.

INDEX TERMS AIS, vessel trajectory, trajectory prediction, maritime surveillance, multimodal data, deep learning, transformer.

I. INTRODUCTION

In the last decades, the development of maritime activities has raised significant concerns relating to Maritime Surveillance (MS) and Maritime Situational Awareness (MSA), with vessel trajectory prediction being a focal point. Anticipating the direction of vessels and their approximate locations at medium-range time horizons, ranging from a few tens of minutes to tens of hours ahead, is at the core of diverse MS and MSA applications, including but not limited to search and rescue [1], [2], traffic control [3], path planning [4], [5], [6], port congestion avoidance [7], [8], [9], pollution monitoring [10].

The Automatic Identification System—AIS provides invaluable information for the monitoring and surveillance of maritime traffic. AIS data provide vessels' kinetic information (the current position indicated by the latitude and longitude coordinates, the current Speed Over Ground—SOG, the current Course Over Ground–COG, *etc.*), the information of the voyages, as well as the static information (the identification number in the format a Maritime Mobile Service Identity—MMSI number, the name of the vessel, *etc.*) of vessels in the vicinity. Vessel trajectory prediction using AIS data has been studied for more than a decade [11], [12], [13], [14], [15], [16], [17], [18]. However, the achievements thus far have been still limited. Most state-of-the-art schemes reach a relevant prediction performance only for short time horizons (ranging from a few minutes to half an hour) [14], [19], or for longer time horizons under particular

The associate editor coordinating the review of this manuscript and approving it for publication was Emanuele Crisostomi.

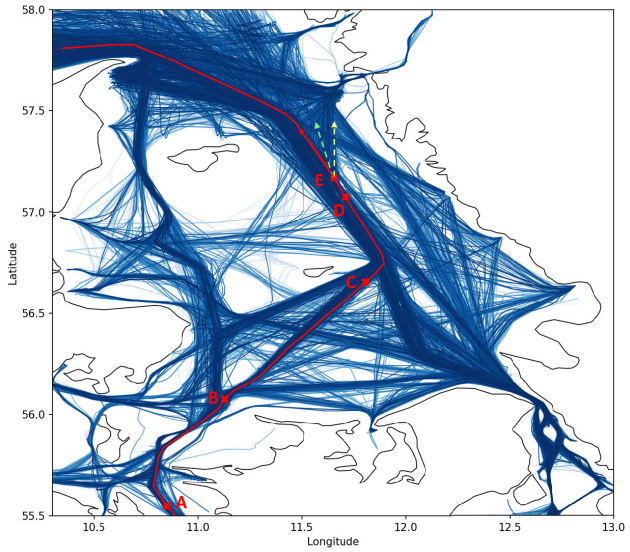

FIGURE 1. Illustration of long-term[1] dependence patterns in AIS vessel trajectories: At E, vessels typically follow one of the two main maritime routes indicated by the red and the yellow dashed arrows. In order to forecast whether a vessel will continue straight ahead (the red path) or turn right (the yellow path), the prediction model may need to roll back several time steps to D, C, B, and A to understand the vessel's previous movements. Moreover, if the prediction model is not multimodal, it may output as a prediction an unusual green dashed path, which is a merged path of the true red and yellow ones.

movement patterns corresponding to predefined maritime routes [20], [21] or unimodal movement patterns [18], [19]. Due to the complexity of vessel movement patterns and the heterogeneous nature of AIS data, forecasting vessel positions above several hours remains highly challenging. As an illustration, we report in Fig. 1 two vessel paths with very similar movement patterns on segment $\mathbf{C}\rightarrow\mathbf{D}$ but heading to two different destinations.

Trajectory prediction has gained significant attention in recent years, particularly in the context of pedestrian and vehicle movement patterns [22], [23], [24], [25]. In this context, deep learning schemes have emerged as the state-of-the-art approach [26], [27], [28], [29], [30]. These recent advances however barely transfer to AIS-based vessel trajectory prediction [31], [32]. First, the targeted space-time scales strongly differ (e.g., meters and a few seconds to a few minutes for pedestrian movements vs. kilometers and hours for vessel movements). Second, while interaction factors are critical to understanding and predicting pedestrian and vehicle trajectories, they have negligible effects on vessels' movements in the open sea. Besides, long-term dependencies are key factors for the latter and need to be explicitly addressed in AIS-based vessel trajectory prediction models.

In the open sea, vessels often follow some common movement patterns in order to optimize fuel consumption and to conform with maritime traffic regulations [33], [34]. However, the analysis of the maritime traffic according to a finite set of interconnected maritime routes using clustering-based approaches [18], [19], [20], [21] appears too simplistic to account for the heterogeneous and multimodal characteristics of real-world AIS data. The core challenge of vessel trajectory prediction for the targeted time horizon in this paper (ranging from half an hour to tens of hours) revolves around accurately predicting the turning direction at the ''intersections''—commonly referred to as *waypoints*—along maritime routes (see Fig. 1). From a mathematical perspective, this involves developing a method that effectively represents the multimodal nature of AIS data at these waypoints, where each turning direction corresponds to a distinct mode of the data distribution. In order to forecast the trajectory correctly, the prediction model may need to backtrack several time steps to know where the vessel comes from and to fully understand large-scale movement patterns. In essence, we can identify two primary methodological challenges: i) learning how to represent maritime traffic flows beyond a fully-structured network of maritime routes; and ii) capturing multi-scale patterns in vessel trajectories.

To address these challenges, we propose a novel deep learning model, referred to as *TrAISformer*. Our key contributions are as follows:

- *TrAISformer* exploits a specific sparse, high-dimensional representation of AIS data and frames the prediction task as a classification problem to explicitly model the heterogeneity of AIS data and the multimodality of vessel trajectories.
- We leverage a probabilistic transformer architecture to capture long-term dependencies in AIS vessel trajectories.
- We benchmark *TrAISformer* w.r.t. state-of-the-art schemes on a real AIS dataset and report a prediction error below 10 nmi (nautical mile) up to 10 hours, which significantly outperforms previous works [14], [18], [19].

The paper is organized as follows. Section II states the problem and gives an overview of the related work and current limitations for AIS-based vessel trajectory prediction. We present the proposed approach in Section III. Section IV details our numerical experiments. We further discuss our main contributions and future work in Section V.

II. PROBLEM STATEMENT AND RELATED WORK

AIS-based vessel trajectory prediction involves forecasting the future positions of vessels over a specific time horizon, using a series of historical AIS data. Formally, let us denote by $\mathbf{x}_t$ an AIS observation at time step t, where $\mathbf{x}_t$ comprises the position of the vessel (indicated by the latitude and the longitude coordinates), its Speed Over Ground—SOG, and its Course Over Ground—COG.[2]

$$\mathbf{x}_t \triangleq [lat, lon, SOG, COG]^T . \tag{1}$$

[1] In this paper, we use the terms ''medium-range time horizon'', ''medium-range forecasting'', *etc*. to indicate the prediction horizons ranging from a few tens of minutes to tens of hours. The terms ''long-term dependency'', ''long-term correlation'', *etc*., on the other hand, indicate the correlations across several time steps in the series.

[2] We let the reader refer to [35] for a more detailed presentation of AIS data streams.

An AIS vessel trajectory is then represented by a series of observations $\{\mathbf{x}_{t_0}, \mathbf{x}_{t_1}, \dots \mathbf{x}_{t_T}\}$ where $t_i < t_j$ if $i < j$. One can use some simple interpolation method to get a series of $T+1$ equally sampled observations $\mathbf{x}_{0:T} \triangleq \{\mathbf{x}_{t_0}, \mathbf{x}_{t_0+\Delta t}, \mathbf{x}_{t_0+2*\Delta t}, \dots, \mathbf{x}_{t_0+T*\Delta t}\}$. The time step Δt is chosen such that the error inherited from the interpolation has a negligible effect on downstream tasks. In this paper, we fix the time step at 10 minutes and omit Δt for the sake of notation simplicity, *i.e.* $\mathbf{x}_{t_0+n*\Delta t} \triangleq \mathbf{x}_{t+n}$.

The L-step-ahead prediction problem comes to predict the trajectory $\mathbf{x}_{T+1:T+L} \triangleq \{\mathbf{x}_{T+1}, \mathbf{x}_{T+2}, \dots, \mathbf{x}_{T+L}\}$ given $T+1$ observations $\mathbf{x}_{0:T} \triangleq \{\mathbf{x}_0, \mathbf{x}_1, \dots, \mathbf{x}_T\}$ up to time T. Given the nature of the prediction problem, it naturally arises as the sampling of the following conditional distribution:

$$p(\mathbf{x}_{T+1:T+L}|\mathbf{x}_{0:T}). \tag{2}$$

We may point out that this probabilistic formulation also covers deterministic prediction models [14], [36], which reduce p to Dirac distributions.

There are two primary categories of approaches to AIS-based vessel trajectory prediction. The first one relies on a state-space formulation, which combines a dynamical prior on vessel movements with a filtering method to infer or sample the posterior (2). Models following this approach have certain limitations. Firstly, the dynamical prior often relies on a simplistic model, such as the Curvilinear Motion Model (CMM) [37], which cannot account for complex patterns including turning points. Secondly, filtering methods, such as the Kalman filter [37], [38] or the particle filter [11], are prone to error propagation issues. Consequently, they seem less suitable for medium-range prediction.

In recent years, the second approach, known as the learning-based approach, has gained substantial popularity [39]. [40], [41], [42] leveraged LSTM (Long Short-Term Memory) and GRU (Gated Recurrent Unit) to learn the temporal patterns in $\mathbf{x}_{0:T}$. However, given the multi-path patterns exhibited in AIS data, such schemes are likely to fail [43]. More sophisticated models take into account the interactions between vessels. [32] used a customized pooling layer—referred to as Collision-Free Social Pooling (CFSP), while [31] employed Graph Convolutional Neural Networks (GCN) to model the interactions between vessels in proximity. These models demonstrate improved prediction performance in dense traffic scenarios with relatively short prediction horizons, typically below one hour. However, in the open sea, where vessel density is significantly lower, and for medium-range horizons (spanning from a few hours to tens of hours), the impacts of the interactions between vessels are less pronounced. Furthermore, since the surrounding environment of a vessel may change as vessels enter or exit the considered zone, it is intractable to explicitly model such interactions for these time horizons (see the Appendix).

To address the heterogeneous and multimodal nature of AIS data, several methods rely on clustering [12], [14], [18], [19], [44]. They assume that maritime traffic in a given area can be represented as a graph, where each node corresponds to a *waypoint* and each edge represents the maritime route between two nodes. The prediction problem then resorts to exploiting a forecasting model over the defined graph. A rich literature exists and exploits among others the constant velocity model and the particle filter [12], Gaussian Processes [13] and neural networks [18], [19]. However, all those schemes face a common limitation: they rely on a route-based representation of maritime traffic, which is viable only when the traffic is highly organized and structured. In real life, a significant fraction of AIS trajectories cannot be assigned to predefined routes [43], [45], limiting the practical application of clustering-based techniques in operational systems.

In this paper, we present a novel model for AIS-based vessel trajectory prediction, referred to as *TrAISformer*. To tackle the complexity and multimodality of AIS vessel trajectories, we propose a new data representation and harness the modeling capabilities of deep learning, specifically transformer architectures [46]. Contrary to clustering-based models which constrain the trajectories to a maritime traffic graph structure [12], [14], [18], [19], [44], *TrAISformer* is applicable to any trajectory within the region of interest, without imposing constraints on an explicit graph of maritime routes. Additionally, we re-frame the prediction as a classification-based learning problem to best forecast the positions of maritime vessels several hours into the future.

III. PROPOSED APPROACH

In this section, we detail the proposed approach. We introduce a new representation of AIS data, derive a new loss function, and provide a brief introduction of the transformer architecture used in *TrAISformer*.

A. DISCRETE AND SPARSE REPRESENTATION OF AIS DATA

One of the primary challenges in trajectory prediction in general, and AIS-based vessel trajectory prediction in particular, is the modeling of the heterogeneous and multimodal nature of motion data given relatively low-dimensional observations. Here, we introduce a novel representation of AIS data, which addresses the heterogeneity aspect. The multimodality part will be addressed in the next subsection with a classification-based training loss.

The most prevalent way to represent an AIS message is to use a 4-dimensional real-valued vector composed of the position and the velocity of the vessel, as in (1).

However, as discussed in [43] and [45], it is challenging to encode complex vessel movement patterns in this feature space. A natural approach is to expand the feature space to a higher dimensional one. Specifically, instead of modeling the conditional probability distribution of the future trajectory given the past $p(\mathbf{x}_{T+1:T+L}|\mathbf{x}_{0:T})$, we consider $p(\mathbf{e}_{T+1:T+L}|\mathbf{e}_{0:T})$, where $\mathbf{e}_t \in \mathbb{R}^{d_e}$ represents a high-dimensional embedding vector of $\mathbf{x}_t$. Recently, variational autoencoders have been very successful in learning such effective mappings that encode $\mathbf{e}_t$ from $\mathbf{x}_t$, and decode $\mathbf{x}_t$ from $\mathbf{e}_t$ [47], [48], [49], [50], [51], [52]. However, when

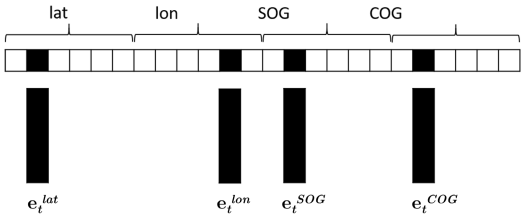

FIGURE 2. Proposed representation of AIS data: To overcome the challenge of representing the heterogeneous and multimodal nature of motion data with relatively low-dimensional observations in AIS-based vessel trajectory prediction, a new representation of AIS data is proposed in this study. For each attribute $att \in \{lat, lon, SOG, COG\}$, the observed value (which is continuous) is discretized into a one-hot vector $\mathbf{h}_t^{att}$. Each $\mathbf{h}_t^{att}$ is then associated with a high dimensional real-valued embedding vector $\mathbf{e}_t^{att}$.

the dimension of $\mathbf{e}_t$ is much higher than that of $\mathbf{x}_t$, the training becomes extremely difficult and prone to overfitting.

To overcome the overfitting problem, we exploit the "four-hot" representation vector $\mathbf{h}_t$, defined in our previous works [43], [45]. Specifically, we discretize the latitude, the longitude, the SOG, and the COG into N_{lat}, N_{lon}, N_{SOG}, and N_{COG} bins, respectively. We then build a one-hot vector for each attribute $\{lat, lon, SOG, COG\}$. The "four-hot" vector $\mathbf{h}_t$ is the concatenation of the four one-hot vectors.

$$\mathbf{h}_t \triangleq [1_t^{lat}, 1_t^{lon}, 1_t^{SOG}, 1_t^{COG}]^T \tag{3}$$

with 1_t^{att} ($att \in \{lat, lon, SOG, COG\}$) being the one-hot vector of att. The details are presented in Algorithm 1.

Each attribute bin of $\mathbf{h}_t$ will be associated with a high dimensional embedding vector, denoted as $\mathbf{e}_t^{att}$. The embedding vector $\mathbf{e}_t$ of an AIS observation $\mathbf{x}_t$ is the concatenation of $\mathbf{e}_t^{att}$. This mapping is illustrated in Fig. 2. The proposed approach ensures that in the embedding space, only $N_{lat} \times N_{lon} \times N_{SOG} \times N_{COG}$ values of $\mathbf{e}_t$ will be used. By imposing this sparsity constraint, we effectively regularize the mapping and avoid overfitting when augmenting the original 4-dimensional AIS observation $\mathbf{x}_t$ to a much higher dimensional space of $\mathbf{e}_t$ [53]. An AIS vessel trajectory is then represented by $\mathbf{e}_{0:T} \triangleq \{\mathbf{e}_0, \mathbf{e}_1, \ldots, \mathbf{e}_T\}$.

Note that the mapping $\mathbf{h}_t \rightarrow \mathbf{e}_t$ is one-to-one, and obtaining $\mathbf{h}_t$ from $\mathbf{x}_t$ is a straightforward process. However, in the reverse direction, it is not possible find the exact $\mathbf{x}_t$ from $\mathbf{h}_t$ because of the discretization. In this paper, we employ a simplifying approximation where we use the mid-points of the bins specified by $\mathbf{h}_t$ to estimate $\mathbf{x}_t$. This approximation introduces an error equivalent to half of the resolution of the $\mathbf{h}_t$ bins in the estimation of $\mathbf{x}_t$, even when the bins estimation is perfect. Nevertheless, we argue that this inherent error is negligible for medium-range vessel trajectory prediction applications such as search and rescue, traffic control, path planning, and port congestion avoidance. To provide an illustration, let's consider a resolution of 0.01° for *lat* and *lon*. The approximation introduces an error of around 0.15 nautical miles (nmi). This level of error does not significantly impact the aforementioned applications, as it remains well within acceptable limits.

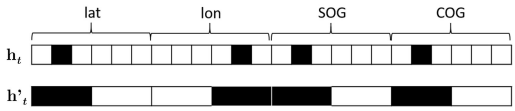

FIGURE 3. Example of multi-resolution "four-hot" vectors for AIS data: The model uses the fine-resolution vector $\mathbf{h}_t$ in the embedding module (see Fig.2), while the loss function uses both $\mathbf{h}_t$ and a coarse-resolution $\mathbf{h}'_t$.

B. TRANSFORMER ARCHITECTURE

As depicted in Fig. 1, in order to forecast the trajectory of a vessel correctly, a prediction model needs to capture possible long-term dependencies in the historical AIS observations. In this regard, transformer neural networks [46] naturally arise as highly suitable candidates. In this work, we adopt a transformer architecture akin to the GPT models [54]. The model's architecture is briefly presented in the following paragraphs. Interested readers are encouraged to refer to [55] for a more complete and formal introduction to transformer.

The transformer network in *TrAISformer* consists of a series of attention layers that are stacked together. Each layer functions as an auto-regressive model that employs the dot-product multiple-head self-attention mechanism:

$$Attention(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_e}}\right)V, \tag{4}$$

where Q, K, V are linear projections of the input sequence (which is $\{\mathbf{e}_t\}$ for the first layer, or the output sequence of the previous layer for other layers), and d_e is the *model size*, *i.e.* the dimension of $\mathbf{e}_t$. At each layer, the input sequence is projected into a new space V, and the output of the attention block is a weighted sum in V, where the weights signify the relative contribution of each time step. These weights are computed as the softmax on the dot product of Q and K, normalized by $\sqrt{d_e}$. The projection operators of Q, K, V are learned during the training phase, and the calculation is performed in parallel. The parallel processing capability allows the model to directly retrieve information from multiple past time steps simultaneously. This is a critical advantage compared to recurrent networks, where the model has to process data sequentially and may not be able to retrieve long-term information.

The output of the transformer's final layer is a vector $\mathbf{l}_t$ with the same dimension as $\mathbf{h}_t$. We will present in the next section how *TrAISformer* uses this output to model $p(\mathbf{h}_{T+l}|\mathbf{e}_{0:T+l-1})$.

C. LEARNING SCHEME

In the learning literature, trajectory prediction is commonly formulated as a regression problem where a model aims to output the best possible continuous-valued $\mathbf{e}_{T+1:T+L}$ (or $\mathbf{x}_{T+1:T+L}$) given the input $\mathbf{e}_{0:T}$ (or $\mathbf{x}_{0:T}$) [14], [18]. Within a deterministic setting, the most common loss function is the mean square error, which measures the squared difference between the predicted and the actual values [14], [19], [56]:

$$\mathcal{L}_{MSE} = \frac{1}{L}\sum_{l=1}^{L} ||\mathbf{x}_{T+l}^{pred} - \mathbf{x}_{T+l}^{true}||_2^2, \tag{5}$$

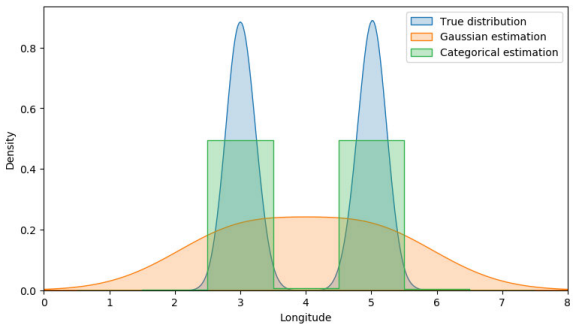

FIGURE 4. Illustration of the loss function $\mathcal{L}_{CE}$ to account for multimodal posterior: Let's consider a scenario where at a specific *waypoint*, half of the vessels in the training set turn left and the other half turn right. The true distribution of the longitude at the next time step forms a bimodal normal distribution, depicted by the blue curve. If we use a real-valued scalar to represent the longitude and use $\mathcal{L}_{MSE}$ as the loss function, the implicit distribution of the model is an unimodal Gaussian distribution. Consequently, the model tends to merge the two modes of the true distribution, as illustrated by the orange curve. By contrast, if we use a one-hot vector to represent the longitude and use $\mathcal{L}_{CE}$ as the loss function, the implicit distribution of the model is a categorical distribution. With this distribution, the model preserves the two modes, as shown by the green curve.

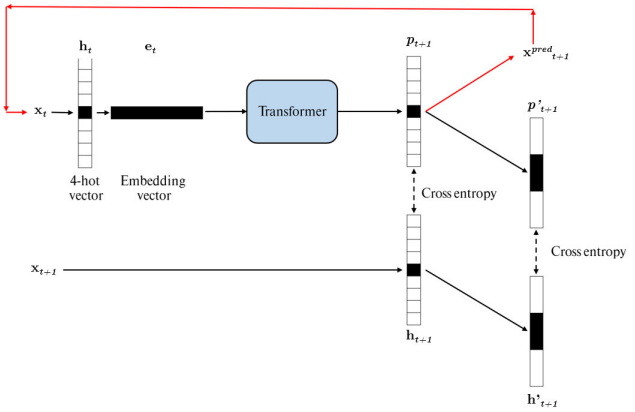

FIGURE 5. Sketch of the *TrAISformer* architecture: each AIS observation $\mathbf{x}_t$ is discretized into a "four-hot" vector $\mathbf{h}_t$ (for visualization purposes, we illustrate a one-hot vector instead of a "four-hot" vector for $\mathbf{h}_t$). Subsequently, each $\mathbf{h}_t$ is paired with a high dimensional real-valued embedding vector $\mathbf{e}_t$. The sequence of embeddings the $\mathbf{e}_{0:t}$ will be fed into a transformer network to predict $\boldsymbol{p}_{t+1} \triangleq p(\mathbf{h}_{t+1}|\mathbf{e}_{0:t})$. During the training phase, the model is optimized to minimize the cross-entropy loss between the true value $\mathbf{h}_{t+1}$ and $\boldsymbol{p}_{t+1}$. To enhance prediction accuracy, we introduce a "multi-resolution" loss. This involves calculating the cross-entropy at different spatial resolutions of $\mathbf{h}_{t+1}$. In the forecasting phase, we generate vessel positions recursively. We sample $\mathbf{h}_{t+1}^{pred}$ from $\boldsymbol{p}_{t+1}$, calculate the "pseudo-inverse" of it to derive $\mathbf{x}_{t+1}^{pred}$. The predicted $\mathbf{x}_{t+1}^{pred}$ is fed back into the network to sample the next position (as shown by the red path in the diagram). This iterative process continues until we reach the desired prediction horizon L.

where $||.||_2$ denotes the Euclidean norm L_2. However, this L_2 loss function (which can be interpreted w.r.t. Gaussian assumption on the conditional likelihood $p(\mathbf{e}_{T+l}|\mathbf{e}_{0:T+l-1})$ or $p(\mathbf{x}_{T+l}|\mathbf{x}_{0:T+l-1})$) may not be appropriate for posterior distributions that exhibit multimodality, as illustrated by vessels' trajectories in Fig.1. To explicitly account for multimodal posteriors, we propose a classification-based formulation that involves modeling $p(\mathbf{e}_{T:T+l}|\mathbf{e}_{0:T+l-1})$ as a concatenation of four categorical distributions, each corresponding to an attribute $\{lat, lon, SOG, COG\}$. Specifically, because the mapping $\mathbf{h}_t \rightarrow \mathbf{e}_t$ is one-to-one, we have:

$$p(\mathbf{h}_{T+l}|\mathbf{e}_{0:T+l-1}) = p(\mathbf{e}_{T+l}|\mathbf{e}_{0:T+l-1}). \tag{6}$$

As $\mathbf{h}_t$ is a "four-hot" vector, this transforms the prediction into a classification problem with four heads, each corresponding to one component of the one-hot vector $\mathbf{h}_t$). Let us denote $p_{T+l} \triangleq p(\mathbf{h}_{T+l}|\mathbf{e}_{0:T+l-1}) = p(\mathbf{e}_{T+l}|\mathbf{e}_{0:T+l-1})$, the loss function is defined as:

$$\mathcal{L}_{CE} \triangleq \sum_{l=1}^{L} CE(\mathbf{h}_{T+l}, p_{T+l}), \tag{7}$$

with CE being the cross-entropy function. The details are presented in Algorithm 2. We demonstrate how the proposed loss function helps maintain the multimodal characteristics of the data in Fig. 4.

Note that $\mathbf{h}_t$ is a discrete representation of the continuous $\mathbf{x}_t$. This discretization can be performed at different resolutions. We empirically observed that the prediction could be marginally improved if we use a multi-resolution version of $\mathcal{L}_{CE}$ as follows (see Fig. 3):

$$\mathcal{L}_{CE} = \sum_{l=1}^{L} CE(p_{T+l}, \mathbf{h}_{T+l}) + \beta CE(p'_{T+l}, \mathbf{h}'_{T+l}). \tag{8}$$

where $p'_{T+l} \triangleq p(\mathbf{h}'_{T+l}|\mathbf{e}_{0:T+l-1})$, $\mathbf{h}'_{T+l}$ is a coarser version of $\mathbf{h}_{T+l}$, β is a scalar balancing the relative importance of the coarse-resolution loss.

The training procedure's specifics are outlined in Algorithm 3. For the sake of notation simplicity, we present the training on a per-series basis. In practice, the model processes the data in batches.

The proposed model is applied recursively. To predict a vessel position at time $T + l$, we sample a "four-hot" vector $\mathbf{h}_{T+l}^{pred}$ from $p(\mathbf{h}_{T+l}|\mathbf{e}_{0:T+l-1})$:

$$\mathbf{h}_{T+l}^{pred} \sim p(\mathbf{h}_{T+l}|\mathbf{e}_{0:T+l-1}) \tag{9}$$

and compute the "pseudo-inverse" of the sampled "four-hot" vector to output the new position $\mathbf{x}_{T+l}^{pred}$. The latter is subsequently fed into the network to sample similarly a position at the next time step. We repeat this sampling procedure until we achieve the desired trajectory length. Multiple runs of this sampling procedure can be performed for a given AIS vessel trajectory to generate different possible predicted paths. This stochastic procedure allows us to address the fact that two vessels currently having similar movement patterns at present may diverge in their trajectories at the next waypoint. The details are presented in Algorithm 4. We demonstrate in Section IV that if we do not sample $\mathbf{h}_{T+l}^{pred}$ from $p(\mathbf{h}_{T+l}|\mathbf{e}_{0:T+l-1})$ according to (9), the performance of the model will degrade.

A sketch of the resulting *TrAISformer* architecture is shown in Fig. 5.

Algorithm 1 fourhot($\mathbf{x}_t$, $\mathbf{R}$, $SOG_{\max}$, $\mathbf{N}$)

Description: Create "four-hot" vector.
Input: AIS observation $\mathbf{x}_t \triangleq [lat, lon, SOG, COG]^T$, the limits of the ROI $\mathbf{R} \triangleq [lat_{\min}, lat_{\max}, lon_{\min}, lon_{\max}]^T$, $SOG_{\max}$, the numbers of bins $\mathbf{N} \triangleq [N^{lat}, N^{lon}, N^{SOG}, N^{COG}]^T$.
Output: "Four-hot" vector $\mathbf{h}_t$.

```
// Create the one-hot vector for
   each attribute.
```

$1_t^{lat} = \texttt{onehot}(\mathbf{x}_t^{lat}, lat_{\min}, lat_{\max}, N^{lat})$
$1_t^{lon} = \texttt{onehot}(\mathbf{x}_t^{lon}, lon_{\min}, lon_{\max}, N^{lon})$
$1_t^{SOG} = \texttt{onehot}(\mathbf{x}_t^{SOG}, 0, SOG_{\max}, N^{SOG})$
$1_t^{COG} = \texttt{onehot}(\mathbf{x}_t^{COG}, 0, 360, N^{COG})$

```
// Concatenate the one-hot vectors
```

$\mathbf{h}_t = [1_t^{lat}, 1_t^{lon}, 1_t^{SOG}, 1_t^{COG}]^T$
Return: $\mathbf{h}_t$

Algorithm 2 ce_loss($\mathbf{h}_t$, $\mathbf{l}_t$, $\mathbf{N}$)

Description: Calculate the cross-entropy loss $\mathcal{L}_{CE}$.
Input: "four-hot" vector $\mathbf{h}_t$, the output of the transformer $\mathbf{l}_t$, the numbers of bins $\mathbf{N}$.
Output: the cross-entropy $CE(\mathbf{h}_t, \mathbf{l}_t)$.

```
// Split h_t back into 4 one-hot
   vectors, each corresponding to an
   attribute of the AIS observation.
```

$1_t^{lat}, 1_t^{lat}, 1_t^{SOG}, 1_t^{COG} = split(\mathbf{h}_t, \mathbf{N})$

```
// Split l_t into 4 heads.
```

$\mathbf{l}_t^{lat}, \mathbf{l}_t^{lon}, \mathbf{l}_t^{SOG}, \mathbf{l}_t^{COG} = split(\mathbf{l}_t, \mathbf{N})$

```
// Calculate the cross-entropy for
   each head.
```

$p_t^{lat} = CE(\texttt{Categorical}(logit = \mathbf{l}_t^{lat}), 1_t^{lat})$
$p_t^{lon} = CE(\texttt{Categorical}(logit = \mathbf{l}_t^{lon}), 1_t^{lon})$
$p_t^{SOG} = CE(\texttt{Categorical}(logit = \mathbf{l}_t^{SOG}), 1_t^{SOG})$
$p_t^{COG} = CE(\texttt{Categorical}(logit = \mathbf{l}_t^{COG}), 1_t^{COG})$

```
// Calculate "total" cross-entropy.
```

$ce\_with\_logit(\mathbf{h}_t, \mathbf{l}_t) = p_t^{lat} * p_t^{lon} * p_t^{SOG} * p_t^{COG}$
Return: $ce\_with\_logit(\mathbf{h}_t, \mathbf{l}_t)$

IV. EXPERIMENTS AND RESULTS

In this section, we report the experimental results of our model on a real AIS dataset introduced in [18]. We include a benchmarking w.r.t state-of-the-art methods for a prediction horizon up to 15 hours. Additionally, we also present an ablation study to assess the relevance of each component of the proposed model. To facilitate the reproducibility of the work in this paper, we chose a publicly available AIS dataset and made the code of the model available at https://github.com/CIA-Oceanix/TrAISformer.

A. EXPERIMENTAL SET-UP

1) DATASET

We tested *TrAISformer* on a public AIS dataset provided by the Danish Maritime Authority (DMA).[3] The dataset comprises AIS observations of cargo and tanker vessels from January 01, 2019 to March 31, 2019. The Region of Interest (ROI) is a rectangle from (55.5°, 10.3°) to (58.0°, 13.0°). Prior to preprocessing, the raw dataset contained approximately 712 million AIS messages. We used AIS data from January 01, 2019 to March 10, 2019 and from March 11, 2019 to March 20, 2019 to train the model and tune the hyper-parameters, respectively. The test set comprises AIS data from March 21, 2019 to March 31, 2019. A subset of this dataset was exploited in [18] to evaluate state-of-the-art models for AIS-based vessel trajectory prediction, including deep learning models.

2) DATA PRE-PROCESSING

AIS data often contain outliers and missing data, which can pose challenges to the prediction. In the training phase, the presence of outliers and missing data introduces additional noise and uncertainty, potentially affecting the convergence of the learning process. During the evaluation phase, missing data prevents us from calculating the prediction errors, while outliers can lead to an inaccurate assessment of prediction accuracy. To mitigate the impact of outliers and missing data, we implemented the following preprocessing steps:

- Remove AIS messages with unrealistic speed values (SOG $\geq$ 30 knots);
- Remove moored or at-anchor vessels;
- Remove AIS observations within 1 nautical mile distance to the coastline;
- Split non-contiguous voyages into contiguous ones. A contiguous voyage [43], [45] is a voyage whose the maximum interval between two consecutive AIS messages is smaller than a predefined value, here 2 hours;
- Remove AIS voyages whose length is smaller than 20 or those that last less than 4h;
- Remove abnormal messages. An AIS message is considered abnormal if the empirical speed (calculated by dividing the distance traveled by the corresponding interval between the two consecutive messages) is unrealistic, here above 40 knots;
- Down-sample AIS data with a sampling rate of 10-minute;
- Split long voyages into shorter ones with a maximum sequence length of 20 hours.

3) HYPER-PARAMETERS

the results reported in this paper were obtained using a transformer architecture with 8 layers. Each layer contains 8 attention heads. The resolution of the "four-hot" vector $\mathbf{h}_t$ was set to 0.01° for *lat* and *lon*, 1 knot for *SOG* and 5° for *COG*. With this resolution, the corresponding sizes of $\mathbf{e}_t^{lat}, \mathbf{e}_t^{lon}, \mathbf{e}_t^{SOG}, \mathbf{e}_t^{COG}$ were 256, 256, 128 and 128 for the ROI reported in this paper. This resulted in a 768-dimensional embedding $\mathbf{e}_t$. The coarse vector $\mathbf{h}_t'$ was

[3] https://dma.dk/safety-at-sea/navigational-information/ais-data

Algorithm 3 trAISformer_train($\{\mathbf{x}_{0:T+L}\}$, Θ, $\mathbf{R}$, SOG_{max}, $\mathbf{N}$, β)

Description: Train *TrAISformer*.
Input: The training set $\{\mathbf{x}_{0:T+L}\}$, the set of *TrAISformer*'s parameters Θ, the limits of the ROI $\mathbf{R}$, SOG_{max}, the numbers of bins $\mathbf{N}$, coefficient β.
Output: The learned set of parameters Θ.

for $\mathbf{x}_{0:T+L}$ *in* $\{\mathbf{x}_{0:T+L}\}$ **do**
  // Create the "four-hot" vectors at different resolutions.
  for t *in* $0:T+L$ **do**
    $\mathbf{h}_t = \texttt{fourhot}(\mathbf{x}_t, \mathbf{R}, SOG_{\text{max}}, \mathbf{N})$
    $\mathbf{h}'_t = \texttt{fourhot}(\mathbf{x}_t, \mathbf{R}, SOG_{\text{max}}, \mathbf{N}/3)$
  end
  // Get the embeddings and apply the transformer.
  $\mathbf{e}_{0:T+L-1} = \texttt{embedding}(\mathbf{h}_{0:T+L-1})$
  $\mathbf{l}_{1:T+L} = \texttt{transformer}(\mathbf{e}_{0:T+L-1})$
  $\mathbf{l}'_{1:T+L} = \texttt{conv1d}(\mathbf{l}_{1:T+L}, kernel = [1/3, 1/3, 1/3]^T, stride = 3)$
  // Calculate the loss.
  $\mathcal{L}_{CE} = 0$
  for l *in* $1:L$ **do**
    $ce(\mathbf{h}_t, \mathbf{l}_t) = \texttt{ce\_loss}(\mathbf{h}_t, \mathbf{l}_t, \mathbf{N})$
    $ce(\mathbf{h}'_t, \mathbf{l}'_t) = \texttt{ce\_loss}(\mathbf{h}'_t, \mathbf{l}'_t, \mathbf{N}/3)$
    $\mathcal{L}_{CE} = \mathcal{L}_{CE} + ce(\mathbf{h}_t, \mathbf{l}_t) + \beta * ce(\mathbf{h}'_t, \mathbf{l}'_t)$
  end
  // Optimize Θ.
  $\Theta = \texttt{AdamW}(\mathcal{L}_{CE}, \Theta, \mathbf{x}_{0:T+L})$
end
Return: Θ

Algorithm 4 trAISformer_predict($\mathbf{x}_{0:T}$, Θ, $\mathbf{R}$, SOG_{max}, $\mathbf{N}$, L)

Description: Use *TrAISformer* to predict vessel trajectory.
Input: The initial segment of the trajectory to predict $\mathbf{x}_{0:T}$, the trained *TrAISformer*'s parameters Θ, the limits of the ROI $\mathbf{R}$, SOG_{max}, the numbers of bins $\mathbf{N}$, the prediction horizon L.
Output: The predicted trajectory $\mathbf{x}_{0:T+L}$.

// Create the "four-hot" vectors of the initial segment.
for t *in* $0:T$ **do**
  $\mathbf{h}_t = \texttt{fourhot}(\mathbf{x}_t, \mathbf{R}, SOG_{\text{max}}, \mathbf{N})$
end
// Get the embeddings of the initial segment.
$\mathbf{e}_{0:T} = \texttt{embedding}(\mathbf{h}_{0:T})$
// Iterate over the prediction horizon.
for l *in* $1:L$ **do**
  $\mathbf{l}_{1:T+l} = \texttt{transformer}(\mathbf{e}_{0:T+l-1})$
  // Split $\mathbf{l}_{T+l}$ into 4 heads.
  $\mathbf{l}^{lat}_{T+l}, \mathbf{l}^{lon}_{T+l}, \mathbf{l}^{SOG}_{T+l}, \mathbf{l}^{COG}_{T+l} = split(\mathbf{l}_{T+l}, \mathbf{N})$
  // Create the categorical distributions and sample $\mathbf{h}^{att}_{T+l}$ from them.
  for att *in* $\{lat, lon, SOG, COG\}$ **do**
    $\mathbf{h}^{att}_{T+l} \sim \texttt{Categorical}(logit = \mathbf{l}^{att}_{T+l})$
  end
  // Get the predicted $\mathbf{h}_{T+l}$, $\mathbf{e}_{T+l}$, and $\mathbf{x}_{T+l}$
  $\mathbf{h}_{T+l} = [\mathbf{h}^{lat}_{T+l}, \mathbf{h}^{lon}_{T+l}, \mathbf{h}^{SOG}_{T+l}, \mathbf{h}^{COG}_{T+l}]^T$
  $\mathbf{e}_{T+l} = \texttt{embedding}(\mathbf{h}_{T+l})$
  $\mathbf{x}_{T+l} = \texttt{pseudo-inverse}(\mathbf{h}_{T+l})$
end
Return: $\mathbf{x}_{0:T+L}$

obtained by merging three consecutive bins of $\mathbf{h}_t$. We noticed that when we reduced or increased the resolution of $\mathbf{h}_t$ by 2, the difference in the results was negligible. The historical sequence length T was set to 3 hours and the prediction horizon L was up to 15 hours. The model was trained using AdamW optimizer [57] with cyclic cosine decay learning rate scheduler [58]. The learning rate was set to $6e^{-4}$. Other implementation details can be found in the GitHub repository that we shared above. We trained the model on a single GTX 1080 Ti GPU over 50 epochs with early stopping. In terms of computational complexity, it took ∼60 minutes to process 10 days of data in the test set, which suggests that the model can run in real-time [59].

4) BENCHMARK MODELS

we compare the performance of *TrAISformer* against different state-of-the-art deep learning models: LSTM seq2seq [14], convolutional seq2seq [60], seq2seq with attention [18], [19], *GeoTrackNet* [45].

It is challenging to conduct a fair quantitative comparison with clustering-based methods [12], [14], [18], [19], [36], [44]. First, those methods did not state clearly how to address clustering noise and small clusters. Second, most of them use a DBSCAN clustering, which is sensitive to hyper-parameters [12], [19], [36]. Different sets of hyper-parameters could lead to very different results. Third, as mentioned in Section II, clustering-based approaches, such as [19], assume vessels' trajectories belong to a predefined graph of maritime routes. This assumption does not hold for the considered dataset. This is the reason why [18] restricted their analysis to a subset of the whole dataset. That subset is composed of tankers' trajectories for a few predefined routes. Though they only involve a subset of trajectories compared with the other benchmarked approaches, we regard the resulting score in [18] as a score under a best-case scenario for clustering-based methods for the considered ROI.

We may point out that, contrary to the whole dataset, that subset has not been made available.

TABLE 1. Mean prediction performance of the benchmarked models (in nautical miles).

Model	1h	2h	3h
LSTM_seq2seq	5.83	8.39	11.64
Conv_seq2seq	4.23	6.77	9.66
LSTM_seq2seq_att	3.35	6.41	9.65
Clustering_LSTM_seq2seq_att[1][18]	*0.78*	*1.93*	*3.66*
GeoTrackNet [45]	0.72	1.59	2.67
TrAISformer	**0.48**	**0.94**	**1.64**
TrAISformer_No-Stoch	1.28	2.88	5.02

[1] The result from [18] was evaluated on a subset of the whole dataset, which comprises only tankers' trajectories for a predefined number of maritime routes. As such, this is regarded as a best-case scenario for clustering-based models.

As the benchmarked models are not public, we conducted an independent implementation and fine-tuned each model to get optimal outcomes.

5) EVALUATION CRITERIA

for each prediction, the prediction error at time step t is calculated as the haversine distance between the true position and the predicted one:

$$d_k = 2R \arcsin\left(\sqrt{\sin^2(\bar{\phi}) + \cos(\phi_1)\cos(\phi_2)\sin^2(\bar{\lambda})}\right), \tag{10}$$

with R the radius of the Earth, $\bar{\phi} \triangleq 0.5(\phi_2 - \phi_1)$, $\bar{\lambda} \triangleq 0.5(\lambda_2 - \lambda_1)$, ϕ_1 and ϕ_2 denote the latitudes, λ_1 and λ_2 denote the longitudes of the predicted position and the true position, respectively.

We used a *best-of-N* criterion, *i.e.* for each model, we sampled N predictions for each target trajectory and reported the best result. In this paper, $N = 16$. This criterion allows us to account for the effect of multimodality.

B. RESULTS

Table 1 shows the average prediction errors evaluated at 1, 2, and 3 hours ahead horizons. The ROI contains several waypoints, rendering prediction for time horizons ranging from 1 to 15 hours highly challenging. In the case of incorrect prediction of turning directions at the waypoints, the prediction errors of a model increase significantly, often above a few nautical miles (nmi), as demonstrated by the benchmarked models in Table 1. *TrAISformer* outperforms all the benchmarked models by a large margin. For instance, for the 2-hour-ahead prediction, it is the only model with an average error below one nautical mile (41% better than the second best model *GeoTracknet*). These results confirm the capability of *TrAISformer* to capture the multimodal nature of vessel trajectories, extract pertinent long-term dependencies, and deliver accurate predictions of vessel paths.

TrAISformer improves by a factor of 2 the performance of the model proposed in [18], which is one of the current

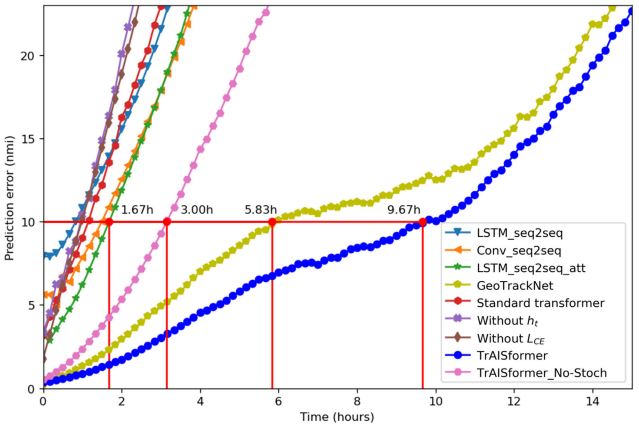

FIGURE 6. Prediction performance w.r.t. prediction time horizon: we plot for each benchmarked model the mean prediction performance for prediction time horizons from 10 minutes to 15 hours. We also highlight the time horizon up to which the performance of a given model remains below the maximum visibility under good weather conditions (i.e., 10 nmi).

state-of-the-art schemes, with respective scores of 0.94 nmi and 1.93 nmi. We may recall that the performance of this clustering-based scheme refers to a best-case scenario, as it only involves tankers' trajectories for a few maritime routes in the case-study region. We also note that the direct application of state-of-the-art deep learning schemes on the 4-dimensional AIS feature vector, namely LSTM seq2seq [14], convolutional seq2seq [60], seq2seq with attention [18], [19], transformer [46], [54] (see Table. 2) leads to poor prediction performances (mean error greater than 6 nmi for a 2-hour-ahead prediction). The second best approach is our previous work *GeoTrackNet* [43], [45]. It shares two key features with *TrAISformer*: i) a similar sparse high-dimensional representation of AIS data and ii) a probabilistic neural-network-based learning scheme. However, *GeoTrackNet* uses a Variational Recurrent Neural Network (VRNN) [61] instead of a transformer architecture to capture the temporal patterns in the AIS data. The improved performance of *TrAISformer* over *GeoTrackNet* suggests that transformers may be a better neural architecture for AIS data than VRNN.

To further highlight the importance of the probabilistic feature of *TrAISformer*, we report the performance of a deterministic version—denoted as *TrAISformer_No-Stoch*—of *TrAISformer*. This model outputs the "four-hot" vector with the highest probability, *i.e.* $\mathbf{h}^{pred}_{T+l} = \text{argmax}_{\mathbf{h}}\, p(\mathbf{h}_{T+l}|\mathbf{e}_{0:T+l-1})$, instead of sampling $\mathbf{h}^{pred}_{T+l}$ like in (9). The decrease in the prediction performance (from 0.94 to 2.88 for the 2-hour-ahead prediction) demonstrates the importance of a multimodal representation of vessels' trajectories for the considered case-study. As pointed out previously, two vessels departing from the same port, having the same current position and velocity, may follow different paths at the next waypoint, making it impossible for the prediction model to produce correct deterministic forecasts all the time. Models that are capable of predicting multiple possibilities are more relevant. Yet, the deterministic version

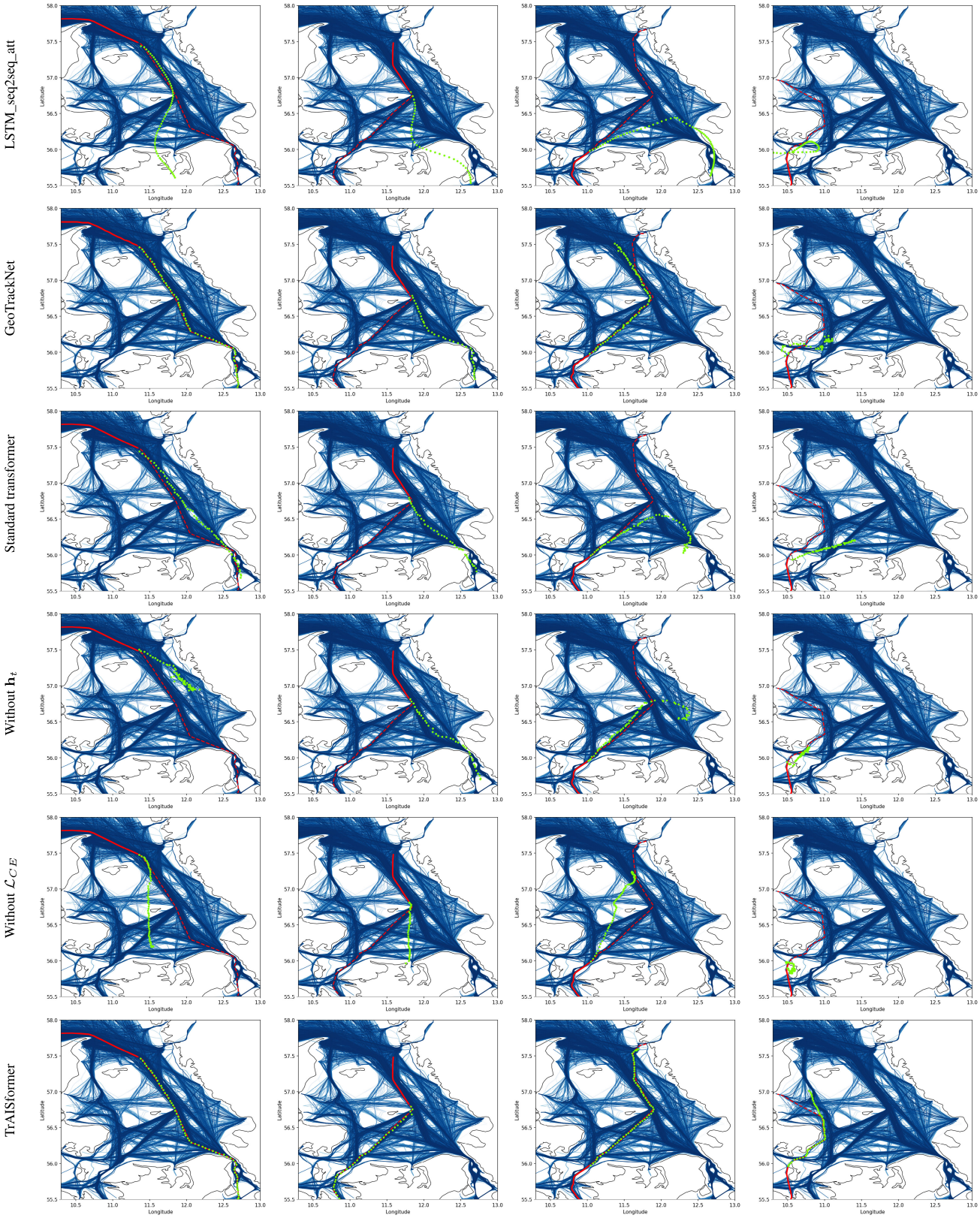

FIGURE 7. Examples of AIS-based vessel trajectory predictions: each column depicts the predictions of a given real vessel trajectory. The rows correspond to the various models used for benchmarking. For each example, we display the AIS observations $\mathbf{x}_{0:T}$ used as the input by all models –, the real vessel trajectory –, and the predicted trajectory ●.

TABLE 2. Mean prediction performance (in nautical miles) of the models in the ablation study.

Model	AIS data representation	Embedding $\mathbf{x}_t \rightarrow \mathbf{e}_k$	Loss function	1h	2h	3h
Without $\mathbf{e}_t$ and $\mathbf{h}_t$ (standard transformer)	$[lat, lon, SOG, COG]^T$	None	$\mathcal{L}_{MSE}$	4.75	8.36	11.40
Without $\mathbf{h}_t$	$[lat, lon, SOG, COG]^T \rightarrow \mathbf{e}_t$	MLP	$\mathcal{L}_{MSE}$	5.02	9.69	15.04
Without the classification loss $\mathcal{L}_{CE}$	"four-hot" vector $\rightarrow \mathbf{e}_t$	Via $\mathbf{h}_t$	$\mathcal{L}_{MSE}$	5.53	10.64	16.06
TrAISformer	"four-hot" vector $\rightarrow \mathbf{e}_t$	Via $\mathbf{h}_t$	$\mathcal{L}_{CE}$	**0.48**	**0.94**	**1.64**

of *TrAISformer* is still much better than standard seq2seq models, which again stresses the relevance of the transformer architecture as well as of the considered representation of AIS feature vector.

In maritime downstream tasks, a crucial factor to consider when using a prediction model is the maximum meaningful prediction horizon, which is the longest time horizon where a prediction is still useful. In some scenarios, such as search and rescue operations, a prediction is deemed helpful if the prediction error is smaller than the visibility, which is generally assumed to be 10 nmi under clear weather conditions [62]. Fig. 6 depicts such prediction horizons of the benchmarked models. In the best case scenario, *TrAISformer* can extend the prediction horizon by a factor of ∼5.8 compared with current state-of-the-art methods (9.67h for *TrAISformer vs*. 1.67h for LSTM_seq2seq_att).

Fig. IV-B displays some examples of the predictions made by *TrAISformer* and the other benchmarked methods. *TrAISformer* successfully samples realistic turning directions to forecast the potential paths taken by vessels. We recall here that for probabilistic models such as *GeoTrackNet*, *TrAISformer*, we report among 16 sampled trajectories the one closest to the real trajectory. We may highlight that the model applies not only to the main maritime routes (the first three columns) but also to less frequent ones (the last column). By contrast, clustering-based methods struggle in such cases. The four examples show the relatively poor performance of the direct application of sequence-to-sequence deep learning models. *GeoTrackNet* samples realistic trajectories for the first three examples, though not as close to the real ones as the ones predicted by *TrAISformer*. However, for the last example, it performs poorly, while *TrAISformer* still succeeds in sampling a realistic path.

We further analyze in Fig. IV-B the behavior of *TrAISformer* through the activation of an attention block in the first layer of *TrAISformer* for the trajectory shown in Fig. 1. Each row shows the relative importance of each time step in the predicted output. Some remarks raised from this analysis:

- On straight lines, only the information from recent time steps is used to predict the next time step, which is similar to constant velocity models [63];
- At the waypoints, the model needs to retrieve information from much earlier time steps, especially at the previous waypoints to predict the next time step. For example, row 40 (the red rectangle) depicts the attention weights to compute the prediction at **E**. The model pays more attention to the inputs at **A**, **B**, **C**, **D**, and **E**. One intuitive explanation is that the model needs to know where the vessel comes from (points **A**, **B**, **C**), what the movement pattern of the vessel is in the current segment (point **D**), as well as the current position and velocity (point **E**) to guess the movement pattern to come.

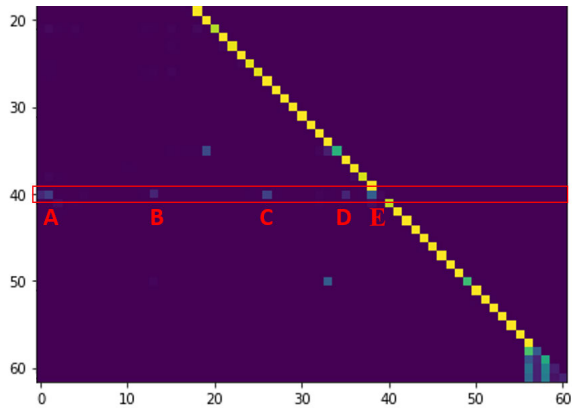

FIGURE 8. Relative importance of each time step in the prediction: Visualization of the activation of one attention block of *TrAISformer* for the trajectory shown in Fig. 1. Horizontal axis: input time step; vertical axis: output time step. Each row shows which parts of the input that the model pays attention to in order to compute the output at the corresponding time step.

As such, this example demonstrates the ability of *TrAISformer* to extract relevant long-term dependencies to predict vessel trajectories.

C. ABLATION STUDY

To evaluate the significance of the different components of *TrAISformer* architecture, we conducted an ablation study:

- Firstly, we removed $\mathbf{e}_t$ and $\mathbf{h}_t$ to demonstrate the significance of the high-dimensional encoding. This is equivalent to applying directly a GPT model [54] to 4-dimensional AIS data streams.
- Secondly, we kept $\mathbf{e}_t$ but removed $\mathbf{h}_t$ to assess the relevance of the sparsity constraint. The embedding $\mathbf{x}_t \rightarrow \mathbf{e}_t$ in this model is a MultiLayer Perceptron (MLP).
- Finally, we tested a model with the same architecture as *TrAISformer* but used a regression loss as the training loss to demonstrate the criticality of the classification loss.

The results in Tab. 2 show that all the ablated models lead to significantly worse performance compared with

TrAISformer. Interestingly, the performance degradation is in the same order of magnitude for the three ablated models, though the impact of the classification-based loss is slightly greater. Overall, these results highlight the importance of integrating all the components of our architecture for achieving the best prediction performance.

V. CONCLUSION AND FUTURE WORK

In this paper, we presented a novel model—referred to as *TrAISformer*, for vessel trajectory prediction using AIS data. The model uses an augmented, sparse, and high-dimensional representation of AIS data as well as a state-of-the-art transformer network architecture to learn complex patterns in vessel trajectories. Using a classification-based training loss, *TrAISformer* can capture the multimodal nature of trajectory data. Experiments on real, public AIS data show *TrAISformer* outperforms existing methods by a significant margin. With a 9-hour-ahead prediction error below 10 nmi on a real AIS dataset in a case-study region involving dense and complex maritime traffic patterns, these results open new research avenues for various applications such as search and rescue, port congestion avoidance, and maritime surveillance.

Through an ablation study, we have shown that all the above-mentioned components of *TrAISformer* have critical roles in the reported performance. Though transformer architectures are likely not fully explainable [64], [65], we have also shown that the intermediate attention weights of the transformer architecture provide a natural way to explore how *TrAISformer* exploits the past AIS data to compute its predictions of the future trajectory. This supports that the learned transformer representation could be of interest beyond the considered application to prediction tasks.

Future work could further improve the architecture and develop the applications of *TrAISformer* framework. Among others, we may cite the learning of conditional *TrAISformer* w.r.t. weather conditions as the latter clearly impact vessels' movement. While we currently omit the influence of vessel interactions, future work could study the possibility of integrating those interactions into the model. We may also stress that the proposed *TrAISformer* architecture is significantly more complex (see Tab. 3) with ∼300 times more parameters than the second most complex architecture among the benchmarked ones. While the greater complexity likely contributes to the significant gain, recent advances in model compression techniques, such as Neural Network Pruning [66] and Knowledge Distillation [67], suggest that we could reduce the model's size typically by a factor of tens to hundreds without compromising its performance. This would promote the assessment and adoption of *TrAISformer* in operational systems. The combination of *TrAISformer* with other learning-based modules for classification and anomaly detections [45] is also of interest. Recent advances in the exploitation of AIS data for the inversion of sea surface parameters [68], [69] may also be an appealing line of research for our future work.

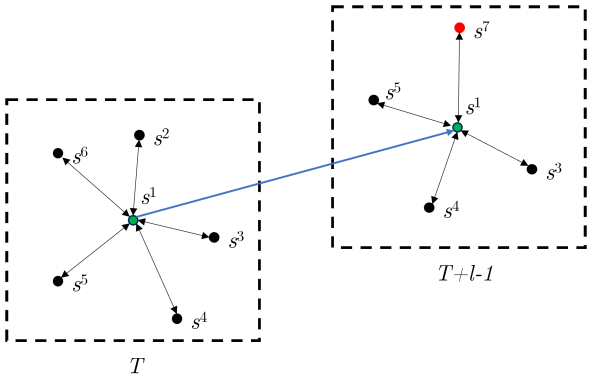

FIGURE 9. Illustration of the intractability of the modeling of the interactions between agents in medium-range trajectory prediction. Consider the scenario where we aim to predict the medium-range trajectory of agent s^1 (denoted by the green dot). At time T, there are six other agents ($s^2, \ldots, s^6$) within its vicinity (enclosed by the dashed rectangle). The interactions between s^1 and these agents are depicted by the double-headed arrows. As we project into the medium-range future at time $T + l - 1$, s^2 and s^6 will have moved away from s^1; and a new, unknown agent s^7 (denoted by the red dot) may appear near s^1. The prediction model would have no knowledge of this new vicinity, making the modeling of interactions between agents intractable.

APPENDIX A
WHY PEDESTRIAN AND VEHICLE TRAJECTORY PREDICTION MODELS DO NOT APPLY TO MEDIUM-RANGE AIS-BASED VESSEL TRAJECTORY PREDICTION

In this appendix, we present a mathematical demonstration explaining why pedestrian and vehicle trajectory prediction models are not directly applicable to medium-range AIS-based vessel trajectory prediction.

Let us denote by $\mathbf{x}_t^{s^i}$ an observation of an agent s^i at time t. For instance, in pedestrian and vehicle trajectory prediction, s^i can represent either a pedestrian or a vehicle, and $\mathbf{x}_t^{s^i}$ corresponds to their respective positions on the map. In AIS trajectory prediction, s^i represents a vessel, and $\mathbf{x}_t^{s^i}$ represents its AIS message. The trajectory of agent s^i from t_1 to t_2 ($t_2 > t_1$) is then represented by a sequence of observations $\mathbf{x}_{t_1:t_2}^{s^i} = \{\mathbf{x}_{t_1}^{s^i}, \mathbf{x}_{t_1+1}^{s^i}, \ldots, \mathbf{x}_{t_2}^{s^i}\}$. At time t, we denote the group of other agents in the vicinity of s^i as $\mathbf{V}_t^i$, and their historical trajectories are denoted as $\mathbf{x}_{0:t}^{\mathbf{V}_t^i}$.

In the context of this paper, using these notations, trajectory prediction refers to forecasting the trajectory of an agent s^i for L timesteps ahead, based on the historical observations of this agent and the others in the vicinity up to time T, by maximizing the likelihood:

$$p(\mathbf{x}_{T+1:T+L}^{s^i} | \mathbf{x}_{0:T}^{s^i}, \mathbf{x}_{0:T}^{\mathbf{V}_t^i}). \tag{11}$$

Here we use p in the broad sense, which includes deterministic models.

The conditioning side of (11) encompasses two crucial components: the $\mathbf{x}_{0:T}^{s^i}$ component embeds the intention of the agent, while the $\mathbf{x}_{0:T}^{\mathbf{V}_t^i}$ part embeds the interactions with the environment. State-of-the-art methods for pedestrian

and vehicle trajectory prediction are centered on effectively modeling and integrating these two terms. Prominent examples include S-LSTM [26], S-GAN [26], S-ATTN [27], SoPhie [28], MATF [29], Trajection [70], Trajectron++ [30], etc. Those models leverage deep neural networks such as LSTM or GAN (Generative Adversarial Network) to capture correlations in historical data, and some pooling techniques to embed the interactions between agents. This idea has been adopted for short-term AIS-based vessel trajectory prediction [31], [32]. Although those methods have shown promising results on the corresponding datasets, they are not suitable for medium-range vessel trajectory prediction. First, the prediction horizons considered in those works are from a few seconds to a few minutes, which are too short for maritime applications. Second, those works address different types of maneuvers, of which the movement of an agent depends highly on the interactions with other agents and the surrounding environment. For maritime traffic contexts, and at medium-range time horizons, the path that a vessel will make depends mainly on where it wants to go. It is barely affected by the interactions with other vessels in the vicinity at the current moment. Mathematically, this means at time T we have the approximation:

$$p(\mathbf{x}^{s^i}_{T+l}|\mathbf{x}^{s^i}_{0:T}, \mathbf{x}^{\mathbf{v}^i_t}_{0:T}) \approx p(\mathbf{x}^{s^i}_{T+l}|\mathbf{x}^{s^i}_{0:T})\Big|_{l\gg 1}. \tag{12}$$

One may argue that we could use the predicted value of $\mathbf{x}^{\mathbf{v}^i_{T+l-1}}_{0:T+l-1}$ and $\mathbf{x}^{s^i}_{0:T+l-1}$ to estimate $\mathbf{x}^{s^i}_{T+l}$. However, in order to predict $\mathbf{x}^{\mathbf{v}^i_{T+l-1}}_{0:T+l-1}$, we need to predict the trajectory of all the agents in the vicinity of s^i at $T+l-1$. This is an expensive or even intractable approach. For example, an unknown agent may join the ROI, as illustrated in Fig. 9).

It's important to note that if we remove $\mathbf{x}^{\mathbf{v}^i_t}_{0:T}$ from (11), this objective function simplifies (2), which is the objective function used in medium-range AIS-based vessel trajectory prediction. Likewise, if we remove the module that encodes the interactions between agents in some of the pedestrian and vehicle trajectory prediction models mentioned above, we get models that have similar architectures to those designed for AIS trajectory prediction. For example, if we remove the interactions between agents part in Trajectron [70], it becomes an LSTM_seq2seq model.

APPENDIX B
MODEL SELECTION STRATEGY

In this work, we used cross-validation as the model selection strategy.

As suggested by Reviewers, another approach is to use information criteria such as the Akaike Information Criterion (AIC) [71], the Bayesian Information Criterion (BIC) [72], the Hannan-Quinn Information Criterion (HQIC) [73].

We have calculated these criteria, the result is shown in Tab.3.

The details of the calculation are as follows:

TABLE 3. AIC SBIC, and HQIC of the studied models.

Model	# parameters	AIC	BIC	HQIC
LSTM_seq2seq	1.2k	-6.4M	-6.4M	-6.4M
Conv_seq2seq	56.3k	-7.7M	-7.1M	-7.5M
LSTM_seq2seq_att	33.7k	-7.7M	-7.4M	-7.6M
GeoTrackNet [45]	179.8k	-3.5M	-1.5M	-3.0M
TrAISformer	57.4M	111.0M	753.9M	292.4M

- For deterministic models (LSTM_seq2seq, Conv_seq2seq, LSTM_seq2seq_att), we calculated the approximated likelihood as $\hat{\mathcal{L}} = RSS/n$, with RSS being the Residual Sum Squares of the fitting and n being the number of observations [74].
- Although GeoTrackNet and TrAISformer are probabilistic models, they use transformed data. Hence, we also used RSS/n to approximate the likelihood in the original data space.
- Since the model in [18] (Clustering_LSTM_seq2seq_att) works in a reduced scope, we did not include this model in this comparison (because we can not compare the information criteria of models trained on different data).
- TrAISformer_No-stoch is TrAISformer with a different inference strategy. The information criteria of TrAISformer_No-stoch are the same as TrAISformer's.

We can see that the values are dominated by the penalty terms and more specifically the number of parameters. The AIC, SBIC, and HQIC in Tab 3 favor models with fewer parameters (LSTM_seq2seq, Conv_seq2seq, LSTM_seq2seq_att). This is in line with the objective of information criteria: to identify overparameterized models [75]. By contrast, neural networks can often be regarded as over-parameterized. This motivates the exploitation of regularization schemes during the training phase to avoid overfitting patterns. To some extent, overparametrization can yield positive effects [76], [77]).

REFERENCES

[1] Z. Ou and J. Zhu, "AIS database powered by GIS technology for maritime safety and security," *J. Navigat.*, vol. 61, no. 4, pp. 655–665, Oct. 2008.
[2] I. Varlamis, K. Tserpes, and C. Sardianos, "Detecting search and rescue missions from AIS data," in *Proc. IEEE 34th Int. Conf. Data Eng. Workshops (ICDEW)*, Apr. 2018, pp. 60–65.
[3] T. Fabbri, R. Vicen-Bueno, R. Grasso, G. Pallotta, L. M. Millefiori, and L. Cazzanti, "Optimization of surveillance vessel network planning in maritime command and control systems by fusing METOC & AIS vessel traffic information," in *Proc. OCEANS*, May 2015, pp. 1–7.
[4] E. Tu, G. Zhang, L. Rachmawati, E. Rajabally, and G.-B. Huang, "Exploiting AIS data for intelligent maritime navigation: A comprehensive survey," *IEEE Trans. Intell. Transp. Syst.*, vol. 19, no. 5, pp. 1559–1582, Sep. 2017.
[5] X. K. Dang, H. N. Truong, and V. D. Do, "A path planning control for a vessel dynamic positioning system based on robust adaptive fuzzy strategy," *Automatika*, vol. 63, no. 3, pp. 580–592, Jul. 2022.
[6] X.-P. Nguyen, X.-K. Dang, V.-D. Do, J. M. Corchado, and H.-N. Truong, "Robust adaptive fuzzy-free fault-tolerant path planning control for a semi-submersible platform dynamic positioning system with actuator constraints," *IEEE Trans. Intell. Transp. Syst.*, vol. 24, no. 11, pp. 12701–12715, Nov. 2023.

[7] J. M. Mou, C. V. D. Tak, and H. Ligteringen, "Study on collision avoidance in busy waterways by using AIS data," *Ocean Eng.*, vol. 37, nos. 5–6, pp. 483–490, Apr. 2010.
[8] C. Liu, J. Liu, X. Zhou, Z. Zhao, C. Wan, and Z. Liu, "AIS data-driven approach to estimate navigable capacity of busy waterways focusing on ships entering and leaving port," *Ocean Eng.*, vol. 218, Dec. 2020, Art. no. 108215.
[9] M. J. Kang, S. Zohoori, M. Hamidi, and X. Wu, "Study of narrow waterways congestion based on automatic identification system (AIS) data: A case study of Houston ship channel," *J. Ocean Eng. Sci.*, vol. 7, no. 6, pp. 578–595, Dec. 2022.
[10] G. Soldi, D. Gaglione, N. Forti, L. M. Millefiori, P. Braca, S. Carniel, A. D. Simone, A. Iodice, D. Riccio, F. C. Daffinà, D. Quattrociocchi, G. Bottini, P. Willett, and A. Farina, "Space-based global maritime surveillance. Part II: Artificial intelligence and data fusion techniques," *IEEE Aerosp. Electron. Syst. Mag.*, vol. 36, no. 9, pp. 30–42, Sep. 2021.
[11] B. Ristic, B. L. Scala, M. Morelande, and N. Gordon, "Statistical analysis of motion patterns in AIS Data: Anomaly detection and motion prediction," in *Proc. 11th Int. Conf. Inf. Fusion*, Jun. 2008, pp. 1–7.
[12] F. Mazzarella, V. F. Arguedas, and M. Vespe, "Knowledge-based vessel position prediction using historical AIS data," in *Proc. Sensor Data Fusion, Trends, Solutions, Appl. (SDF)*, Oct. 2015, pp. 1–6.
[13] H. Rong, A. P. Teixeira, and C. G. Soares, "Ship trajectory uncertainty prediction based on a Gaussian process model," *Ocean Eng.*, vol. 182, pp. 499–511, Jun. 2019.
[14] N. Forti, L. M. Millefiori, P. Braca, and P. Willett, "Prediction of vessel trajectories from AIS data via sequence-to-sequence recurrent neural networks," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, May 2020, pp. 8936–8940.
[15] T. A. Volkova, Y. E. Balykina, and A. Bespalov, "Predicting ship trajectory based on neural networks using AIS data," *J. Mar. Sci. Eng.*, vol. 9, no. 3, p. 254, Feb. 2021.
[16] B. Murray and L. P. Perera, "Ship behavior prediction via trajectory extraction-based clustering for maritime situation awareness," *J. Ocean Eng. Sci.*, vol. 7, no. 1, pp. 1–13, Feb. 2022.
[17] J. Park, J. Jeong, and Y. Park, "Ship trajectory prediction based on bi-LSTM using spectral-clustered AIS data," *J. Mar. Sci. Eng.*, vol. 9, no. 9, p. 1037, Sep. 2021.
[18] S. Capobianco, L. M. Millefiori, N. Forti, P. Braca, and P. Willett, "Deep learning methods for vessel trajectory prediction based on recurrent neural networks," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 57, no. 6, pp. 4329–4346, Dec. 2021.
[19] B. Murray and L. P. Perera, "An AIS-based deep learning framework for regional ship behavior prediction," *Rel. Eng. Syst. Saf.*, vol. 215, Nov. 2021, Art. no. 107819.
[20] L. M. Millefiori, G. Pallotta, P. Braca, S. Horn, and K. Bryan, "Validation of the Ornstein–Uhlenbeck route propagation model in the Mediterranean Sea," in *Proc. OCEANS*, May 2015, pp. 1–6.
[21] L. M. Millefiori, P. Braca, K. Bryan, and P. Willett, "Modeling vessel kinematics using a stochastic mean-reverting process for long-term prediction," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 52, no. 5, pp. 2313–2330, Oct. 2016.
[22] A. Rudenko, L. Palmieri, M. Herman, K. M. Kitani, D. M. Gavrila, and K. O. Arras, "Human motion trajectory prediction: A survey," *Int. J. Robot. Res.*, vol. 39, no. 8, pp. 895–935, Jul. 2020.
[23] H. Zhao, J. Gao, T. Lan, C. Sun, B. Sapp, B. Varadarajan, Y. Shen, Y. Shen, Y. Chai, C. Schmid, C. Li, and D. Anguelov, "TNT: Target-driven trajectory prediction," in *Proc. Conf. Robot Learn.*, Oct. 2021, pp. 895–904.
[24] F. Leon and M. Gavrilescu, "A review of tracking and trajectory prediction methods for autonomous driving," *Mathematics*, vol. 9, no. 6, p. 660, Mar. 2021.
[25] J. Gu, C. Sun, and H. Zhao, "DenseTNT: End-to-end trajectory prediction from dense goal sets," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2021, pp. 15283–15292.
[26] A. Gupta, J. Johnson, L. Fei-Fei, S. Savarese, and A. Alahi, "Social GAN: Socially acceptable trajectories with generative adversarial networks," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit.*, Jun. 2018, pp. 2255–2264.
[27] A. Vemula, K. Muelling, and J. Oh, "Social attention: Modeling attention in human crowds," in *Proc. IEEE Int. Conf. Robot. Autom. (ICRA)*, May 2018, pp. 4601–4607.
[28] A. Sadeghian, V. Kosaraju, A. Sadeghian, N. Hirose, H. Rezatofighi, and S. Savarese, "SoPhie: An attentive GAN for predicting paths compliant to social and physical constraints," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 1349–1358.
[29] T. Zhao, Y. Xu, M. Monfort, W. Choi, C. Baker, Y. Zhao, Y. Wang, and Y. N. Wu, "Multi-agent tensor fusion for contextual trajectory prediction," in *Proc. IEEE/CVF Conf. Comput. Vis. Pattern Recognit. (CVPR)*, Jun. 2019, pp. 12118–12126.
[30] T. Salzmann, B. Ivanovic, P. Chakravarty, and M. Pavone, "Trajectron++: Dynamically-feasible trajectory forecasting with heterogeneous data," 2020, *arXiv:2001.03093*.
[31] R. W. Liu, M. Liang, J. Nie, Y. Yuan, Z. Xiong, H. Yu, and N. Guizani, "STMGCN: Mobile edge computing-empowered vessel trajectory prediction using spatio-temporal multigraph convolutional network," *IEEE Trans. Ind. Informat.*, vol. 18, no. 11, pp. 7977–7987, Nov. 2022.
[32] R. W. Liu, M. Liang, J. Nie, W. Y. B. Lim, Y. Zhang, and M. Guizani, "Deep learning-powered vessel trajectory prediction for improving smart traffic services in maritime Internet of Things," *IEEE Trans. Netw. Sci. Eng.*, vol. 9, no. 5, pp. 3080–3094, Sep. 2022.
[33] P. Coscia, P. Braca, L. M. Millefiori, F. A. N. Palmieri, and P. Willett, "Multiple Ornstein–Uhlenbeck processes for maritime traffic graph representation," *IEEE Trans. Aerosp. Electron. Syst.*, vol. 54, no. 5, pp. 2158–2170, Oct. 2018.
[34] I. Varlamis, K. Tserpes, M. Etemad, A. S. JÃúnior, and S. Matwin, "A network abstraction of multi-vessel trajectory data for detecting anomalies," in *Proc. EDBT/ICDT Workshops*, 2019. [Online]. Available: https://ceur-ws.org/Vol-2322/
[35] R. Bošnjak, L. Šimunović, and Z. Kavran, "Automatic identification system in maritime traffic and error analysis," *Trans. Maritime Sci.*, vol. 1, no. 2, pp. 77–84, Oct. 2012.
[36] Y. Suo, W. Chen, C. Claramunt, and S. Yang, "A ship trajectory prediction framework based on a recurrent neural network," *Sensors*, vol. 20, no. 18, p. 5133, Sep. 2020.
[37] L. P. Perera, P. Oliveira, and C. G. Soares, "Maritime traffic monitoring based on vessel detection, tracking, state estimation, and trajectory prediction," *IEEE Trans. Intell. Transp. Syst.*, vol. 13, no. 3, pp. 1188–1200, Sep. 2012.
[38] S. Fossen and T. I. Fossen, "Extended Kalman filter design and motion prediction of ships using live automatic identification system (AIS) data," in *Proc. 2nd Eur. Conf. Electr. Eng. Comput. Sci. (EECS)*, Dec. 2018, pp. 464–470.
[39] X. Zhang, X. Fu, Z. Xiao, H. Xu, and Z. Qin, "Vessel trajectory prediction in maritime transportation: Current approaches and beyond," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 11, pp. 19980–19998, Nov. 2022.
[40] C. Wang, H. Ren, and H. Li, "Vessel trajectory prediction based on AIS data and bidirectional GRU," in *Proc. Int. Conf. Comput. Vis., Image Deep Learn. (CVIDL)*, Jul. 2020, pp. 260–264.
[41] W. Li, C. Zhang, J. Ma, and C. Jia, "Long-term vessel motion predication by modeling trajectory patterns with AIS data," in *Proc. 5th Int. Conf. Transp. Inf. Saf. (ICTIS)*, Jul. 2019, pp. 1389–1394.
[42] H. Tang, Y. Yin, and H. Shen, "A model for vessel trajectory prediction based on long short-term memory neural network," *J. Mar. Eng. Technol.*, vol. 21, no. 3, pp. 136–145, Sep. 2019.
[43] D. Nguyen, R. Vadaine, G. Hajduch, R. Garello, and R. Fablet, "A multi-task deep learning architecture for maritime surveillance using AIS data streams," in *Proc. IEEE 5th Int. Conf. Data Sci. Adv. Anal. (DSAA)*, Oct. 2018, pp. 331–340.
[44] G. Pallotta, M. Vespe, and K. Bryan, "Vessel pattern knowledge discovery from AIS data: A framework for anomaly detection and route prediction," *Entropy*, vol. 15, no. 12, pp. 2218–2245, Jun. 2013.
[45] D. Nguyen, R. Vadaine, G. Hajduch, R. Garello, and R. Fablet, "GeoTrackNet—A maritime anomaly detector using probabilistic neural network representation of AIS tracks and a contrario detection," *IEEE Trans. Intell. Transp. Syst.*, vol. 23, no. 6, pp. 5655–5667, Jun. 2022.
[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Å. Kaiser, and I. Polosukhin, "Attention is all you need," in *Advances Neural Information Processing Systems*, vol. 30. Red Hook, NY, USA: Curran Associates, 2017.
[47] Y. LeCun, Y. Bengio, and G. Hinton, "Deep learning," *Nature*, vol. 521, no. 7553, pp. 436–444, May 2015.
[48] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. Cambridge, MA, USA: MIT Press, 2016.
[49] D. P. Kingma and M. Welling, "Auto-encoding variational Bayes," 2013, *arXiv:1312.6114*.

[50] D. Jimenez Rezende and S. Mohamed, "Variational inference with normalizing flows," 2015, *arXiv:1505.05770*.
[51] Y. Pu, Z. Gan, R. Henao, X. Yuan, C. Li, A. Stevens, and L. Carin, "Variational autoencoder for deep learning of images, labels and captions," in *Advances Neural Information Processing Systems*, vol. 29, D. D. Lee, M. Sugiyama, U. V. Luxburg, I. Guyon, and R. Garnett, Eds. Red Hook, NY, USA: Curran Associates, 2016, pp. 2352–2360.
[52] A. Vahdat and J. Kautz, "NVAE: A deep hierarchical variational autoencoder," in *Advances Neural Information Processing Systems*, vol. 33. Red Hook, NY, USA: Curran Associates, 2020, pp. 19667–19679.
[53] A. Ng, "Sparse autoencoder," *CS294A Lect. Notes*, vol. 72, pp. 1–19, Jan. 2011.
[54] A. Radford, K. Narasimhan, T. Salimans, and I. Sutskever, "Improving language understanding by generative pre-training," OpenAI, Tech. Rep., Jan. 2018. [Online]. Available: https://openai-assets.s3.amazonaws.com/research-covers/language-unsupervised/language_understanding_paper.pdf
[55] M. Phuong and M. Hutter, "Formal algorithms for transformers," 2022, *arXiv:2207.09238*.
[56] P. Dijt and P. Mettes, "Trajectory prediction network for future anticipation of ships," in *Proc. Int. Conf. Multimedia Retr. (ICMR)*. New York, NY, USA: Association for Computing Machinery, Jun. 2020, pp. 73–81.
[57] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," 2017, *arXiv:1711.05101*.
[58] I. Loshchilov and F. Hutter, "SGDR: Stochastic gradient descent with warm restarts," 2016, *arXiv:1608.03983*.
[59] D. Nguyen, M. Simonin, G. Hajduch, R. Vadaine, C. Tedeschi, and R. Fablet, "Detection of abnormal vessel behaviours from AIS data using GeoTrackNet: From the laboratory to the ocean," in *Proc. 21st IEEE Int. Conf. Mobile Data Manage. (MDM)*, Jun. 2020, pp. 264–268.
[60] J. Gehring, M. Auli, D. Grangier, D. Yarats, and Y. N. Dauphin, "Convolutional sequence to sequence learning," in *Proc. Int. Conf. Mach. Learn.*, Jul. 2017, pp. 1243–1252.
[61] J. Chung, K. Kastner, L. Dinh, K. Goel, A. Courville, and Y. Bengio, "A recurrent latent variable model for sequential data," in *Proc. Adv. Neural Inf. Process. Syst.*, Jun. 2015, pp. 2980–2988.
[62] C. A. Blance, *Norie's Nautical Tables*. St Ives, U.K.: Imray, Laurie, Norie and Wilson Ltd, 2019.
[63] 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, May 2020.
[64] A. Adadi and M. Berrada, "Peeking inside the black-box: A survey on explainable artificial intelligence (XAI)," *IEEE Access*, vol. 6, pp. 52138–52160, 2018.
[65] A. B. Arrieta, N. Díaz-Rodríguez, J. D. Ser, A. Bennetot, S. Tabik, A. Barbado, S. Garcia, S. Gil-Lopez, D. Molina, R. Benjamins, R. Chatila, and F. Herrera, "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Inf. Fusion*, vol. 58, pp. 82–115, Jun. 2020.
[66] P. Molchanov, A. Mallya, S. Tyree, I. Frosio, and J. Kautz, "Importance estimation for neural network pruning," 2019, *arXiv:1906.10771*.
[67] J. Gou, B. Yu, S. J. Maybank, and D. Tao, "Knowledge distillation: A survey," *Int. J. Comput. Vis.*, vol. 129, no. 6, pp. 1789–1819, Jun. 2021.
[68] S. Benaïchouche, C. Le Goff, Y. Guichoux, F. Rousseau, and R. Fablet, "Unsupervised reconstruction of sea surface currents from AIS maritime traffic data using learnable variational models," in *Proc. IEEE Int. Conf. Acoust., Speech Signal Process. (ICASSP)*, Jun. 2021, pp. 4100–4104.
[69] S. Benaïchouche, C. Legoff, Y. Guichoux, F. Rousseau, and R. Fablet, "Unsupervised reconstruction of sea surface currents from AIS maritime traffic data using trainable variational models," *Remote Sens.*, vol. 13, no. 16, p. 3162, Aug. 2021.
[70] B. Ivanovic and M. Pavone, "The trajectron: Probabilistic multi-agent trajectory modeling with dynamic spatiotemporal graphs," in *Proc. IEEE/CVF Int. Conf. Comput. Vis. (ICCV)*, Oct. 2019, pp. 2375–2384.
[71] H. Akaike, "Information theory and an extension of the maximum likelihood principle," in *Selected Papers of Hirotugu Akaike*. New York, NY, USA: Springer, 1998, pp. 199–213.
[72] G. Schwarz, "Estimating the dimension of a model," *Ann. Statist.*, vol. 6, no. 2, pp. 461–464, Mar. 1978.
[73] E. J. Hannan and B. G. Quinn, "The determination of the order of an autoregression," *J. Roy. Stat. Soc. B, Methodol.*, vol. 41, no. 2, pp. 190–195, Jan. 1979.
[74] K. P. Burnham and D. R. Anderson, *Model Selection and Multimodel Inference: A Practical Information-Theoretic Approach*, vol. 2. New York, NY, USA: Springer, 2004.
[75] U. Anders and O. Korn, "Model selection in neural networks," *Neural Netw.*, vol. 12, no. 2, pp. 309–323, Mar. 1999.
[76] S. Arora, N. Cohen, and E. Hazan, "On the optimization of deep networks: Implicit acceleration by overparameterization," in *Proc. Int. Conf. Mach. Learn.*, 2018, pp. 244–253.
[77] Z. Allen-Zhu, Y. Li, and Z. Song, "A convergence theory for deep learning via over-parameterization," in *Proc. Int. Conf. Mach. Learn.*, 2019, pp. 242–252.

DUONG NGUYEN (Member, IEEE) received the Diplôme d'Ingénieur degree in machine learning from Télécom Bretagne, the M.Sc. degree (summa cum laude) in signal and image processing from the University of Rennes 1, and the Ph.D. degree in machine learning from IMT Atlantique.

He has published several articles on vessel trajectory analysis and anomaly detection. One of them has gained widespread adoption in industry and is being used as the foundation model to build maritime traffic anomaly detection systems. His research interests include the study of machine learning (deep learning) for time series modeling and analysis, with applications to dynamical system identification and maritime traffic surveillance.

RONAN FABLET (Senior Member, IEEE) received the M.Sc. and Engineering degrees in space aeronautics and applied mathematics from ISAE/SUPAERO, in 1997, and the Ph.D. degree in signal processing and computer vision from the University of Rennes/INRIA, in 2001. After an INRIA Postdoctoral Fellowship with Brown University, RI, USA, he was a full-time Research Scientist with Ifremer (the French Institute for Sea Research), from 2003 to 2007. He joined IMT Atlantique, in 2008. He is currently a Professor with the Mathematical and Electrical Engineering Department. He has been developing research activities at the interface between signal processing, machine learning, and marine science. He has led national and international programs (e.g., EU STREP AFISA, ANR MN EMOCEAN, and ANR ASTRID SESAME) and coauthored more than 200 articles and communications in peer-reviewed conferences and journals. His current research interests include physics-informed deep learning, data-driven approaches to inverse problems, ocean remote sensing, and maritime surveillance.