Data in Brief 65 (2026) 112629

Contents lists available at ScienceDirect

Data in Brief

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

Data Article

MMDEC: Multimodal maritime dataset on the English channel

Tristan Averty[a,*], Ioannis Nasios[b], Cyril Ray[a], Nikos Piliouras[b]

[a] Institut de Recherche et d'Études Navales (IRENav), ENSAM / École navale, BCRM de Brest, CC 600, 29240 Brest Cedex 9, France
[b] Nodalpoint Systems, Pireos 203-205, Athens, Greece

ARTICLE INFO

Article history:
Received 28 January 2026
Revised 19 February 2026
Accepted 20 February 2026
Available online 25 February 2026

Dataset link: MMDEC: Multimodal Maritime Dataset on the English Channel (Original data)

Keywords:
Multi-sensor data
Maritime mobility dataset
Trajectory reconstruction
AIS
Sentinel-1 SAR
SAR-based vessel detection

ABSTRACT

The rapid proliferation of tracking sensors—ranging from vessel and vehicle tracking systems to smartwatches, cameras, and Earth observation sensors—has led to an unprecedented influx of high-frequency, high-volume data. Yet, despite this abundance, many trajectories remain incomplete, contain errors, or are entirely missing. A vast reservoir of tracking data remains unexplored or underutilized, holding valuable insights that could enhance monitoring and decision-making. The MMDEC dataset aims to serve as a resource for both teaching and the research community, and is specifically designed to support advanced maritime trajectory analysis and inference. Its core contribution lies in the explicit integration of SAR-based, AI-driven ship detections and attribute estimates with AIS vessel reports and multiple contextual data layers, enabling direct comparison between transmitted vessel positions and independently observed radar targets. MMDEC aggregates and harmonizes multi-source data including AIS streams, satellite imagery, meteorological and oceanographic fields, port locations, and sea state and marine protected area boundaries over western Celtic Sea, the English Channel, and part of the North Sea. The dataset covers a continuous three-month period from 1 July to 30 September 2023. Its multimodal content and explicit spatial–temporal alignment provide a multi-sensor benchmark for evaluating

* Corresponding author.
E-mail address: tristan.averty@ecole-navale.fr (T. Averty).

https://doi.org/10.1016/j.dib.2026.112629

algorithms designed to enhance, validate, or infer maritime trajectories from heterogeneous observational sources.

Specifications Table

Subject	Computer Sciences
Specific subject area	Maritime surveillance, maritime mobility, ship detection
Type of data	Table, Image, Geometric objects
Data collection	Integration of Sentinel-1 SAR satellite imagery (Copernicus) with AIS streams, meteorological records (ERA5), oceanographic fields (CMEMS), bathymetry, CROSS maritime operations, marine protected areas, traffic separation schemes, wind farms, and submarine cables.
Data source location	Western Celtic Sea, the English Channel, and part of the North Sea.
Data accessibility	Repository name: MMDEC: Multimodal Maritime Dataset on the English Channel Data identification number: 10.5281/zenodo.17491518 Direct URL to data: https://doi.org/10.5281/zenodo.17491518
Related research article	none

1. Value of the Data

- The dataset provides a unique combination of AIS and Sentinel-1 SAR data collected over the English Channel between July and September 2023. This integration allows a direct link between vessel position reports and radar backscatter imagery, offering a valuable reference for maritime monitoring, Earth observation, and data fusion research.
- The dataset provides a strong foundation for the development of artificial intelligence training datasets focused on ship detection and classification. By combining verified AIS information with corresponding Sentinel-1 SAR imagery, it offers the essential components needed to create labeled samples, evaluate preprocessing methods, and design annotation strategies for machine learning and computer vision applications in maritime surveillance.
- The simultaneous integration of AIS and SAR data allows detailed studies of AIS reliability, signal gaps, and vessel identification accuracy by comparing transmitted vessel positions with radar-detected targets in one of the world's busiest maritime regions.
- The data can be reused to analyze maritime traffic density, vessel routes, and seasonal activity patterns, supporting environmental assessments, maritime safety analysis, and transport planning studies.
- The dataset follows open-access and FAIR principles, is well-documented, and organized in standard formats (.csv, .jpeg, .geojson, .shp), enabling easy reuse, reproducibility, and integration into existing remote sensing and machine-learning research workflows.
- The integration of AIS data, SAR-based AI ship detections, and operational and environmental context enables the study of specific maritime events, such as search-and-rescue operations or anomalous vessel behavior, and supports vessel monitoring applications relevant to environmental oversight, including the analysis of illegal, unreported, and unregulated (IUU) fishing activities and the identification of vessels operating without AIS ("dark vessels").

2. Background

The rapid proliferation of tracking and Earth observation technologies has generated massive streams of spatio-temporal data. However, despite this abundance, mobility analysis is still

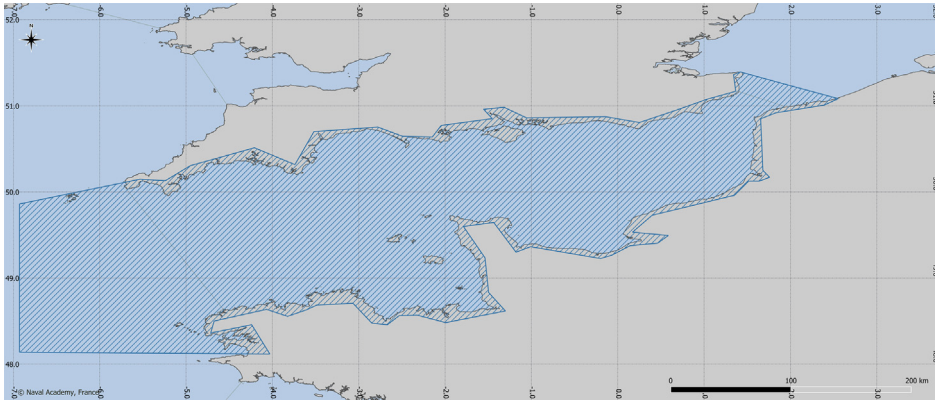

Fig. 1. Area of interest retained for the creation of the MMDEC dataset.

constrained by incomplete trajectories, heterogeneous data quality, and fragmented sensor coverage. Large volumes of potentially valuable tracking information therefore remain underexploited. Within this context, the MUSIT[1] project develops advanced AI-driven and spatio-temporal fusion techniques to reconstruct, refine, and semantically enrich trajectories derived from diverse sensing infrastructures. Its objective is to unlock latent information contained in multimodal tracking archives and to support robust monitoring and analytics across maritime and urban mobility domains [1].

This paper describes the creation of an integrated satellite-centric maritime dataset over a well-defined Area of Interest, illustrated by the Fig. 1, and covering the western Celtic Sea, the English Channel, and the southern North Sea. The reason we are interested in this area is the wealth of events and elements found there: significant maritime traffic supplying English, Dutch, and French commercial ports, illegal immigration, illegal fishing in several marine protected areas, the existence of wind farms, and the presence of submarine cables. Focusing on a 3-month period (July, August and September) in 2023, we integrated free Sentinel-1 SAR data processed through the SatShipAI model with a broad set of contextual layers including AIS records, meteorological conditions, search-and-rescue activity, port locations, protected areas, submarine cable routes, and traffic separation schemes. Satellite data preparation (acquisition, chunking, model inference, and structuring) and ancillary data preprocessing were carried out jointly, ensuring thorough documentation and interoperability.

3. Data Description

Table 1 summarizes the MMDEC dataset components grouped into satellite-derived data, vessel tracking data (AIS), operational and regulatory layers, and environmental layers, while the following subsections describe each component in detail, beginning with the Sentinel-1 SAR–based satellite datasets.

3.1. Area of interest

The area of interest described in section “BACKGROUND” and illustrated by Fig. 1 has been added as a GeoJSON file (`AOI.geojson`) containing 1 polygon.

[1] https://cordis.europa.eu/project/id/101182585.

Table 1
Overview of MMDEC dataset components by data category (Satellite, AIS, Operations, Infrastructure and Environment).

Category	Dataset component	Format	Primary purpose
SAT	`Dataset_S1_chunks`	JPEG	Sentinel-1 SAR image chunks (VV, VH, averaged polarization) for localized ship analysis.
SAT	`Dataset_S1_chunks_corners`	CSV	Geographic corner coordinates of SAR image chunks for spatial referencing.
SAT	`Dataset_S1_SatShipAI_outputs`	CSV	AI-based ship detections and attributes derived from Sentinel-1 SAR imagery.
SAT	`Dataset_S1_SatShipAI_excluding_areas`	SHP / GeoJSON	Wind-farm exclusion zones used to filter SAR-based detections.
AIS	`Dataset_AIS_POS`	Parquet	AIS position messages for vessel trajectory reconstruction.
AIS	`Dataset_AIS_SPEC`	Parquet	AIS status messages providing vessel identity and characteristics.
AIS	`Dataset_AIS_ShipTypes`	CSV	Mapping between AIS numerical ship type codes and vessel categories.
OPS	`Dataset_SECMAR`	Parquet	Maritime surveillance and rescue operations for event-based analysis.
INFRA	`Dataset_PORTS`	Parquet	Port locations supporting origin–destination and proximity analyses.
INFRA	`Dataset_TSS`	Parquet	Traffic separation schemes defining regulated navigation corridors.
INFRA	`Dataset_WIND_FARMS`	Parquet	Locations and attributes of offshore wind farms.
INFRA	`Dataset_MPA`	Parquet	Marine protected areas supporting regulatory and environmental studies.
INFRA	`Dataset_CABLES`	Parquet	Submarine power and telecommunication cables representing critical infrastructure.
ENV	`Dataset_ERA5`	Parquet	Meteorological variables providing atmospheric context.
ENV	`Dataset_CMEMS_PHY`	Parquet	Physical oceanographic variables (temperature, salinity, currents).
ENV	`Dataset_CMEMS_WAV`	Parquet	Wave-based oceanographic variables describing sea state conditions.
ENV	`Dataset_BATHYMETRY`	Parquet	High-resolution bathymetric depth data for seabed context.

3.2. Sea areas

The area of interest described in section “BACKGROUND” actually covers three clearly defined maritime areas as defined by the International Hydrographic Organization (IHO): the Celtic Sea, the English Channel, and the North Sea. It was in our interest to consider the Celtic Sea to include the Ushant Separation Scheme and the North Sea to obtain information off the coast of Dunkirk. We then extracted these three marine areas from the IHO data [2], which resulted in a Parquet file (`Dataset_SEAS.parquet`) containing 3 polygons. The attributes of these entities are basically the bounding boxes and the area of the regions. Furthermore, it is worthy to note that polygons are not intersections (i.e. sea areas that “touch” our area of interest are retained and are kept intact).

3.3. Satellite dataset

The satellite dataset is based on imagery from the Copernicus Sentinel-1 Synthetic Aperture Radar (SAR) satellites and is organized into several ZIP files, each containing specific data products and associated information for the area of interest and over the three-month period de-

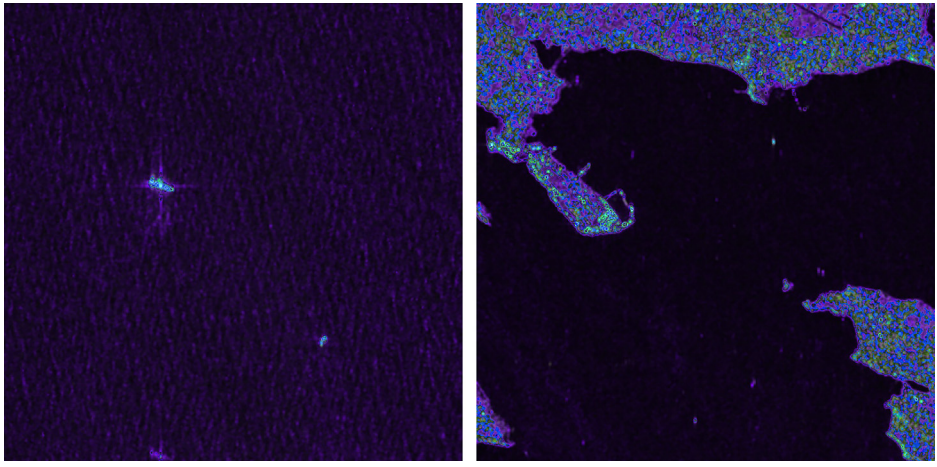

Fig. 2. Two example image chunks from different locations within the same Sentinel-1 SAR product are shown. Large vessels in the left image are readily detected, whereas the smaller vessels in the right image are more challenging to identify.

scribed in section "BACKGROUND". For information, in this space-time window, this represents 230 satellite products. We then have the following files:

- `Dataset_S1_chunks.zip` (230 folders)
  ZIP archive structured by satellite product name, with each subfolder containing multiple JPEG image chunks derived from the original Sentinel-1 Ground Range Detected (GRD) products. Each image has dimensions of 800 × 800 × 3 pixels, though some edge chunks vary slightly due to product boundaries. The three channels represent VV, VH, and (VV + VH)/2 polarization values (examples depicted in Fig. 2).
- `Dataset_S1_chunks_corners.zip` (230 files)
  ZIP archive containing one CSV file for each satellite product, providing the geospatial corner coordinates of all corresponding image chunks. Each CSV file lists the image chunk name along with the latitude and longitude (in decimal degrees) of its four corner points, starting from the top-left corner and proceeding clockwise.
- `Dataset_S1_SatShipAI_outputs.zip` (231 files)
  One CSV file per satellite product, presenting the AI-based ship detection results generated by the SatShipAI models [3,4]. For every detected ship, the ship's latitude and longitude (in decimal degrees), its X and Y position within the original Sentinel-1 product, the detection probability output from the model, the estimated ship length (in meters), the predicted ship type (one of Cargo, Fishing, Other, Passenger, or Tanker), and the associated classification probability are provided. It is important to emphasize that the SAR detections and vessel classifications included in this dataset are outputs of the SatShipAI deep learning models and do not constitute manually validated ground truth. The detections represent probabilistic model predictions derived from Sentinel-1 SAR imagery and may include false positives, missed detections, or classification uncertainties. Users should therefore treat these outputs as model-generated annotations rather than authoritative labels.
  An additional CSV file (`extra_info.csv`) is included, providing metadata for all Sentinel-1 products used. The recorded parameters comprise the product name, image dimensions (width and height in pixels), acquisition date and time, North azimuth, vertical pixel spacing, and the geographic coordinates (latitude and longitude in decimal degrees) of the four image corners, listed sequentially from the top-left corner in a clockwise direction. Three supplementary parameters are also provided for incidence angle estimation: a longitude multiplier, a latitude multiplier, and an additive offset value.

- `Dataset_S1_SatShipAI_excluding_areas.zip` (5 files)
  Three wind-farm exclusion zones within the AOI. For convenience, the data are provided in both Shapefile and GeoJSON formats. These zones were used to exclude detections located within or in close proximity to offshore wind-farm boundaries. It should be noted that these manually created exclusion zones are very similar to the official wind farms described in subsection "Location of wind farms", but their presence is explained by the fact that the detections were made before the complete dataset was created.

3.4. AIS dataset

The Automatic Identification System (AIS) is a maritime communication system used to enhance the safety and efficiency of navigation at sea. It allows vessels to automatically exchange information such as their identity, position, speed, and course with other ships and coastal authorities. By providing real-time data, AIS helps reduce the risk of collisions, improves traffic monitoring, and supports maritime security and search and rescue operations. Today, AIS has become an essential tool for vessel trajectories monitoring.

There are many types of AIS messages (27), but the main ones that are interesting within the project are position messages (types 1, 2, 3, 18, 19, and 27) and status messages (types 5 and 24).

3.5. Positions messages (types 1, 2, 3, 18, 19 and 27)

The complete AIS database we had consisted of all messages for the year 2023 worldwide. As we focused on a specific time period and an area of interest covering the English Channel for the reasons mentioned above, we pre-processed the position messages (types 1, 2, 3, 18, 19 and 27) geographically and temporally to retain only those that met the criteria. This is possible because, by definition, latitude and longitude are fields contained in position messages. The result of this processing is a Parquet file (`Dataset_AIS_POS.parquet`) of 19,014,229 messages for 25,130 unique MMSI.[2] All the features contained in these messages are well described on the Navigation Center (U.S. Department of Homeland Security) website [5] and transcribed in the Table 2. Moreover, a density map of all the position messages is depicted in Fig. 3. Thus, the main commercial routes taken by ships can be easily seen.

3.6. Status messages (types 5 and 24)

Status messages (types 5 and 24) contain information about the ships. For example, information such as the vessel name, destination with estimated time of arrival, and physical characteristics including length, width, and draught can be found. The Parquet file (`Dataset_AIS_SPEC.parquet`) contains 13,558,007 status messages for 23,958 unique MMSI. All features contained in these messages are well described on the Navigation Center (U.S. Department of Homeland Security) website [5] are transcribed in Table 3.

3.7. Ship types

In the `ShipType` attribute of AIS status messages, the values are numerical, ranging from 0 to 255, and each numerical value corresponds to a specific type of vessel. To make this mapping

[2] The MMSI (Maritime Mobile Service Identity) is a unique 9-digit code used to identify maritime mobile stations and ships at sea.

Table 2
Description of the features contained in the AIS position messages dataset.

Features	Description
`Date`	Date when the report was generated by the electronic position system
`Source`	Type of AIS receiver • "eee-land": Terrestrial • "eee-sat": Satellite
`MessageType`	• 1: Scheduled position report; Class A shipborne mobile equipment • 2: Assigned scheduled position report; Class A shipborne mobile equipment • 3: Special position report, response to interrogation; Class A shipborne mobile equipment • 18: Standard position report for Class B shipborne mobile equipment to be used instead of Messages 1, 2, 3 • 19: Extended position report for Class B shipborne mobile equipment; contains additional static information • 27: Class A and Class B "SO" shipborne mobile equipment outside base station coverage
`Mmsi`	MMSI number of the ship sending the position report
`NavigationStatus`	• 0: under way using engine • 1: at anchor • 2: not under command • 3: restricted maneuverability • 4: constrained by her draught • 5: moored • 6: aground • 7: engaged in fishing • 8: under way sailing • 9: reserved for future amendment of navigational status • 10: reserved for future amendment of navigational status • 11: power-driven vessel towing astern (regional use) • 12: power-driven vessel pushing ahead or towing alongside (regional use) • 13: reserved for future use • 14: AIS-SART (active), MOB-AIS, EPIRB-AIS • 15: undefined (default)
`Latitude`	Latitude (in degrees)
`Longitude`	Longitude (in degrees)
`PositionAccuracy`	Position accuracy (PA) in accordance with the rules: • True (PA $\leq$ 10 m) • False (PA $>$ 10 m or default)
`CourseOverGroundDegrees`	Course over ground (in degrees) : 511 indicates not available (default)
`SpeedOverGround`	Speed over ground (in knots)
`RateOfTurn`	• 0 to $+126°$: turning right at up to 708 deg per min or higher • 0 to $-126°$: turning left at up to 708 deg per min or higher • $+$ 127°: turning right at $>$5 deg per 30 s (No TI available) • -127: turning left at $>$5 deg per 30 s (No TI available) • -128 (80 hex) indicates no turn information available (default).
`TrueHeadingDegrees`	True heading (in degrees) : 511 indicates not available (default)
`chunk_folder`	Satellite product (folder) name that contains chunks in which the ship declaring its position can be found. `NaN` if such a satellite product does not exist.
`id_chunk`	List of chunk names (without the ".jpg") in which the ship declaring its position can be found. `NaN` if such chunks do not exist.

possible, we have provided a CSV file (`Dataset_AIS_ShipTypes.csv`) containing 256 lines from a transcription of a file from the Marine Cadastre Project [6].

The Fig. 4 depicts all the ship types that can be declared in an AIS status message as well as the proportion of each ship type in the AIS status message dataset (`Dataset_AIS_SPEC.parquet`). We only kept one message per MMSI, so we assumed that ships do not change type over time.

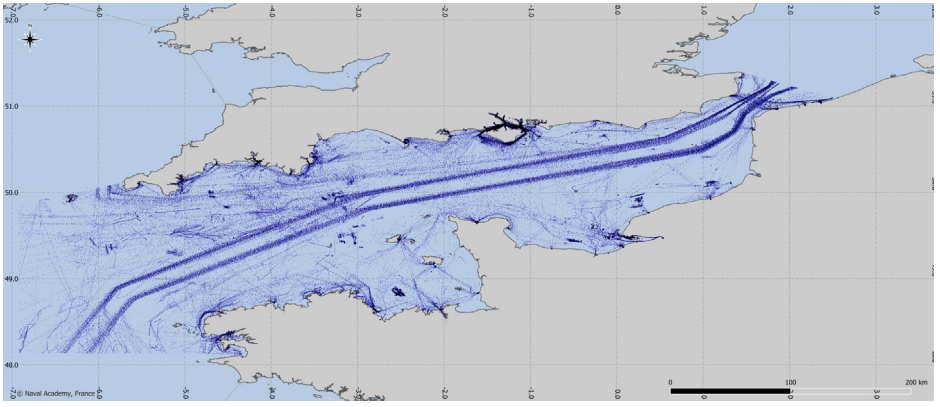

Fig. 3. Overview of vessel positions along the 3-month period.

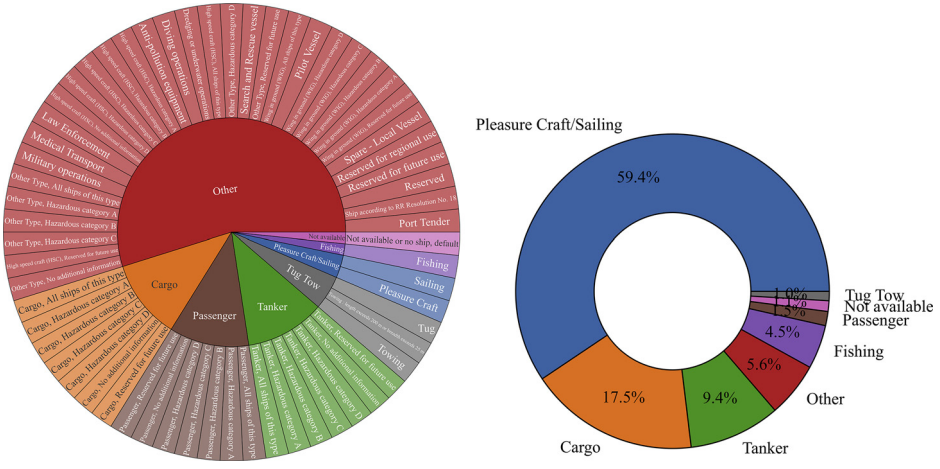

Fig. 4. Left: Sunburst chart showing all the ship types that can be declared in an AIS status message. / Right: Pie chart showing the proportion of each ship type in the MMDEC dataset (based on the status messages).

3.8. Location of ports

Adding port locations to the MMDEC dataset can significantly enrich the analysis of ship trajectories constructed from AIS data or satellite images. We extracted the ports within the area of interest from the worldwide list of seaports [7], which resulted in a Parquet file (`Dataset_PORTS.parquet`) containing 399 geographic entities. Among the attributes of each port, one can find the geographical location and the name of the port as well as their referencing (by the United Nations Economic Commission for Europe (UNECE) and/or Electronic Recording and Reporting (ERS)).

3.9. Maritime operations of regional operating surveillance and rescue Centres

The objective of the Regional Operating Surveillance and Rescue Centres[3] is to ensure safety in maritime areas under French responsibility. Their main missions are Search And Rescue (SAR)

[3] The French acronym is CROSS for “Centres régionaux opérationnels de surveillance et de sauvetage”.

Table 3
Description of the features contained in the AIS status messages dataset.

Features	Description
`Date`	Date when the report was generated by the electronic position system
`Source`	Type of AIS receiver • "eee-land": Terrestrial • "eee-sat": Satellite
`MessageType`	• 5: Scheduled static and voyage related vessel data report, Class A shipborne mobile equipment • 24: Additional data assigned to an MMSI Part A: Name Part B: Static Data
`Mmsi`	MMSI number of the ship sending the position report
`ImoNumber`	• 0: not available (default) • 0000,000,001–0000,999,999: not used • 0001,000,000–0009,999,999: valid IMO number; • 0010,000,000–1073,741,823: official flag state number.
`CallSign`	6 or 7 characters • @@@@@@@: not available (default) • Craft associated with a parent vessel should use "A" followed by the last 6 digits of the MMSI of the parent vessel.
`VesselName`	Maximum 20 characters • @@@@@@@@@@@@@@@@@@@@: not available (default) • The name should be as shown on the station radio license.
`ShipType`	Types described in the `Dataset_AIS_ShipType.csv`
`DimensionToBow`	In meters
`DimensionToStern`	In meters
`DimensionToPort`	In meters
`DimensionToStarboard`	In meters
`Draught10thMetres`	Draught in 1/10 m • 255: draught of 25.5 m or greater • 0: not available (default)
`Destination`	Maximum 20 characters • @@@@@@@@@@@@@@@@@@@@: not available
`EtaMonth`	Estimated month of arrival (1–12) • 0: not available (default)
`EtaDay`	Estimated day of arrival (1–31) • 0: not available (default)
`EtaHour`	Estimated hour of arrival (0–23) • 24: not available (default)
`EtaMinute`	Estimated minute of arrival (0–59) • 60: not available (default)
`PositionFixType`	Type of electronic position fixing device • 0: undefined (default) • 1: GPS • 2: GLONASS • 3: combined GPS/GLONASS • 4: Loran-C • 5: Chayka • 6: integrated navigation system • 7: surveyed • 8: Galileo • 9–14: not used • 15: internal GNSS

at sea, monitoring maritime traffic (e.g. provide assistance to vessels in distress) or state action at sea (Illegal fishing, clandestine immigration, illegal trafficking of goods, etc.). Since 2018, operations, as well as the humans statistics and the vessels involved in them, have been freely available on the French public data platform and are updated daily [8]. In the case of our MMDEC dataset, it seemed worthwhile to add these operations. Indeed, spatial and temporal joins could be performed between the recorded operations and the AIS positions/statuses reported by vessels in the area, which would also be potentially visible on SAR satellite image chunks.

We extracted the operations carried out during the three-month period and in the area of interest described in section "BACKGROUND", which resulted in a Parquet file (`Dataset_SECMAR.parquet`) containing 1769 operations. All the features are described on

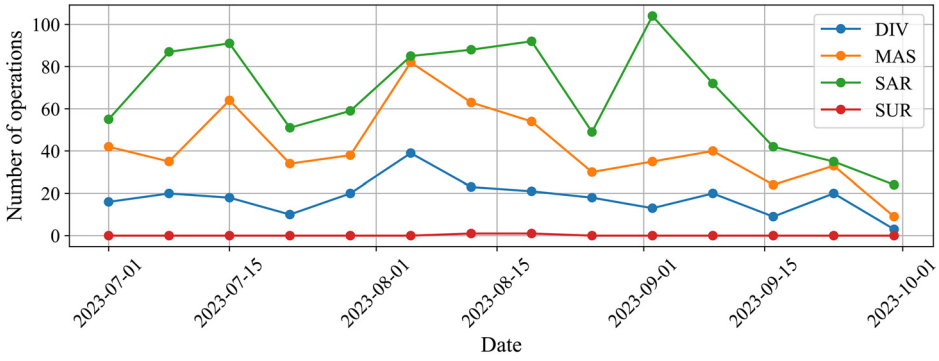

Fig. 5. Weekly evolution of the number of CROSS operations by category.

the GitHub of the French Ministry of Ecological Transition.[4] Fig. 5 provides an overview of the weekly evolution of SAR (Search And Rescue), MAS (Maritime Assistance Service), SUR (Ship safety) and DIV (Others) operations contained in the dataset.

3.10. Traffic separation schemes

Traffic separation schemes (TSS) are used to facilitate, organize, and secure vessel traffic in areas where it is dense or dangerous. As such, there is generally one route for each direction of traffic with specific areas for fishing or for vessels to cross paths. This greatly reduces the number of collisions or groundings in sensitive areas. Given the traffic density in the studied area of interest, a few TSS are present in the region. We therefore retained only the TSS that intersect our area of interest from the full set of TSS listed by Naval Hydrographic and Oceanographic Service[5] and published on the French public data platform [9]. The resulting Parquet file (`Dataset_TSS.parquet`) contains 19 geographic entities representing the areas separating two traffic lanes or a traffic lane from a coastal navigation area or the central area of a roundabout.

3.11. Location of wind farms

Wind farms were included in this dataset primarily because they are areas where navigation is restricted, while they may also be relevant in other contexts, such as instances of protests by fishermen. We therefore extracted data from the EMODnet project on wind farms [10] in our area of interest, which resulted in the creation of a Parquet file (`Dataset_WIND_FARMS.parquet`) containing 3 polygons for the 3 wind farms in the area: Rampion (UK), Fécamp (FR) and Saint-Brieuc (FR). Among the attributes of this dataset, the number of turbines, the total power, and the year in which the wind turbines began to deliver electricity. It should be noted that the two French wind farms began operating in 2024 but were under construction during the three months of 2023 selected. Therefore, the constraints associated with their construction, as well as their visibility on satellite images, must be taken into account.

[4] https://mtes-mct.github.io/secmar-documentation/schema.html#operations.
[5] The French acronym is SHOM for “Service hydrographique et océanographique de la Marine”.

Table 4
List of weather features with units.

Column name	Attribute	Unit
sp	Surface pressure	hPa
msl	Mean sea level pressure	hPa
tcc	Total cloud cover	[0, 1]
u10	10 m U wind component	m/s
v10	10 m V wind component	m/s
t2m	2 m temperature	°C
d2m	2 m dewpoint	°C

3.12. Marine protected areas

Marine protected areas (MPA) serve to protect biodiversity by preserving sensitive habitats (coral reefs, seagrass beds, mangroves) in order to restore fishery resources and strengthen resilience to climate change. As such, in many MPA, fishing is regulated or prohibited, and destructive gear is banned. It is therefore interesting to add them to a dataset that aims to study such use cases. We extracted the MPA contained within the area of interest from the Protected Planet website [11], which resulted in a Parquet file (`Dataset_MPA.parquet`) of 168 geographic entities. All features are described on the Protected Planet website [11].

3.13. Telecommunication and power submarine cables

Submarine cables play an essential role in the modern world. Laid at the bottom of the oceans, communication cables transmit >95 % of total data volume, enabling telephone calls, internet access, financial transactions, and data sharing on a global scale. Their usefulness is therefore strategic, as satellites are not currently a viable alternative. As for power supply cables, their usefulness is obvious when it comes to supplying electricity to islands, for example. The vulnerability of these cables to accidents, human activity, or acts of sabotage makes them critical infrastructure, the protection of which has become an international issue. These considerations highlight the importance of including this information in a dataset focused on maritime traffic.

For this purpose, we combined two data sources from the European EMODnet project, one for power cables [12] and one for telecommunications cables [13]. We only kept the cables located within the area of interest, which made it possible to create a Parquet file (`Dataset_CABLES.parquet`) containing 272 geographic entities. For each entity (cable), there are 75 attributes as each European provider uses its own attribute names with the primary attribute representing the type of the cable (POWER or COMM).

3.14. Meteorology

Weather information can be important when studying maritime trajectories. Purely wind-powered activities are only feasible when there is wind, ships may be diverted in the event of a storm, navigation statistics can be studied according to temperature or cloud cover, etc. For this reason, we downloaded the weather data from the Copernicus Climate Change Service [14] for the 3-month time period within the area of interest. Initially in NetCDF format, the file was decompressed to obtain a Parquet file (`Dataset_ERA5.parquet`) containing 537,510 weather observations. The spatial resolution is 0.25° (∼ 28 km) and the temporal resolution is hourly. In addition to the location (latitude and longitude), weather attributes we have retained are listed in the Table 4. Additionally, the box plots of these attributes shown in Fig. 6 are provided to give a statistical overview of the data contained inside this file.

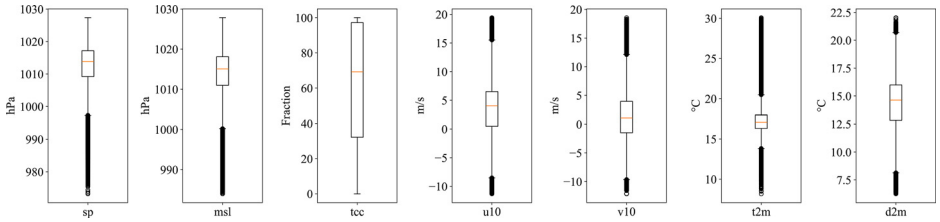

Fig. 6. Box plots of the 7 numerical features in the weather dataset.

Table 5
List of oceanographical (physical-based) features with units.

Column name	Attribute	Unit
`so`	Sea surface salinity	PSU
`thetao`	Sea surface temperature	°C
`uo`	Eastward surface velocity	m/s
`vo`	Northward surface velocity	m/s
`zos`	Sea surface height	m

3.15. Oceanography

In addition to weather data, and primarily for studying sea conditions, we have added oceanography files available from the E.U. Copernicus Marine Service Information.[6] On this platform, oceanographic data is divided into two categories: physical-based and wave-based.

3.16. Physical-based

These data, which basically contain the salinity and temperature of the water, but also the height and surface velocities, are generated from the E.U. Copernicus Marine Service Information [15]. Initially in GRIB format, the file was decompressed to obtain a Parquet file (`Dataset_CMEMS_PHY.parquet`) containing 5007,803 oceanographic observations. The spatial resolution is 0.083° (∼ 9 km) and the temporal resolution is hourly. In addition to the location (latitude and longitude), physical-based oceanographical attributes we have retained are listed in the Table 5.

3.17. Wave-based

These data, which basically contain the height, direction, and period of primary and secondary waves as well as swell waves, are generated from the E.U. Copernicus Marine Service Information [16]. Initially in GRIB format, the file was decompressed to obtain a Parquet file (`Dataset_CMEMS_WAV.parquet`) containing 1670,779 oceanographic observations. The spatial resolution is 0.083° (∼ 9 km) and the temporal resolution is 3-hourly. In addition to the location (latitude and longitude), the wave-based oceanographical attributes we retained are listed in the Table 6.

[6] https://marine.copernicus.eu.

Table 6
List of oceanographical (wave-based) features with units.

Column name	Attribute	Unit
VCMX	Maximum crest trough wave height	m
VHM0	Spectral significant wave height (Hm0)	m
VHM0_SW1	Spectral significant primary swell wave height	m
VHM0_SW2	Spectral significant secondary swell wave height	m
VHM0_WW	Spectral significant wind wave height	m
VMDR	Mean wave direction from	degree
VMDR_SW1	Mean primary swell wave direction from	degree
VMDR_SW2	Mean secondary swell wave direction from	degree
VMDR_WW	Mean wind wave direction from & degree	degree
VMXL	Height of the highest crest	m
VPED	Wave principal direction at spectral peak	degree
VSDX	Stokes drift U	m/s
VSDY	Stokes drift V	m/s
VTM01_SW1	Spectral moments (0,1) primary swell wave period	s
VTM01_SW2	Spectral moments (0,1) secondary swell wave period	s
VTM01_WW	Spectral moments (0,1) wind wave period	s
VTM02	Spectral moments (0,2) wave period	s
VTM10	Spectral moments (−1,0) wave period	s
VTPK	Wave period at spectral peak / peak period	s

3.18. Bathymetry

An additional layer of interest is the bathymetry of the area. Thus we added the data of Naval Hydrographic and Oceanographic Service from their HOMONIM project [17], which has a spatial resolution of 0.001° (∼ 111 m). We then created a Parquet file (`Dataset_BATHYMETRY.parquet`) containing the 12,386,244 data points extracted to only retain those in the area of interest.

4. Experimental Design, Materials and Methods

The design for the MMDEC dataset is centered on the aggregation of multi-source data, specifically integrating AIS streams and Sentinel-1 SAR satellite imagery over a continuous three-month period in 2023. This satellite dataset was prepared using the SatShipAI platform to automate the retrieval and processing of satellite products. The pre-processing of AIS data has been done in Python and consisted of geographical and temporal filtering to limit messages to the area and period of interest, while performing a spatial join with satellite image segments and preserving the initial position data unmodified.

4.1. Satellite dataset

The satellite dataset was generated through the SatShipAI platform, which automates key processing stages including Sentinel-1 product retrieval, preprocessing, AI-based inference, and post-processing, all according to the defined AOI and time period. Details on the model performance and platform implementation are provided in [4]. The imagery used originates from the Copernicus Sentinel-1 SAR constellation. Sentinel-1 GRD products were acquired from the Copernicus Open Access Hub for the English Channel AOI, covering the period July–September 2023. The data were acquired in Interferometric Wide (IW) swath mode, providing a swath width of approximately 250 km, and include both VV and VH polarization channels.

Satellite data preprocessing was carried out using the ESA Sentinel Application Platform (SNAP [18]) software. The workflow included the application of the “Apply Orbit File” operator

for precise orbit correction, followed by "Thermal Noise Removal" and "Radiometric Calibration" to sigma-nought backscatter values. These steps ensured geometric and radiometric consistency across all products.

After satellite preprocessing, each product was divided into 800 × 800 × 3 pixel tiles (image chunks), where the three channels correspond to VV, VH, and (VV + VH)/2. To generate the chunks from the full satellite image, an overlap of 50 pixels was applied. All images were scaled to the 0–255 range and converted to the `int8` data type. Only chunks located over sea areas within the defined AOI were retained. This procedure reduced the size of each Sentinel-1 product from approximately 2 GB to about 50 MB. During the model inference stage, zero-padding was applied when necessary to maintain the required image dimensions for the detection model, which was used to identify ship locations. For each detected ship, smaller 64 × 64 × 3 pixel image patches were processed by regression and classification models to estimate ship length and ship type, respectively.

During the post-processing phase, several exclusion criteria were applied to enhance the reliability of the detections. Ship positions were first cross-checked against land-mask polygons to remove detections located over land or within 500 m of the coastline. Additional filtering steps excluded detections outside the AOI boundaries and within predefined wind-farm zones, ensuring that only valid maritime detections were retained in the final dataset.

The three wind-farm exclusion areas were created using the Copernicus Browser.[7] Within the platform, the "Draw polygonal AOI for image downloads and timelapse" tool, available under "Create an area of interest" was used to manually delineate the polygons after setting the image date to fall within the 2023 observation period.

4.2. AIS messages

In the `Dataset_AIS_POS.parquet`, we have added two particularly relevant attributes that did not exist initially: `chunk_folder` and `id_chunk`. Actually, they constitute a spatial join between position messages and satellite image chunks whose coordinates are listed in the `Dataset_S1_chunks_corners.zip` archive. In other words, in these columns, it will be possible to find the names of the chunks (one or more due to the chunks overlapping mentioned above) in which the vessel reporting its AIS position is theoretically located, if such chunks exist.

Status messages do not contain geographical positions; therefore, simple geographical extraction was not possible. We therefore extracted, for each day, status messages for which the declared MMSI (and thus the associated vessel) sent at least one position in the area during the day. However, there is not a full correspondence between position and status messages across all ships because status messages are sent less frequently than position messages. That is why the number of unique MMSI in `Dataset_AIS_SPEC.parquet` is smaller than the number of unique MMSI in `Dataset_AIS_POS.parquet`.

Apart from the addition of `chunk_folder` and `id_chunk`, we decided to keep the AIS data as is to give users the opportunity to process the "raw" data. To go into a little more detail, we used only two Python libraries (PyArrow[8] and GeoPandas[9]) and the algorithm procedure is rather simple. Indeed, for each day in the period, we

- created an empty list L
- read all the positions messages (worldwide)
- extracted only the chosen features (columns)
- kept only the position messages located in `AOI.geojson`
- stored the unique MMSIs in L
- extended `Dataset_AIS_POS.parquet` read all the status messages (worldwide)

[7] https://browser.dataspace.copernicus.eu.
[8] https://arrow.apache.org/docs/python/index.html.
[9] https://geopandas.org/en/stable/.

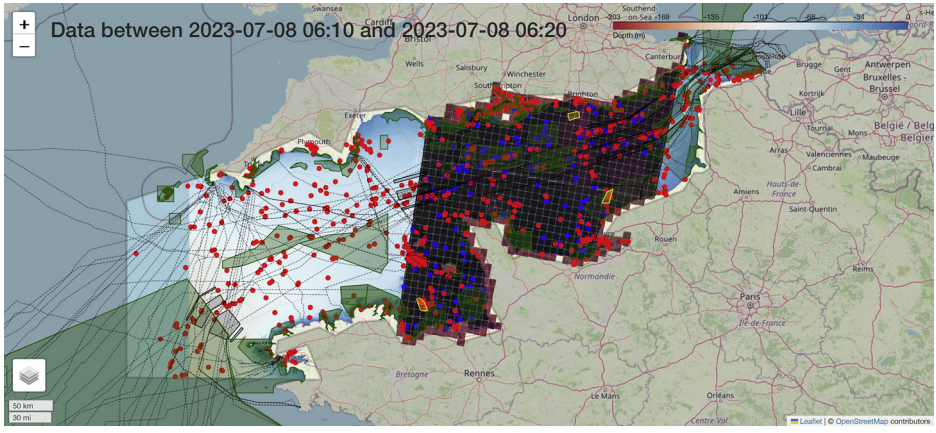

Fig. 7. Screenshot of the visualization map. The bathymetry has been placed in the background, four satellite products acquired between 2023 and 07–08 06:10 and 2023–07–08 06:20 have been added in transparency with the white grid to see the boundaries of the chunks, the AIS position messages are in red, the AI detections by the SatShipAI model are in blue, the wind farms are the yellow areas, the marine protected areas are depicted in green, the TSS are the black areas and the telecommunications and power cables are the grey dashed lines.

- extracted only the chosen features (columns)
- kept only the status messages emitted by a MMSI stored in L
- extended `Dataset_AIS_SPEC.parquet`

4.3. Visualization

In order to visualize all these data sources, an interactive map, created using the Folium Python library, utilizing a subset of all data sources of the MMDEC dataset between 2023–07–08 06:10 and 2023–07–08 06:20, has been included in a standalone HTML file (`VISUALIZATION.html`). Fig. 7 depicts a screenshot of this interactive map with some of the layers been displayed.

5. Example Use Case and Guidance for Data Reuse

As a simple illustrative use case, the dataset can be used to associate AIS vessel positions with the SAR-based ship detections derived from Sentinel-1 imagery. A typical workflow involves selecting a SAR satellite product and its corresponding image chunks, retrieving SatShipAI detections, and extracting AIS position messages within a temporal window around the satellite overpass, followed by spatial matching to link AIS-reported positions with radar-detected targets. This enables basic analyses such as short-term trajectory reconstruction around satellite acquisitions, identification of AIS gaps, or comparison between transmitted and observed vessel positions. An example of AIS-SAR matching is shown in Fig. 8, which presents two SAR image chunks with overlaid AI detections and temporally nearby AIS position reports selected within $a \pm 20$-minute window. In the left image, two vessels are detected (red bounding boxes), with three AIS positions for one vessel (green markers connected by a dashed local trajectory) and two for the other (yellow markers), while in the right image, located closer to the shoreline, a single detected vessel is associated with five AIS positions. The apparent displacement between AIS-reported positions and SAR detections illustrates known Doppler-related effects in SAR imagery, highlighting the need for flexible temporal and spatial matching criteria when linking AIS and SAR observations.

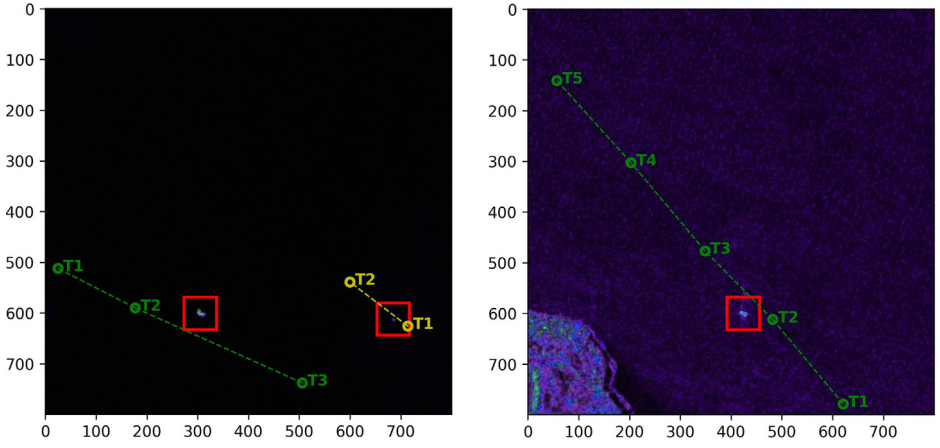

Fig. 8. Examples of AIS–SAR matching using Sentinel-1 SAR image chunks. Red rectangles indicate ship detections produced by the SatShipAI models. AIS position messages recorded within ±20 min of the SAR acquisition are overlaid as colored markers and connected by dashed lines to illustrate local trajectory segments. The apparent offset between AIS-reported positions and the SAR detection illustrates Doppler-related displacement effects commonly observed in SAR imagery.

This displacement is primarily governed by the component of vessel velocity directed toward or away from the satellite (i.e., along the radar line-of-sight). In practice, this component can be approximated by combining AIS-derived speed and course information with the satellite viewing geometry provided in the satellite dataset metadata (`extra_info.csv`), such as azimuth and incidence angle. These parameters enable users to estimate the expected azimuth offset when performing spatial matching between AIS and SAR observations. A detailed treatment of SAR motion compensation and displacement correction is beyond the scope of this data paper, however, the dataset contains the necessary information for users who wish to apply such corrections in their own analyses.

The matching of AIS positions with SAR detections is itself a processing task and is not enforced in the dataset. Users are expected to define their own matching criteria depending on the application. Based on exploratory analysis, most vessels in the dataset transmit AIS position messages every 9–16 min (approximately 65 %), with 11–13 min being the most frequent interval (approximately 39 %). Selecting AIS position messages within $a \pm 20$-minute window around the SAR acquisition time for instance for this dataset would typically yield between 1 and 5 AIS trajectory points per vessel. Spatial correspondence may then be evaluated using a distance threshold while a conservative value of no >3 km could serve as an initial reference.

For new users, a recommended entry point is the combined use of SAR image chunks, SatShipAI detection outputs, and AIS position messages, which together support trajectory reconstruction, trajectory enhancement, and trajectory inference tasks. In this context, trajectory reconstruction refers to assembling vessel paths directly from AIS observations, trajectory enhancement to improving or validating these paths using SAR detections or contextual data, and trajectory inference to estimating vessel behavior or movement when observations are sparse or missing, such as during AIS gaps or for non-transmitting vessels.

Limitations

Users should be aware of several limitations inherent to the dataset. AIS reports may contain temporal gaps or inaccuracies, and small temporal mismatches between AIS transmissions and Sentinel-1 SAR acquisition times may affect vessel correspondence. In addition, some ves-

sels in the AIS position message dataset that do not have any corresponding AIS specification messages. It represents 1172 vessels. This is due to the frequency with which both types of messages (position or specifications) are sent and our desire to keep the raw data intact. Thus, asynchrony necessarily arises. Due to the construction of the dataset, the reverse is not true, i.e. any vessel that has sent a specification message has at least one associated position. Although the information contained in `Dataset_AIS_SPEC` is not necessarily required for the study of simple trajectories, it is true that this information can be useful if one wishes to semantically enrich the trajectories in order to offer an in-depth study. For this purpose, information on the length, ship name, and flag could be extracted from websites such as MarineTraffic[10] or Marine Vessel Traffic[11] in order to complete this missing data. Furthermore, from a purely practical point of view, using the `merge_asof` function from the Pandas library is particularly useful for enriching each position message with information about the ship extracted from the closest specification message in time.

Moreover, the SAR-based ship detections, classifications, and length estimates provided by the SatShipAI model are model-derived outputs rather than ground truth and may include false positives or false negatives, particularly for small vessels, closely spaced targets, or challenging sea states. In addition, differences in spatial resolution across data layers (e.g., Sentinel-1 SAR, AIS point data, and auxiliary weather data at ∼28 km resolution) should be considered when performing multi-source analyses. Finally, the dataset covers a three-month period (July–September 2023), which limits its suitability for long-term or seasonal trend studies.

Ethics Statement

The authors have read and follow the ethical requirements for publication in Data in Brief and confirming that the current work does not involve human subjects, animal experiments, or any data collected from social media platforms.

While this study does not involve human subjects, we recognize the importance of ethical considerations regarding the use of maritime data. The Automatic Identification System (AIS) is primarily designed for maritime safety and collision avoidance, and its data are mandated by international maritime regulations to be publicly broadcast for safety purposes. The MMSI numbers included in our dataset are publicly available identifiers that do not constitute personal data, as they only refer to vessels. These data are openly accessible through various institutional and commercial providers and are widely used for maritime traffic analysis, coastal management, and scientific research. Our use of this data is consistent with established practices in the maritime research community. The data processing and analysis were conducted solely for scientific research purposes, and no sensitive operational information that could compromise vessel security or commercial interests has been disclosed.

CRediT Author Statement

- **Tristan Averty**: Conceptualization, Software, Data Curation, Writing – Original Draft, Visualization
- **Ioannis Nasios**: Software, Data Curation, Writing – Original Draft, Visualization
- **Cyril Ray**: Conceptualization, Writing – Original Draft, Supervision
- **Nikos Piliouras**: Writing – Review & Editing, Supervision

[10] https://www.marinetraffic.com/.
[11] https://www.marinevesseltraffic.com.

Data Availability

MMDEC: Multimodal Maritime Dataset on the English Channel (Original data) (Zenodo).

Acknowledgements

This work was supported by the MUSIT Project through the European Union's Horizon Europe Framework Programme (HORIZON), under Marie Sklodowska-Curie grant agreement no. 101182585. The work only reflects the authors' views; the EU Agency is not responsible for any use of the information it contains.

Declaration of Competing Interest

The authors declare that has no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

[1] C. Ray, A. Troupiotis-Kapeliaris, I. Kontopoulos, V. Andronikou, I. Nasios, N. Piliouras, T. Chevallier, V. Delmas, K. Tserpes, D. Zissis, C. Renso, E. Carlini, Multi-sensor inferred trajectories (MUSIT) for vessel mobility, Oceans 2025 Brest (2025).
[2] Flanders Marine Institute, "IHO sea areas, version 3," 2018. [Online]. Available: https://doi.org/10.14284/323.
[3] SatShipAI, "Ship detection, type classification, and length estimation with artificial intelligence over SAR imagery," 2018. [Online]. Available: https://satshipai.eu/.
[4] I. Nasios, K. Vogklis, AI at sea, year six: performance evaluation, failures, and insights from the operational meta–analysis of SatShipAI, a sensor-fused maritime surveillance platform, Electronics 14 (18) (2025) 3648.
[5] Navigation Center, United States Coast Guard, U.S. Department of Homeland Security, "AIS messages," 2026. [Online]. Available: https://www.navcen.uscg.gov/ais-messages.
[6] U.S. Coast Guard, NOAA, BOEM, "AIS vessel type and group codes used by the marine Cadastre project," 2018. [Online]. Available: https://coast.noaa.gov/data/marinecadastre/ais/VesselTypeCodes2018.pdf.
[7] C. Merrien, "Worldwide list of seaports," 2021. [Online]. Available: https://doi.org/10.12770/59ab5f6f-79ea-425d-830e-be5ecdb7bdbe.
[8] Ministère de la Transition écologique, "Opérations coordonnées par les CROSS," 2026. [Online]. Available: https://www.data.gouv.fr/datasets/operations-coordonnees-par-les-cross/.
[9] SHOM, "Dispositifs de séparation du trafic," 2021. [Online]. Available: https://www.data.gouv.fr/datasets/dispositifs-de-separation-du-trafic-1.
[10] Centro Tecnológico del Mar - Fundación CETMAR, "EMODnet human activities, energy, wind farms," 2025. [Online]. Available: https://emodnet.ec.europa.eu/geonetwork/srv/api/records/8201070b-4b0b-4d54-8910-abcea5dce57f.
[11] UNEP-WCMC and IUCN, "Protected Planet: the World Database on Protected Areas (WDPA)," 2025. [Online]. Available: www.protectedplanet.net.
[12] Cogea, "EMODnet human activities, cables, power, actual routes," 2023. [Online]. Available: https://emodnet.ec.europa.eu/geonetwork/srv/api/records/41b339f8-b29c-4550-b787-3d68f08fdbcc.
[13] Cogea, "EMODnet human activities, cables, telecommunication, actual routes," 2023. [Online]. Available: https://emodnet.ec.europa.eu/geonetwork/srv/api/records/39ebe289-410b-4a5d-88a4-51bfcde538de.
[14] Copernicus Climate Change Service, Climate Data Store, "ERA5 hourly data on single levels from 1940 to present," 2023. [Online]. Available: https://doi.org/10.24381/cds.adbb2d47.
[15] E.U. Copernicus Marine Service Information, "Global ocean physics analysis and forecast," 2024. [Online]. Available: https://doi.org/10.48670/moi-00016.
[16] E.U. Copernicus Marine Service Information, "Global ocean waves analysis and forecast," 2023. [Online]. Available: https://doi.org/10.48670/moi-00017.
[17] SHOM, "MNT bathymétrique de façade Atlantique (Projet HOMONIM)," 2015. [Online]. Available: 10.17183/MNT_ATL100m_HOMONIM_WGS84.
[18] European Space Agency, "SNAP - {ESA} sentinel application platform v5.0.0," 2016. [Online]. Available: http://step.esa.int.