


# A Framework for Classifying Illegal Vessel Activities in the Maritime Domain

Vinicius D. do Nascimento <sup>1,†,‡</sup> , Tiago A. O. Alves <sup>2,‡</sup>, Claudio M. de Farias <sup>1,‡</sup> and Diego L. C. Dutra <sup>1,\*</sup>

<sup>1</sup> Affiliation 1; e-mail@e-mail.com

<sup>2</sup> Affiliation 2; e-mail@e-mail.com

\* Correspondence: e-mail@e-mail.com; Tel.: (optional; include country code; if there are multiple corresponding authors, add author initials) +xx-xxxx-xxx-xxxx (F.L.)

† Current address: Affiliation.

‡ These authors contributed equally to this work.

**Abstract:** A single paragraph of about 200 words maximum. For research articles, abstracts should give a pertinent overview of the work. We strongly encourage authors to use the following style of structured abstracts, but without headings: (1) Background: place the question addressed in a broad context and highlight the purpose of the study; (2) Methods: describe briefly the main methods or treatments applied; (3) Results: summarize the article's main findings; (4) Conclusions: indicate the main conclusions or interpretations. The abstract should be an objective representation of the article, it must not contain results which are not presented and substantiated in the main text and should not exaggerate the main conclusions.

**Keywords:** keyword 1; keyword 2; keyword 3 (List three to ten pertinent keywords specific to the article; yet reasonably common within the subject discipline.)

## 1. Introduction

O tráfego marítimo desempenha um papel crucial na conectividade global e no comércio internacional, servindo como uma artéria vital para a economia mundial. Milhões de toneladas de mercadorias são transportadas pelos oceanos todos os anos, impulsionando o crescimento econômico e o desenvolvimento dos países costeiros [1]. A importância estratégica do transporte marítimo é inegável, uma vez que é responsável por mais de 90% do comércio global [2]. Além disso, o tráfego marítimo é uma fonte essencial de emprego e receita para muitas nações, sustentando comunidades costeiras e promovendo o desenvolvimento socioeconômico.

No entanto, essa rede complexa de atividades marítimas também apresenta desafios significativos, como a necessidade de garantir a segurança das rotas marítimas, a proteção do meio ambiente marinho e a prevenção de atividades ilícitas. Portanto, o monitoramento eficaz do tráfego marítimo pelas autoridades marítimas (MA) locais torna-se fundamental para garantir a eficiência, a segurança e a sustentabilidade das operações marítimas em todo o mundo.

Através da consciência do domínio marítimo (MDA), as MA podem monitorar suas zonas econômicas exclusivas (EEZ) para garantir tráfego marítimo seguro e a exploração econômica da costa pelo país.

## 2. Materials and Methods

Nesta seção, apresentaremos a metodologia adotada neste trabalho, juntamente com uma discussão breve sobre conceitos essenciais, incluindo ensemble models, active learning e o framework JDL. Detalharemos a implementação do framework, os comportamentos de navegação, as regras definidas com base no conhecimento de especialistas, e a modelagem de active learning. Além disso, descreveremos os dados utilizados para treinamento e

**Citation:** Lastname, F.; Lastname, F.; Lastname, F. A Framework for Classifying Illegal Vessel Activities in the Maritime Domain. *Sensors* **2024**, *1*, 0. <https://doi.org/>

Received:

Revised:

Accepted:

Published:

**Copyright:** © 2024 by the authors. Submitted to *Sensors* for possible open access publication under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

teste, bem como o delineamento do experimento realizado. Por último, apresentaremos os principais resultados alcançados.

### 2.1. Trajectories

Uma trajetória é uma sequência de pontos no espaço e no tempo que descreve o movimento de uma embarcação. Cada ponto na trajetória é caracterizado por coordenadas espaciais (latitude e longitude) e um carimbo de tempo. A trajetória fornece uma visão detalhada do caminho percorrido pela embarcação ao longo do tempo. Além disso, temos os dados de velocidade (SOG) e a direção (COG) para onde a embarcação está indo. Uma trajetória considerada normal segue um padrão de movimento esperado baseado em comportamentos históricos ou regras estabelecidas. Por exemplo, por motivos de economia de combustível, normalmente embarcações de carga seguem uma rota comercial pré-definida em velocidade constante e sem desvios bruscos. No caso de embarcações de pesca, estas podem apresentar trajetórias características de pesca no locais onde estão efetivamente pescando. No caso de uma trajetória anômala, esta pode apresentar desvios significativos do padrão esperado, indicando comportamentos atípicos ou suspeitos. Na próxima seção iremos abordar esses tipos de comportamentos e como detectá-los.

Por fim, para facilitar o acesso aos dados das trajetórias, estas serão armazenadas como objetos serializados em um banco de dados SQLite. Este formato de armazenamento é eficiente e permite consultas rápidas e flexíveis. Cada trajetória é convertida em um formato de objeto que encapsula todos os pontos e informações associadas, e então é serializada e armazenada no banco de dados.

### 2.2. Vessel's Behaviour

Os comportamentos de embarcações referem-se às respostas observáveis de uma embarcação a estímulos externos e internos, sem necessariamente estar vinculadas a um objetivo específico. These behaviours can be routine, such as following a specific route for trade, or they can be indicative of specific activities, both legal and illicit. Understanding and detecting these behaviours is crucial for maritime safety, security, and surveillance.

The detection and monitoring of vessel behaviours primarily rely on various sensors. The most prominent among these is the AIS, which is a tracking system most used on ships and by vessel traffic services. AIS transponders automatically broadcast information, such as their position, speed, and navigational status, at regular intervals via a VHF transmitter. Other sensors include radars, which detect and map the position of ships based on reflected radio waves, and satellite imagery, which provides a view of maritime regions and can capture vessel movements.

Several behaviors have been identified and studied in maritime literature. In this study, we will include the most common behaviors, with a brief explanation of each in the next subsections. Cada modelo de comportamento ou regra apresentada a seguir foi implementada, terá como entrada trajetórias de embarcações e terão como saída um valor do tipo float representando a probabilidade do comportamento detectado. Nas próximas subseções apresentaremos breves explicações sobre cada comportamento e regra implementada no nosso framework.

#### 2.2.1. AIS Spoofing

Malicious alteration of AIS signals to disguise a vessel's identity or location. Some research has demonstrated that spoofing AIS messages is possible [3] and can mask or favor illegal activities, disrupt monitoring systems, and create navigational risks. Alguns trabalhos tentaram detectar o AIS spoofing como em [4] an architecture for detecting AIS data spoofing in an online data stream was proposed. The architecture stores in a hash the MMSI of the vessels and calculates the average speed between the vessel trajectory points. Similarmente, para detectar o AIS spoofing nós iremos fazer esse mesmo cálculo, só que ao invés de agrupar por MMSI faremos por trajetória.

### 2.2.2. Encounters at Sea

Two or more vessels meeting in open waters, possibly for the transfer of goods or people. This can be a sign of trade but also of activities like smuggling. Apesar de um encontro ser uma comportamento suspeito, nem sempre ele é. Em áreas de exploração de petróleo existem muitas embarcações de apoio que fazem transporte de pessoal, petróleo e material de apoio. É muito comum nessas áreas o chamado ship to ship, que é transferência de carga entre embarcações no mar. Nestes casos a aproximação entre embarcações é considerada uma situação normal, assim como próximo aos portos, onde as embarcações podem ficar atracadas próxima as outras.

Alguns trabalhos como [5–8] propuseram soluções para detecção desses encontros levando em consideração a intenção desses encontros. No entanto, por questão de simplicidade, neste caso nós iremos utilizar somente as posições AIS e os timestamps para detectar as embarcações que se posicionaram a menos de 200 metros entre si. Para isso nós utilizamos um sistema de codificação baseado em geohash. Geohash é um sistema de codificação que transforma coordenadas geográficas (latitude e longitude) em uma cadeia de caracteres alfanuméricos [9]. A principal característica do Geohash é que ele divide o espaço geográfico em uma grade hierárquica de células quadradas de diferentes tamanhos. Cada caractere adicional na cadeia aumenta a precisão da localização. Uma implementação em python de geohash é a biblioteca H3 [10]. A H3 é uma biblioteca desenvolvida pela Uber para particionar o globo terrestre em uma grade hexagonal hierárquica. O H3 combina conceitos de geohashing com a flexibilidade e eficiência das grades hexagonais, que são preferidas por várias propriedades geométricas favoráveis em comparação com grades quadradas. Portanto, para a detecção de encontros neste trabalho, nós agrupamos os hashes gerados por cada posição do AIS utilizando a resolução 9 (200 metros) e utilizamos a janela de tempo de 4h.

### 2.2.3. Dark Ships

Vessels that turn off their AIS transponders to avoid detection, often indicating an intent to engage in illegal activities. Essa prática algumas vezes pode ser confundida com situações onde o navio passa áreas com interferência de sinal ou de mal tempo. Portanto, essas áreas precisam ser conhecidas para podermos inferir esse tipo de comportamento.

In the literature, we found a variety of attempts to solve this problem, mainly by using active sensors for detection. For non-collaborative systems, works such as [11,12] used radars, which when fused with AIS have both AIS and radar trajectories associated and an anomaly can be detected if they diverge. Outros trabalhos tentaram resolver o problema utilizando somente o sensor AIS [13,14]. No entanto, utilizando somente o sensor AIS, o que podemos saber são os gaps de transmissão da trajetória e não se realmente o navio desligou o transponder AIS. Desta forma, em nosso trabalho nós consideraremos esses gaps de transmissão juntamente com os históricos de transmissão dos geohashes para saber se o navio deveria ter transmitido um sinal AIS daquela posição ou não.

### 2.2.4. Fishing Trajectories

Fishing vessels in situations of normal movement from one region to another, they normally behave as cargo vessels, trying to follow an optimized trajectory. However, when the ship arrives in fishing spots and starts its target activity, it needs to perform specific maneuvers. For each type of fishing activity, there are characteristic maneuvers of the vessel like a example that can be seen in Figure ?? . Patterns of movement indicate that a vessel is engaged in fishing and this can be used to monitor fishing activities in restricted areas. The detection fishing trajectories was addressed by several works using AIS sensors data in Machine Learning models [15–21]. Neste trabalho, nós usamos uma solução similar proposta por [17,18,21] onde uma rede neural recorrente (RNN) é utilizada na série temporal da trajetória. As dimensões utilizadas no modelo são a velocidade, a diferença angular entre os pontos, a diferença de tempo, a diferença de distância e a aceleração.

A distância de entre dois pontos pode ser calculada pela distância de haversine, que é usada para calcular a distância entre dois pontos na superfície de uma esfera, dada a latitude e a longitude desses pontos. Logo a distância em milhas náuticas pode ser calculada pela fórmula 1.

$$d = 2r \cdot \arcsin \left( \sqrt{\sin^2 \left( \frac{\varphi_2 - \varphi_1}{2} \right) + \cos(\varphi_1) \cos(\varphi_2) \sin^2 \left( \frac{\lambda_2 - \lambda_1}{2} \right)} \right) \quad (1)$$

Onde:

- $d$  é a distância entre os dois pontos, em milhas náuticas.
- $r$  é o raio da Terra, aproximadamente 3.440 milhas náuticas.
- $\varphi_1$  e  $\varphi_2$  são as latitudes dos dois pontos, em radianos.
- $\lambda_1$  e  $\lambda_2$  são as longitudes dos dois pontos, em radianos.

Com relação a diferença de tempo entre dois pontos, nós utilizamos a unidade de minuto. Para calcularmos a velocidade e aceleração em nós, basta dividirmos a distância e a velocidade respectivamente pelo tempo em horas. Finalmente, para calcularmos a diferença angular entre os pontos, nós podemos utilizar a fórmula 2.

$$\Delta\theta = \arctan 2(\sin(\theta_1 - \theta_2), \cos(\theta_1 - \theta_2)) \quad (2)$$

Where  $\theta_1$  and  $\theta_2$  are the directions at each point, and  $\arctan 2$  is a variation of the arc tangent function that takes into account the sign of both the sine and cosine components to determine the quadrant of the resulting angle. Essa mesma estratégia para detecção de trajetórias de pesca foi utilizada em [21].

Após o criação do modelo de para detecção de trajetórias de pesca, nós utilizamos como dataset para treinamento as trajetórias de pesca do Global Fishing Watching (GFW) [22].

#### 2.2.5. Loitering

Loitering involves a ship or boat spending an abnormally long time in the same area or performing slow movements around a larger area without displacing itself a significant distance. Vessels that remain stationary or move in circles, indicating possible waiting for another ship or engaging in activities like illegal fishing.

Several works in loitering detection have generally been addressed as detecting an object on the scene for a period of time that exceeds a given threshold [23–25]. No entanto, para detectarmos uma trajetória de loitering, nós inovadoramente criamos um método utilizando modelo RNN, assim como na detecção de trajetórias de pesca. Primeiramente, utilizando o dataset do marine cadastre [26] com dados AIS da costa dos estados unidos, nós detectamos encontros entre navios. Para navios se encontrarem, estes primeiramente precisam desacelerar até parar totalmente. Esta desaceleração e até a permanência com baixa velocidade por algumas horas em uma localidade são características de trajetória loitering que podem preceder um encontro entre embarcações. Detectando as trajetórias que realizaram esses encontros, nós extraímos essas trajetórias para realizar o treinamento do nosso modelo RNN. Em seguida, nós criamos uma arquitetura RNN com as mesma dimensões que utilizamos para detectarmos comportamentos de pesca (velocidade, diferença de distância, diferença de tempo, diferença angular e aceleração). Finalmente, treinamos o modelo com as trajetórias extraídas dos encontros do dataset do marine cadastre.

#### 2.2.6. Anomalous Navigation

Desvios dos padrões típicos de navegação podem indicar que uma embarcação está perdida, enfrentando problemas técnicos ou envolvida em atividades ilícitas. A maioria dos estudos na literatura tratam as anomalias nos trajetos de embarcações de forma binária, verificando os trajetos mais comuns e considerando anômalas as trajetórias que se desviam desses padrões [27–30].

No presente estudo, ao invés de monitorar uma área específica, iremos observar toda a costa. Portanto, seria necessário detectar o padrão de normalidade de cada canal de navegação ao longo da costa. Para simplificar, não utilizaremos o conceito de anomalia como um comportamento, mas sim como uma atividade.

Algumas combinações de comportamentos descritos anteriormente, quando associadas a determinadas regras definidas por especialistas, podem ser consideradas atividades anômalas. Por exemplo, uma embarcação navegando com um MMSI inválido e exibindo comportamento de spoofing pode ser classificada como envolvida em uma atividade anômala. Desta forma, adotaremos uma classe de atividade anômala para situações onde uma combinação de comportamentos e regras, julgadas como anômalas pelos especialistas ocorrerem.

### 2.3. Knowledge based rules

Em diversos sistemas de monitoramento do domínio marítimo é comum a utilização de conjuntos de regras baseadas em conhecimento, desenvolvidas por especialistas na área. Essas regras são formuladas com base em anos de experiência e conhecimento aprofundado sobre padrões de navegação, comportamentos esperados das embarcações, sinais indicadores de atividades suspeitas e leis.

As regras baseadas em conhecimento normalmente são criadas para identificar desvios dos padrões normais de operação marítima. Elas abrangem uma variedade de critérios, como trajetórias inesperadas, variações anômalas de velocidade, uso de identificadores inválidos (MMSI), e comportamentos indicativos de spoofing ou operações ilícitas. Os comportamentos cinemáticos, que referem-se ao padrão de movimento de uma embarcação, incluindo aspectos como velocidade, direção e quaisquer mudanças em seu curso ou velocidade ao longo do tempo [31]. Algumas agências, como a Agência Europeia de Segurança Marítima (EMSA) [32] possuem modelos chamados de Monitoramento Automatizado de Comportamento (ABM), alguns exemplos de comportamentos podem ser vistos na imagem ?? . O objetivo do ABM é detectar e destacar automaticamente comportamentos da embarcação que possam ser de interesse ou preocupação. No entanto, o uso excessivo desse tipo regra pode tornar inviável já que os especialistas precisam verificar cada disparo de regra. Portanto, em sistemas de monitoramento marítimo podem-se criar milhares de regras cinemáticas facilmente, sendo o operador responsável por verificar cada disparo e a atualização dessas regras.

Entretanto, essas regras quando combinadas com outros comportamentos de navegação, podem-se tornar ferramentas importantes na detecção de uma atividade ilegal. Por exemplo, o especialista sabe sobre áreas onde a pesca é proibida, então este pode criar polígonos delimitando essas áreas e caso alguma embarcação com identificação de pesca adentre essas áreas apresentando uma trajetória de pesca pode-se inferir então que há uma atividade de pesca ilegal.

Neste trabalho, nós iremos utilizar as seguintes regras criadas pelo especialista a serem aplicadas em cada trajetória:

- A validação do MMSI do navio, tanto na sintaxe quanto na semântica;
- Um crawler para buscar na internet através do MMSI e identificar o nome, bandeira e tipo. Desta forma, podemos criar os seguintes atributos: *flag\_brazil*, *flag\_other*, *flag\_unknow*, *type\_fishing*, *type\_other*, *type\_unknow*;
- Passagem por dentro de área de ancoragem;
- Distância da Costa;
- Passagem por dentro de Área de Proteção Ambiental;
- Passagem dentro de área de FPSO;
- Média de velocidade da trajetória;

Todas as trajetórias são aplicadas a essas regras e as saídas são normalizadas. No caso de respostas da sim ou não, os valores são convertidos para 1 ou 0 respectivamente. Finalmente, após serem aplicadas essas regras as trajetórias, essas são inseridas no banco de dados Sqlite para posterior utilização no metamodelo e active learning.



## 2.4. Vessel's Activities

As Autoridades Marítimas (MA) são normalmente responsáveis por uma variedade de atividades em suas águas jurisdicionais, incluindo monitoramento costeiro, proteção da vida humana no mar, aplicação da lei, proteção de Áreas de Preservação Ambiental (APA), proteção de infraestrutura crítica, proteção ambiental e resposta a problemas ambientais crimes [33]. Dado o grande número de atividades a serem monitoradas, são necessárias soluções em larga escala, pois é praticamente impossível para agentes humanos realizar essas tarefas para toda a zona econômica exclusiva (ZEE). Por exemplo, a ZEE do Brasil é a décima maior do mundo, com 3,5 milhões de quilômetros quadrados [34], e consequentemente, o uso de sensores e aplicações automatizadas para alertar os operadores humanos tornou-se fundamental.

As atividades ilegais praticadas por embarcações em seu domínio marítimo são preocupações das MA pelo mundo, podemos citar algumas, por exemplo: a pesca ilegal, proteção de Áreas Marinhas Protegidas (MAPAs), proteção de infraestruturas críticas, combate ao tráfico (armas, drogas, seres humanos e dinheiro) e situações de segurança de navegação. No entanto, identificar essas atividades de forma automatizada é um grande desafio, pois não basta somente identificar o comportamento do navio, mas também a intenção humana (comandante, tipulação, armador, etc) para qual atividade o navio está sendo usado.

Em [35] foi proposta uma abordagem baseada em ontologia para detectar a atividade da embarcação com base no comportamento da embarcação. O autor defende que a modelagem e exploração de informação incerta é um facilitador fundamental para o reconhecimento da atividade da embarcação, dada a grande variabilidade do comportamento da embarcação e a incerteza inerente à sua intenção. Portanto, o comportamento da embarcação faz parte de sua atividade observada. Desta forma, neste trabalho consideramos o comportamento das embarcações detectadas em conjunto com as regras criadas pelo especialista para detectar a atividade da embarcação.

Detectar todas essas atividades seria um escopo inviável para este trabalho, tanto pela falta de dados históricos e sensores, quanto pelo amplo escopo. Portanto, nós iremos focar nas atividades onde existem dados acadêmicos para que possamos treinar nossos modelos e apresentar testes que demonstrem que o nosso framework funciona e pode ser escalável. Assim, este trabalho classificará as atividades como:

- *Pesca Ilegal*: A pesca é considerada ilegal quando uma embarcação realiza pesca dentro de uma área onde a pesca não é permitida; ou; Quando navios realizam pesca dentro de ZEE estrangeira a seu país;
- *Atividade Suspeita*: consideraremos como suspeita, atividades que tem potencial de ser uma atividade ilegal. Por exemplo, navios realizando encontros a mais de 12 milhas náuticas da costa podem estar transferindo carga ou pessoas. Um outro exemplo seria, navios sem identificação apresentando trajetórias de pesca entre 180 e 200 milhas náuticas;
- *Atividade Anômala*: uma atividade é considerada anômala quando um conjunto de comportamentos e/ou regras são considerados fora do normal, mas que a princípio não indicam uma atividade ilícita;
- *Atividade Normal*: quando um a uma atividade de navegação é considerada normal para aquela localidade;

Ao longo de todo o trabalho, iremos nos referir à essas 4 atividades como os principais rótulos de classificação das atividades. Na seção de caso de estudo definiremos critérios objetivos para que o especialista possa classificar cada atividade.

## 2.5. Ensemble models

Ensemble models é uma técnica de aprendizado de máquina que envolve a combinação de vários modelos para aumentar a precisão e o desempenho na resolução de problemas complexos. A ideia central por trás da aprendizagem em conjunto é que, ao combinar as previsões de vários modelos, é possível obter resultados mais precisos e ro-

bustos do que qualquer modelo único poderia fornecer. Historicamente, o conceito de aprendizagem em conjunto começou a ganhar destaque na década de 1990, com o desenvolvimento de algoritmos como bagging e boosting. Estes métodos demonstraram que a agregação de múltiplos modelos de aprendizagem, cada um contribuindo com a sua própria perspectiva ou experiência, pode levar a um desempenho significativamente melhorado, particularmente em tarefas complexas de classificação e regressão. A premissa básica é que embora um único modelo possa ter suas limitações e estar sujeito a erros específicos, a combinação de vários modelos pode compensar essas fraquezas, resultando em maior precisão e confiabilidade nas previsões [36]. Como um dos nossos objetivos é reduzir as incertezas ao classificar as atividades, esta estratégia irá nos ajudar, permitindo que os pontos fortes de cada modelo de comportamento compensem os pontos fracos dos outros, gerando previsões com menor incerteza.

A nossa principal estratégia será utilizar uma técnica ensemble model conhecida como stacking. No stacking, as previsões geradas por um conjunto de modelos, chamados modelos básicos, são usadas como entrada para um novo modelo, chamado metamodelo ou modelo de segundo nível. Este metamodelo é treinado especificamente para fazer a previsão final. A eficácia do empilhamento é particularmente notável em cenários onde existe uma heterogeneidade significativa entre os modelos base. Essa diversidade permite que o metamodelo desenvolva uma capacidade aprimorada de reconhecer a maneira mais eficiente de integrar previsões individuais dos modelos básicos, resultando em precisão superior na previsão final. Como metamodelo, nós iremos utilizar o modelo random forest neste trabalho.

No nosso caso, iremos utilizar como modelos base os comportamentos de navegação apresentados na seção 2.2 e as regras criadas pelo especialista na seção 2.3. A saída desses modelos serão utilizados como entrada no metamodelo, que no nosso caso será o modelo random forest. No entanto, mesmo combinando os modelos de comportamentos com as regras dos especialista, ainda teremos o problema da falta de rótulos para as atividades. Na próxima seção, onde falaremos sobre active learning, abordaremos melhor esse problema.

## 2.6. Active learning

Avanços significativos na tecnologia de coleta e armazenamento de dados permitem o acúmulo de grandes quantidades de dados em uma ampla gama de aplicações do mundo real. Rotular estes dados, por outro lado, é caro porque requer esforço e conhecimento humano. A aprendizagem semissupervisionada preocupa-se com métodos para melhorar o desempenho da aprendizagem, utilizando dados não rotulados, além de dados rotulados automaticamente. A ideia básica por trás do aprendizado semissupervisionado é que, ao combinar uma pequena quantidade de dados rotulados com uma grande quantidade de dados não rotulados, o modelo pode aprender a fazer previsões melhores do que faria apenas com dados rotulados [37][38].

A aprendizagem ativa é um processo no qual métodos semissupervisionados são combinados com a intervenção de um especialista humano. O modelo de aprendizado de máquina no aprendizado ativo seleciona os exemplos mais informativos do conjunto de dados não rotulados e pede a um especialista humano para rotulá-los. Isto é especialmente útil quando a rotulagem de dados é cara ou demorada. No cenário de aprendizado ativo, o algoritmo de aprendizado de máquina não apenas aprende com dados previamente rotulados, mas também tem a capacidade de solicitar rótulos para exemplos específicos do conjunto de dados não rotulado. A ideia é que o modelo melhore mais rapidamente se conseguir obter rótulos para os exemplos que considera mais incertos ou informativos. Algoritmos de aprendizado ativo normalmente usam algum tipo de estratégia de consulta para tornar isso eficiente. Eles podem, por exemplo, selecionar exemplos onde as previsões do modelo são mais incertas ou exemplos que estão na fronteira de decisão entre duas classes. Quando temos um grande conjunto de dados não rotulados e um especialista humano que pode fornecer rótulos precisos, mas o tempo do especialista é limitado, o aprendizado ativo pode ser muito eficaz [37].

Dadas as suas características únicas, nosso problema apresenta uma oportunidade excepcional para a aplicação da Aprendizagem Ativa. Primeiramente começamos com dados não rotulados e o especialista precisará realizar um "cold start" nos rótulos para fornecer um contexto. Este é um cenário comum em muitas aplicações do mundo real, onde a coleta de dados é abundante, mas a rotulagem é dispendiosa devido à necessidade de experiência humana. Após o cold start realizado pelo especialista, o active learning já pode ser treinado e iniciar as consultas por rótulos para o usuário. O modelo selecionará dados que possuem a maior probabilidade aumentar a acurácia da predição do modelo. Desta forma, a cada rótulo respondido pelo usuário os pesos do modelo são atualizados e consequentemente a acurácia do modelo aumenta. No nosso caso, o modelo utilizado no active learning foi o random forest. Na próxima seção apresentaremos como foi organizado o nosso framework baseado no JDL, desde a aquisição de dados até a predição da atividade.

## 2.7. *A framework for vessel illegal activities classification*

O modelo JDL (Joint Directors of Laboratories Data Fusion Model) é uma estrutura conceitual desenvolvida para guiar a integração e a fusão de dados provenientes de múltiplas fontes. Este modelo foi introduzido pelo Joint Directors of Laboratories, um grupo de pesquisa e desenvolvimento do Departamento de Defesa dos Estados Unidos, para padronizar e melhorar os processos de fusão de dados em sistemas de vigilância e defesa [39].

Como o nosso problema envolve a utilização de sensores, fusão de modelos de comportamento e regras criadas pelo especialista, uma certa complexidade é criada e requer o uso formal de um framework para organizar o nosso conhecimento. Nos baseamos no modelo JDL devido à sua capacidade de fornecer uma abordagem estruturada e hierárquica para a fusão de dados. Este modelo permite que informações provenientes de múltiplas fontes sejam integradas de maneira eficiente, resultando em um quadro situacional mais abrangente e preciso. Tal característica é fundamental em contextos onde decisões críticas precisam ser tomadas com base em dados coletados de diversos sensores e sistemas. Desta forma, nosso framework apresenta 5 níveis: pré-processamento, fusão de nível de objeto, consciência situacional, avaliação de impacto e apoio a decisão. Na figura 1 podemos o processo do framework.

Portanto, nas próximas seções apresentaremos detalhes sobre cada nível do nosso framework.

### 2.7.1. *Preprocessing*

O pré-processamento dos dados é uma etapa crucial para garantir a qualidade e a integridade das análises subsequentes. Os dados podem ser obtidos de diferentes fontes e sensores, como por exemplo, AIS, radares, imagens de satélites e etc. Os dados utilizados neste estudo são oriundos do AIS em sua forma bruta. Cada linha do dataset é composta pelos dados de latitude, longitude, speed over ground (SOG), course over ground (COG), mmsi, nome do navio e um timestamp. Desta forma, temos as principais etapas do pré-processamento realizado:

- Carregamento dos dados AIS cru: Os dados brutos do AIS são inicialmente carregados diretamente dos arquivos disponíveis utilizando a biblioteca pandas do python;
- Limpeza de dados faltantes ou inválidos: Após o carregamento, é realizada uma limpeza das linhas que possuem dados faltantes ou incorretos. Uma trajetória será composta por ao menos 3 pontos, portanto os casos faltantes serão removidos;
- Transformação dos dados AIS em geo dataframes: Os dados limpos são então transformados em geo dataframes utilizando a biblioteca geopandas. Esta transformação permite uma manipulação mais eficiente e precisa das informações geoespaciais, facilitando a análise e visualização dos dados;
- Conversão em trajetórias com moving pandas: Finalmente, os geo dataframes são convertidos em trajetórias utilizando a biblioteca MovingPandas. Esta conversão é essencial para manipular os movimentos das embarcações de forma temporal e



Framework for vessel illegal activity classification

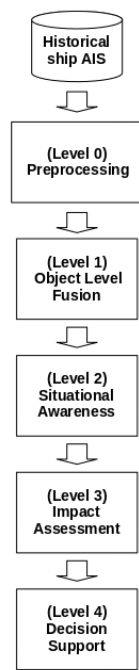


Figure 1. Framework process.

espacial, permitindo análises detalhadas das suas trajetórias e comportamentos de navegação.

Este pré-processamento assegura que os dados utilizados estejam no formato de trajetórias geopandas e geo dataframes de forma a serem utilizados como entrada nos processos posteriores.

2.7.2. Object level fusion

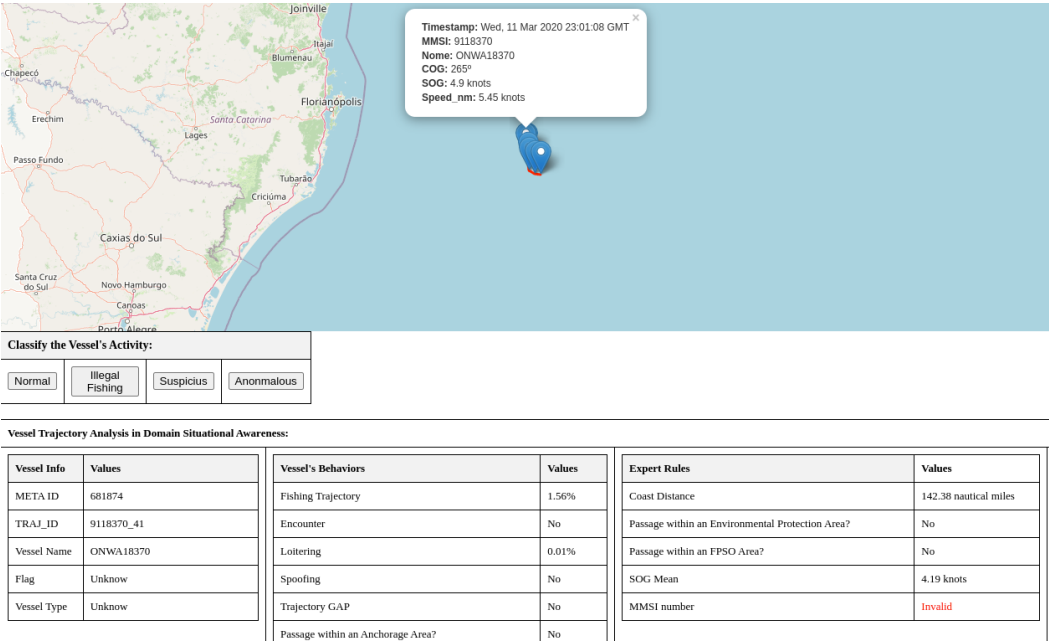
Neste nível recebemos como entrada as trajetórias das embarcações e aplicamos os modelos dos diferentes tipos detecção de comportamentos da seção 2.2 e também as regras dos especialista da seção 2.3. Caso estivéssemos utilizando outro tipo de sensor, como por exemplo o radar, poderíamos realizar a fusão em nível de objeto, fundindo dados do AIS com radar em relação à um mesmo contato.

Desta forma, os diferentes algoritmos aplicados sobre as trajetórias e gerarão uma saída com os respectivos resultados de detecção, gerando uma correlação. Esses resultados e essa correlação serão utilizados no próximo nível.

2.7.3. Situational awareness

Este nível é responsável por combinar as saídas dos comportamentos e regras do especialista, e dar um contexto global da situação da trajetória. Este nível também é responsável por armazenar os dados em um banco de dados, preparar os dados para serem utilizados no active learning e preparar a apresentação dos dados de forma intuitiva para o usuário.

Para combinar os dados, nós primeiramente criamos um dataframe onde cada linha possui colunas com dados sobre as saídas de cada modelo e regra do especialista, correlacionada com a respectiva trajetória. Em seguida, esses dados são armazenamos em um banco de dados (sqlite). A utilização de um banco de dados facilita a manipulação e agiliza a carga desses dados para execução dos modelos. Feito isto, o próximo nível pode usar como entrada esses dados para sua execução.



**Figure 2.** An user interface for trajectory evaluation by expert.

2.7.4. Impact assessment

O nível de avaliação de impacto foca na interpretação dos dados pelo usuário, treinamento/atualização do metamodelo e a previsão das atividades. Primeiramente, este nível é responsável por criar o active learning e o cold start do especialista. Os dados do metamodelo são carregados do banco de dados, caso não haja ao menos 100 rotulações feita pelo especialista a interface do usuário para cold start é utilizada, selecionando 100 trajetórias aleatoriamente e apresentando ao usuário para que este as rotule. Ao atingir 100 trajetórias rotuladas, a interface do usuário para active learning será utilizada, onde primeiramente o modelo random forest será treinado com os dados rotulados e então o modelo selecionará trajetórias que possuem maior probabilidade de aumentar a acurácia do sistema.

A estratégia de consulta para selecionar os dados mais informativos que utilizamos no activie learning foi a *uncertainty sampling*. Nesta estratégia, o modelo seleciona as amostras sobre os quais está mais incerto. Em seguida, as trajetórias são apresentadas para usuário para que este as rotule. Na figura 2 podemos ver um exmeplo da interface com o usuário. Após cada rotulagem feita pelo usuário, o modelo é atualizado. Após um número considerado adequado de classificações (discutiremos isso na seção de caso de estudo), os modelo pode inferir as atividades para cada trajetória recebida. Desta forma, agora o especialista pode trabalhar somente com as trajetórias que geraram predições de interesse, como por exemplo, as de pesca ilegal e atividades suspeitas. Na próxima seção, abordaremos esse tópico.

2.7.5. Decision support

Explicar o que é nível de Decision support; O que é feito? citar exemplos;

2.8. Synthetic dataset

Explicar o que é um dataset sintético; Explicar em que situações ele pode ser utilizado; Explicar no nosso caso porque utilizaremos (falta de rotulos para ativiade ilegais); Explicar como construímos nosso Dataset: -trajetorias de pesca do GFW tranlatadas aleatoriamente para regioao da costa do brasil; -trajetorias de encontros do marinecadastre tranlatadas para regioao da costa do brasil; -Atualizacao da data do AIS; -Concatenacao com AIS da costa sudeste do brasil do periodo de 2019 a 2021; - Mostrar figura com dataset e suas utilizacoes;

## 2.9. Case study

Para testar o nosso framework iremos aplicalo em um caso prático na Costa do Brasil; Iremos utilizar dados AIS coletados no periodo de 2019 a 2020 da costa brasileira;

- Mostrar figura com o mapa do Brasil, com as situações de Atividades de pesca ilegal, suspeitas e anomalias detectadas;
- Mostrar tabela com embarcações: MMSI, regras e comportamentos apresentados;
- Explicar as situações de cada atividade e pq é esse framework é útil;

## 2.10. Activities Prediction

No contexto de atividades ilegais de embarcações encontradas na literatura, podemos citar algumas como pesca ilegal, tráfico (drogas, armas e pessoas), poluição hídrica, pirataria, invasão de áreas protegidas e etc;

## 2.11. Active Learning Model Training

Um operador ira avaliar 100 trajetorias aleatorias como cold start; Em seguida, o algoritmo do active learning será treinado e sugerirá trajetórias para o especialista para este classifique a trajetória; O especialista avaliará 500 trajetórias sugeridas pelo active learning;

## 2.12. Test of framework Vessel's Illegal Activity Detection in Maritime Domain

Após o treinamento, serão inferidas todas as trajetórias do dataset, incluída as trajetórias da costas sul, nordeste e norte; Serão selecionadas 150 trajetórias de cada classe inferidas pelo framework (AL); Mostrar o mapa com as predicoes de cada classe; O especialista avaliará cada das 600 trajetórias e as classificará; Com base na inferencia das atividades pelo modelo e as respostas do especialista, podemos avaliar os resultados a seguir;

- 
- apresentar o funcionamento com o operador;
- 

## 2.13. Results

Tabela com precision, recall e f1 de cada classe; Mostrar imagem da matriz de confusão;

## 3. Discussion

Mostrar graficos shaps e explicar o pq dos resultados; O modelo se comportou com acc acima de 90% para detectar as classes pesca ilegal, suspeita e normal; No caso da classe anomala, ficou proximo de 98%. Devido quantidade de boias e farois sendo confundidos com navios proxima a costa, nós eliminamos trajetórias abaixo de 2 knots e a uma distancia de 10 mn da costa. Dados incoerentes também são apontados como anomalias, como por exemplo, velocidade e posicoes impossiveis. Pelo grafico shap, pode-se ver que a regra de "mmsi valid" tem alta influencia na classificacao de anomalo; Normalmente boias e farois nao tem mmsi valido, assim com alguns navios que saem da normalidade;

## 4. Conclusions

## 5. Patents

This section is not mandatory, but may be added if there are patents resulting from the work reported in this manuscript.

**Author Contributions:** For research articles with several authors, a short paragraph specifying their individual contributions must be provided. The following statements should be used “Conceptualization, X.X. and Y.Y.; methodology, X.X.; software, X.X.; validation, X.X., Y.Y. and Z.Z.; formal analysis, X.X.; investigation, X.X.; resources, X.X.; data curation, X.X.; writing—original draft preparation, X.X.; writing—review and editing, X.X.; visualization, X.X.; supervision, X.X.; project administration, X.X.; funding acquisition, Y.Y. All authors have read and agreed to the published version of the manuscript.”, please turn to the [CRediT taxonomy](#) for the term explanation. Authorship must be limited to those who have contributed substantially to the work reported.

**Funding:** Please add: “This research received no external funding” or “This research was funded by NAME OF FUNDER grant number XXX.” and and “The APC was funded by XXX”. Check carefully that the details given are accurate and use the standard spelling of funding agency names at <https://search.crossref.org/funding>, any errors may affect your future funding.

**Institutional Review Board Statement:** In this section, you should add the Institutional Review Board Statement and approval number, if relevant to your study. You might choose to exclude this statement if the study did not require ethical approval. Please note that the Editorial Office might ask you for further information. Please add “The study was conducted in accordance with the Declaration of Helsinki, and approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval).” for studies involving humans. OR “The animal study protocol was approved by the Institutional Review Board (or Ethics Committee) of NAME OF INSTITUTE (protocol code XXX and date of approval).” for studies involving animals. OR “Ethical review and approval were waived for this study due to REASON (please provide a detailed justification).” OR “Not applicable” for studies not involving humans or animals.

**Informed Consent Statement:** Any research article describing a study involving humans should contain this statement. Please add “Informed consent was obtained from all subjects involved in the study.” OR “Patient consent was waived due to REASON (please provide a detailed justification).” OR “Not applicable” for studies not involving humans. You might also choose to exclude this statement if the study did not involve humans.

Written informed consent for publication must be obtained from participating patients who can be identified (including by the patients themselves). Please state “Written informed consent has been obtained from the patient(s) to publish this paper” if applicable.

**Data Availability Statement:** We encourage all authors of articles published in MDPI journals to share their research data. In this section, please provide details regarding where data supporting reported results can be found, including links to publicly archived datasets analyzed or generated during the study. Where no new data were created, or where data is unavailable due to privacy or ethical restrictions, a statement is still required. Suggested Data Availability Statements are available in section “MDPI Research Data Policies” at <https://www.mdpi.com/ethics>.

**Acknowledgments:** In this section you can acknowledge any support given which is not covered by the author contribution or funding sections. This may include administrative and technical support, or donations in kind (e.g., materials used for experiments).

**Conflicts of Interest:** Declare conflicts of interest or state “The authors declare no conflicts of interest.” Authors must identify and declare any personal circumstances or interest that may be perceived as inappropriately influencing the representation or interpretation of reported research results. Any role of the funders in the design of the study; in the collection, analyses or interpretation of data; in the writing of the manuscript; or in the decision to publish the results must be declared in this section. If there is no role, please state “The funders had no role in the design of the study; in the collection, analyses, or interpretation of data; in the writing of the manuscript; or in the decision to publish the results”.

**Abbreviations**

The following abbreviations are used in this manuscript:

MDPI	Multidisciplinary Digital Publishing Institute
DOAJ	Directory of open access journals
TLA	Three letter acronym
LD	Linear dichroism

**Appendix A**

*Appendix A.1*

The appendix is an optional section that can contain details and data supplemental to the main text—for example, explanations of experimental details that would disrupt the flow of the main text but nonetheless remain crucial to understanding and reproducing the research shown; figures of replicates for experiments of which representative data are

shown in the main text can be added here if brief, or as Supplementary Data. Mathematical proofs of results not central to the paper can be added as an appendix.

**Table A1.** This is a table caption.

Title 1	Title 2	Title 3
Entry 1	Data	Data
Entry 2	Data	Data

**Appendix B**

All appendix sections must be cited in the main text. In the appendices, Figures, Tables, etc. should be labeled, starting with “A”—e.g., Figure A1, Figure A2, etc.

**References**

1. Stopford, M. *Maritime Economics*; Routledge, 2009.

2. Forum, W.E. Worlds busiest ocean shipping routes, 2024.

3. Androjna, A.; Perkovic, M.; Pavic, I.; Miskovic, J. Ais data vulnerability indicated by a spoofing case-study. *Applied Sciences (Switzerland)* **2021**, *11*. <https://doi.org/10.3390/app11115015>.

4. Kontopoulos, I.; Spiliopoulos, G.; Zissis, D.; Chatzikokolakis, K.; Artikis, A. Countering Real-Time Stream Poisoning: An Architecture for Detecting Vessel Spoofing in Streams of AIS Data. *IEEE*, 8 2018, pp. 981–986. <https://doi.org/10.1109/DASC/PiCom/DataCom/CyberSciTec.2018.00139>.

5. Shahir, H.Y.; Glasser, U.; Shahir, A.Y.; Wehn, H. Maritime situation analysis framework: Vessel interaction classification and anomaly detection. Institute of Electrical and Electronics Engineers Inc., 12 2015, pp. 1279–1289. <https://doi.org/10.1109/BigData.2015.7363883>.

6. Matossian, M.; Laurila, P.; Blanchet, C. Detecting dark vessels: Radar satellite-based monitoring of illegal activities at sea. *Sea Technology* **2020**, *61*, 30 – 32. Cited by: 0.

7. Uney, M.; Millefiori, L.M.; Braca, P. Prediction of Rendezvous in Maritime Situational Awareness. In Proceedings of the 2018 21st International Conference on Information Fusion (FUSION), 2018, pp. 622–628. <https://doi.org/10.23919/ICIF.2018.8455816>.

8. Sharma, A.; Shekhar, S. Analyzing Trajectory Gaps to Find Possible Rendezvous Region. *ACM Trans. Intell. Syst. Technol.* **2022**, *13*. <https://doi.org/10.1145/3467977>.

9. Niemeyer, G. Geohash. <http://geohash.org>, 2008. Accessed: 2024-05-20.

10. Uber Technologies, I. H3: A Hexagonal Hierarchical Spatial Index. <https://h3geo.org>, 2018. Accessed: 2024-05-20.

11. Zhang, T.; Zhao, S.; Cheng, B.; Chen, J. Detection of AIS closing behavior and MMSI spoofing behavior of ships based on spatiotemporal data. *Remote Sensing* **2020**, *12*. <https://doi.org/10.3390/rs12040702>.

12. Nikolic, D.; Stojkovic, N.; Popovic, Z.; Tosic, N.; Lekic, N.; Stankovic, Z.; Doncov, N. Maritime over the horizon sensor integration: HFSWR data fusion algorithm. *Remote Sensing* **2019**, *11*. <https://doi.org/10.3390/RS11070852>.

13. Mazzarella, F.; Vespe, M.; Alessandrini, A.; Tarchi, D.; Aulicino, G.; Vollero, A. A novel anomaly detection approach to identify intentional AIS on-off switching. *Expert Systems with Applications* **2017**, *78*, 110–123. <https://doi.org/10.1016/j.eswa.2017.02.011>.

14. Kontopoulos, I.; Chatzikokolakis, K.; Zissis, D.; Tserpes, K.; Spiliopoulos, G. Real-time maritime anomaly detection: detecting intentional AIS switch-off. *International Journal of Big Data Intelligence* **2020**, *7*, 85–96, <https://www.inderscienceonline.com/doi/pdf/10.1504/IJBIDI.2020.107375>. <https://doi.org/10.1504/IJBIDI.2020.107375>.

15. Kontopoulos, I.; Makris, A.; Tserpes, K. A deep learning streaming methodology for trajectory classification. *ISPRS International Journal of Geo-Information* **2021**, *10*. <https://doi.org/10.3390/ijgi10040250>.

16. Arasteh, S.; Tayebi, M.A.; Zohrevand, Z.; Glässer, U.; Shahir, A.Y.; Saeedi, P.; Wehn, H. Fishing Vessels Activity Detection from Longitudinal AIS Data. Association for Computing Machinery, 11 2020, pp. 347–356. <https://doi.org/10.1145/3397536.3422267>.

17. Zhang, Z.; Huang, L.; Peng, X.; Wen, Y.; Song, L. Loitering behavior detection and classification of vessel movements based on trajectory shape and Convolutional Neural Networks. *Ocean Engineering* **2022**, *258*. <https://doi.org/10.1016/j.oceaneng.2022.111852>.

18. Ferreira, M.D.; Spadon, G.; Soares, A.; Matwin, S. A Semi-Supervised Methodology for Fishing Activity Detection Using the Geometry behind the Trajectory of Multiple Vessels. *Sensors* **2022**, *22*. <https://doi.org/10.3390/s22166063>.

19. Zhao, L.; Shi, G. Maritime Anomaly Detection using Densitybased Clustering and Recurrent Neural Network. *The Journal of Navigation* **2019**, *72*. <https://doi.org/10.1017/S0373463319000031>.

20. Pedroche, D.S.; Amigo, D.; García, J.; Molina, J.M. Architecture for trajectory-based fishing ship classification with AIS data. *Sensors (Switzerland)* **2020**, *20*, 1–21. <https://doi.org/10.3390/s20133782>.

21. Do Nascimento, V.D.; Alves, T.A.O.; Dutra, D.L.C.; Kundu, S. A Comparative Study of Fishing Activity Detection Approaches in Maritime Surveillance. In Proceedings of the 2023 Congress in Computer Science, Computer Engineering, & Applied Computing (CSCE), 2023, pp. 173–179. <https://doi.org/10.1109/CSCE60160.2023.00033>.



22. Watch, G.F. Global Fishing Watch Fishing Effort Dataset. <https://globalfishingwatch.org/datasets-and-code/fishing-effort/>, 2024. Accessed: 2024-05-20. 592
23. Li, W.; Zhang, D.; Sun, M.; Yin, Y.; Shen, Y. Loitering Detection Based on Trajectory Analysis. *Institute of Electrical and Electronics Engineers Inc.*, 5 2016, pp. 530–533. <https://doi.org/10.1109/ICICTA.2015.136>. 593
24. Lu, R.; Yang, H.; Zhu, J.; Wu, S.; Wang, J.; Bull, D. Hierarchical video summarization with loitering indication. *Institute of Electrical and Electronics Engineers Inc.*, 4 2016. <https://doi.org/10.1109/VCIP.2015.7457818>. 594
25. Patino, L.; Ferryman, J. Loitering Behaviour Detection of Boats at Sea. *IEEE Computer Society*, 8 2017, Vol. 2017-July, pp. 2169–2175. <https://doi.org/10.1109/CVPRW.2017.269>. 595
26. Cadastre, M. Marine Cadastre National Viewer Datasets. <https://marinecadastre.gov/data/>, 2024. Accessed: 2024-05-20. 596
27. Sidibé, A.; Shu, G. Study of automatic anomalous behaviour detection techniques for maritime vessels. *Journal of Navigation* **2017**, 70, 847–858. <https://doi.org/10.1017/S0373463317000066>. 597
28. Pan, X.; Wang, H.; Cheng, X.; Peng, X.; He, Y. Online detection of anomaly behaviors based on multidimensional trajectories. *Information Fusion* **2020**, 58, 40–51. <https://doi.org/https://doi.org/10.1016/j.inffus.2019.12.009>. 598
29. Laxhammar, R.; Falkman, G. Conformal Prediction for Distribution-Independent Anomaly Detection in Streaming Vessel Data. In *Proceedings of the Proceedings of the First International Workshop on Novel Data Stream Pattern Mining Techniques*, New York, NY, USA, 2010; *StreamKDD '10*, p. 47–55. <https://doi.org/10.1145/1833280.1833287>. 599
30. Lu, N.; Liang, M.; Yang, L.; Wang, Y.; Xiong, N.; Liu, R.W. Shape-Based Vessel Trajectory Similarity Computing and Clustering: A Brief Review. In *Proceedings of the 2020 5th IEEE International Conference on Big Data Analytics (ICBDA)*, 2020, pp. 186–192. <https://doi.org/10.1109/ICBDA49040.2020.9101322>. 600
31. Weintrit, A. *Marine Navigation and Safety of Sea Transportation: Navigational Problems*; CRC Press: New York, NY, USA, 2013. 601
32. European Maritime Safety Agency (EMSA). <https://www.emsa.europa.eu>. Accessed: 2024-05-24. 602
33. de Faria, J.A.P.M. A consciência situacional marítima (CSM) e a Marinha do Brasil. *NAVAL WAR COLLEGE JOURNAL; Vol 18, No 1 (2012): REVISTA DA ESCOLA DE GUERRA NAVAL* **2012**. 603
34. Amazonia Azul. <https://www.marinha.mil.br/delareis/?q=amazoniazul>. Accessed: 2024-05-27. 604
35. Zocholl, M.; Iphar, C.; Joussemme, A.L.; Ray, C. Ontology-based approach for vessel activity recognition. *Institute of Electrical and Electronics Engineers Inc.*, 2021, Vol. 2021-September. <https://doi.org/10.23919/OCEANS44145.2021.9705824>. 605
36. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer Series in Statistics, Springer New York, 2009. <https://doi.org/10.1007/978-0-387-84858-7>. 606
37. Zhou, Z.H. *Ensemble Methods: Foundations and Algorithms*; Chapman and Hall/CRC, 2012. 607
38. Zhu, X.; Goldberg, A.B. *Introduction to Semi-Supervised Learning*; Morgan & Claypool Publishers, 2009. 608
39. Hall, D.L.; Llinas, J. An introduction to multisensor data fusion. *Proceedings of the IEEE* **1997**, 85, 6–23. 609

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content. 610