

MMSI Based Anomaly Classification in AIS: A Rules-First Baseline from International Numbering Standards

*Dhya Aqilla Pramudhita*¹, *Adi Novitarini Putri*¹, and *Totok Yulianto*^{1*}

¹Department of Naval Architecture, Institut Teknologi Sepuluh Nopember, Surabaya, Indonesia

Abstract. Research on Automatic Identification System (AIS) anomalies has largely focused on vessel trajectories and kinematics, while identifier validity is often assumed. This study fills that gap by using the nine-digit Maritime Mobile Service Identity (MMSI) to build a rules-first baseline for anomaly classification. Validation rules are derived from ITU-R M.585-9 and operational guidance (USCG NA VCEN, AMSA), covering format constraints, category and prefix patterns, and MID ranges. The same rules are applied to two public data sources (Global Fishing Watch and NOAA/Access AIS). The pipeline assigns per-record labels for validity, category, and diagnostic notes, and defines a taxonomy of identity anomalies: invalid format, misclassification or misuse, MID and policy inconsistencies, and spatiotemporal “cloned MMSI” detected via overlap tests when positions are available. Results indicate that identity screening reduces noise, highlights priority cases, and produces cleaner inputs for downstream behavioral models without relying on speed or trajectories. Contributions include a reproducible MMSI rule set, an anomaly taxonomy, and a per-source evaluation protocol to avoid misleading generalizations. The approach is transparent, computationally efficient, and easy to integrate as a first-stage filter in maritime analytics pipelines.

1 Introduction

The Automatic Identification System (AIS), which allows authorities to track vessel movements in almost real time, has emerged as a key component of modern maritime safety and monitoring. AIS data now serves a variety of purposes beyond its initial safety-of-life goal, such as environmental protection, fisheries management, and maritime security analytics. However, all of these applications implicitly rely on the dependability and consistency of the vessel identities encoded in Maritime Mobile Service Identities (MMSI) across platforms and jurisdictions. By analyzing the MMSI numbering standard and suggesting rules-first validation techniques that can be methodically applied to extensive AIS data, this paper challenges that assumption.

* Corresponding author: tyulianto1970@its.ac.id

1.1 Background

AIS is increasingly central to maritime safety, law enforcement at sea, and large-scale maritime analytics. Most mainstream AIS anomaly studies concentrate on trajectory and kinematics-based cues such as route deviations, speed anomalies, or loitering. In contrast, the validity of identifiers, especially the nine-digit MMSI, including its structure and consistency, is usually presumed rather than screened upfront. Without early checks, identity inconsistencies can cascade through later modeling stages; therefore, a transparent, lightweight, and reproducible identity-based screen should be the first step. [8]

The ITU-R M.585-9 standard formalizes the MMSI structure, including the definition of Maritime Identification Digits (MID) and numbering schemes for various categories, such as ship and shore stations, Aids to Navigation (AtoN), handheld DSC, and specialized devices such as AIS SART, MOB/MSLD, and EPIRB-AIS. This recommendation also regulates the assignment, use, and reuse of identities at the administrative level, including provisions on numbering consistency and management when devices or vessels change ownership or flag states. With this foundation, anomalies can be formulated deterministically, such as invalid digit patterns, inappropriate category prefixes, or MID inconsistencies, as strong and easily auditable validation rules [1].

At the operational level, the USCG NAVCEN underscores the imperative of timely MMSI registration to enable effective SAR operations and ensure regulatory compliance while providing a synoptic overview of the MMSI structure and applicable MID ranges. Correspondingly, the AMSA delineates the requirements and procedural pathways for MMSI allocation to AIS/DSC and AtoN units, emphasizing the centrality of MMSI to maritime safety and accountable identity management [2], [4].

On the data side, MMSI-based identities are documented and curated by two popular public sources, Global Fishing Watch (GFW) [5] and NOAA (U.S. National Oceanic and Atmospheric Administration) [6]. Variations in their curation practices are crucial for reliable cross-source evaluation and result interpretation. After filtering and harmonizing MMSI fields and vessel attributes, GFW releases a global static dataset of apparent fishing effort based on AIS [6]. On the other hand, downstream users bear the majority of the burden of MMSI validation since NOAA's Marine Cadastre AIS archive exposes raw message fields with little pre-screening [7]. To put GFW and NOAA records on an equal footing before analyses or cross-validation exercises, a clear rules-first approach is therefore required in this work.

1.2 Objectives

The objectives of this study are:

1. Formulated and implemented replicable MMSI validation rules derived from international standards (ITU-R M.585-9) and operational guidelines (USCG NAVCEN, AMSA) to establish a rule-based anomaly classification baseline for AIS data.
2. Identity-based anomaly indicators are defined as follows: (i) format anomalies, (ii) misclassification/misuse, (iii) "cloned MMSI" or cross-spatiotemporal identity collisions, and (iv) policy anomalies.
3. Evaluate the baseline on two data subsets (GFW and NOAA) to quantify detection rates, assess ease of replication, and gauge utility as a preprocessing step prior to trajectory- or kinematics-based modelling.

1.3 Problem Limitations

The problem limitations of this study are:

1. Rules and labels refer to ITU-R M.585-9 (format and categories) and NAVCEN/AMSA guidelines; variations in national policies outside these references are not explored in depth.
2. The case study data sources are subsets of GFW and NOAA, and the findings depend on the curation or definition policies of each provider (e.g., likely_gear, spoofing); therefore, generalization to other AIS providers needs to be tested.
3. Cloned or identity collision detection is treated as an identity-based indicator (requiring temporal overlap testing for the same MMSI in distant locations) and does not yet integrate cross-sensor fusion (e.g., VMS/VIIRS/SAR). Methodological reinforcement may follow identity verification practices in operational documentation.

2 Theoretical Basis

2.1 International Telecommunication Union (ITU)

The ITU is a specialized agency of the United Nations that oversees the radiocommunication sector (ITU-R) with a mandate to ensure the rational, equitable, efficient, and economical use of the radio frequency spectrum by all radiocommunication services. ITU-R technical recommendations are developed through studies in working groups and adopted by the Radiocommunication Assembly. In the context of maritime identity, ITU-R M.585-9 provides the scope, key terms, and appendix structure that describe the identity format and management guidelines.

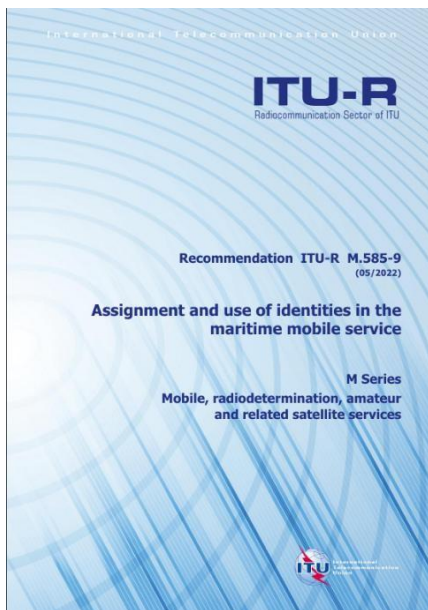


Fig. 2.1 Cover of ITU-R document M.585-9: assignment and use of identities in the maritime mobile service identity (MMSI).

In essence, M.585-9 specifies the method of assignment, use, and conservation of identities in the maritime mobile service (MMS), covering formats for ship/coast/group, AIS

Aids to Navigation (AtoN), crafts associated with a parent ship, SAR aircraft, and special AIS SART/MOB/EPIRB-AIS devices. This document also emphasizes the availability of identity data for authorized entities such as Rescue Coordination Centers (RCC) and publication in the MARS database, while providing guidelines in Annex 3 for administration: assignment procedures, updates to the Radiocommunication Bureau, and reuse criteria (e.g., after two consecutive editions of List V or ≥ 2 years of absence) so that MID/MMSI resources are not depleted [1].

2.2 Australian Maritime Safety Authority (AMSA)

AMSA is Australia's national maritime safety authority that regulates accident prevention, SAR preparedness, and maritime radio equipment compliance. In the context of identity, AMSA manages the licensing or registration of devices that transmit identities on DSC/AIS, including the MMSI application process and checking the conformity of owner, vessel, and device category data. AMSA also explains the application of the MID block "503" (Australia) to various classes, providing readers with concrete examples of how ITU standards are translated into administrative practices in specific countries, from ship/coast/group numbering to specialized devices. The policy on the AMSA website helps researchers distinguish valid identities from misclassified cases, such as handheld devices being treated as vessels, and understand the relevant validation points for MMSI-based screening [4]. From a practical point of view, using AMSA's guidance in the rules-first baseline keeps the screening criteria close to real regulatory practice rather than arbitrary assumptions. Linking the ITU recommendations to how a national authority actually assigns and uses MMSI codes also makes the validation rules easier to interpret and more likely to be reused beyond the Australian context.

2.3 U.S. Coast Guard Navigation Center (NAVCEN)

NAVCEN is the official portal of the U.S. Coast Guard for maritime navigation information and radio communication identity services. In the context of MMSI, NAVCEN summarizes identity formats according to international recommendations, including class mapping such as ship/coast/group, SAR aircraft, handheld DSC, and AtoN, and links to national registration policies (e.g., through the FCC in the United States). NAVCEN also provides a range of valid MIDs (201–775) and explanatory materials to help vessel owners, radio operators, and authorities understand the relationship between MMSI and DSC/AIS and Search and Rescue (SAR) operations [2]. For data researchers/practitioners, the NAVCEN page is very useful for compiling summary tables of MMSI patterns and validation checklists prior to behavioral analysis. Examples of patterns that are often used as references include: coast 00MIDXXXX, group 0MIDXXXX, SAR aircraft 111MIDXXX, handheld 8MIDXXXXXX, and AtoN pattern 99MIDXXXX. This material is practical to use as the basis for identity-based rules (format check, category check, MID consistency) in the pre-processing stage, so that administrative/format anomalies can be filtered out early before the kinematic model is applied [3].

2.4 MMSI Structure (9 digits) & Identity Class

The MMSI is a 9-digit numerical identity used by stations in maritime mobile services, such as ships, coastal stations, AtoN, SAR aircraft, and safety devices for identification purposes on AIS/DSC and integration with maritime safety numbering schemes. The first three digits are Maritime Identification Digits (MID) that represent the administration or authority that assigns the identity; the remaining six digits follow a pattern according to the station or device

class. MMSIs are issued by national administrations or recognized authorities and are intended to be globally unique; they should be updated or retired when a vessel’s flag, ownership, or equipment configuration changes. MID values indicate the assigning administration rather than nationality and are distinct from ISO country codes. Moreover, the leading digit constrains the station class (e.g., 0/00/99/8/9 prefixes for special use). Basic automated screening verifying nine-digit length, numeric content, permissible prefixes, and MID validity against allocation tables can eliminate many identity errors prior to behavioral analysis [1].

The official MMSI format details are specified by ITU-R M.585-9, including the following:

1. Ship station: MID + 6 digits (the first digit of the MMSI is in the range 2–7)
2. Group ship call: 0 + MID + XXXX
3. Coast/base station: 00 + MID + XXXX
4. Handheld VHF DSC: 8 + MID + XXXXX (handheld DSC/GNSS device)
5. Prefix "9" (special devices), including
6. AIS-SART 970..., MOB/MSLD 972..., EPIRB-AIS 974...
7. AIS AtoN 99 + MID + XXXX (with examples of grouping: 99MID1XXX physical, 99MID6XXX virtual, 99MID8XXX mobile)
8. Craft associated with a parent ship 918 + MID + XXXX

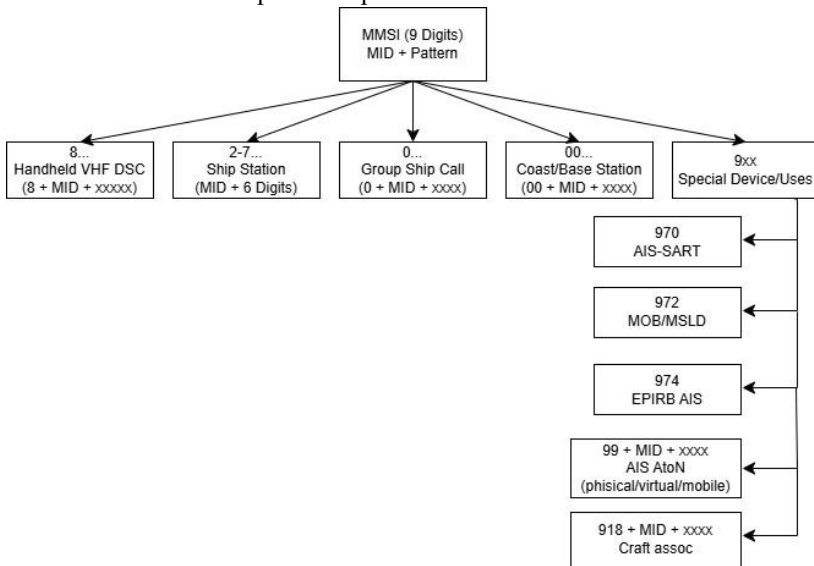


Fig. 2.2 MMSI Prefix Tree and Class Mapping (Derived from ITU R M.585-9)

Implications for validation or anomalies:

Because the numerical pattern of each class is deterministic (e.g., 0/00 = group/coast, 2–7 = ship, 8 = handheld, 9xx = special device), validation rules can capture (i) invalid format (non-digit or ≠9 digits), (ii) mis-categorization/misuse (e.g., prefix 8 but behaves like a commercial vessel), (iii) MID mismatch with flag/area of operation, and (iv) indication of identity collision/cloned MMSI (same identity appears simultaneously in distant locations; requires time overlap test). The management and reuse guidelines in Annex 3 (e.g., reuse after absence from two consecutive editions of List V or ≥2 years) can be used as a basis for policy anomalies[1].

3 Research Methodology

3.1 Research Design

This study adopted a computational survey method with a rules-first validation approach to MMSI identities in AIS data. The rules are derived from the MMSI numbering standard (9-digit format, class/prefix, and MID range 201–775) and then implemented as a validation script that generates validity labels, MMSI categories, and reasons or notations.

For the two data sources (NOAA and GFW), the design includes two validation paths that are consistent yet flexible with regard to column schemes.

1. GFW Path: utilizes `validate_mmsi_indonesia_GFW.py`, which contains the essential methods `analyze_csv()`, which exports to Excel with automatic flagging of invalid rows, and `validate_mmsi()`, which returns `is_valid`, `category`, and `note`.
2. NOAA Path: uses a generic validator package (validation module) via the `validate_mmsi()` function (`boolean + reason`), `validate_rows()`, and `sort_mmsi_records()` to generate separate CSV files (`valid/invalid`) and a `.txt` summary. These functions are exposed via `__init__.py` and are also called by the CLI `__main__.py`.

The core cross-source process includes:

1. MMSI preprocessing (trim, numeric check, and 9-digit length);
2. Prefix pattern-based validation (`ship/group/coast/handheld/9xx`);
3. MID policy checks (201–775).
4. Category labeling + reason notation and export (Excel for GFW; CSV+summary for NOAA). The rule implementation, including the `MID_MIN/MID_MAX` limits, is uploaded in the module.

3.2 Research Time and Place

The computational experiment was conducted offline in a data processing environment (Python + pandas). Implementation period October 2025. The AIS datasets used in this study were obtained from the following sources:

1. GFW: previously downloaded `validate_mmsi_indonesia_GFW.py`. AIS subset; processed with `validate_mmsi_indonesia_GFW.py`.
2. NOAA: equivalent NOAA AIS archive (CSV format); processed with the validation module and `mmsi_validator` CLI.

3.3 Data Collection Methods

The data collection methods for this study are:

1. Primary data (computational results):
 - GFW: Excel output containing the columns `mmsi_valid`, `mmsi_category`, `mmsi_note`; invalid rows are automatically highlighted by `analyze_csv()` (`openpyxl` `PatternFill`).
 - NOAA: two sorted CSV files (`*_valid.csv`, `*_invalid.csv`) plus a `summary*_mmsi_summary.txt` from `sort_mmsi_records()`.
2. (USCG) summary, and AMSA used as the basis for formulating validation rules (without changing the AIS data content).
3. AIS data source and execution:
 - GFWinput: AIS CSV (e.g., `sar_vessel_detections_pipev3_20250922.csv`), called with the `--input` argument and exported to Excel (`--output`).

- Input NOAA: CSV AIS NOAA (mis. AIS_YYYY_MM_DD.csv); via CLI `mmsi_validator --input <file> [--output-dir <folder>]` which prints the summary to stdout.

3.4 Required Data

The data required for this study is:

1. Key attributes: The minimum required input is the MMSI column, represented as a 9-digit string. During pre-processing, values are cleaned (trimming spaces and removing non-digit characters) and rejected if the length $\neq 9$ or if any non-digits remain. In this study, MMSI values were extracted from two AIS sources, Global Fishing Watch (GFW) AIS-based apparent fishing effort v3 and the NOAA Office for Coastal Management (MarineCadastre) AIS archive.
2. Validation parameters: Rules using parameters: MID_MIN = 201, MID_MAX = 775, dan VALID_FIRST_DIGITS = {2,3,4,5,6,7} for ship stations. The range and structure follow the international MMSI standard [1].
3. Category mapping (MMSI pattern):
 - Ship station: MID + 6 digit (first digit of MMSI $\in \{2-7\}$).
 - Group ship call: 0 + MID + XXXX.
 - Coast/Base station: 00 + MID + XXXX.
 - Handheld VHF DSC: 8 + MID + XXXXX.
 - Special devices/uses (prefix "9"): AIS-SART 970..., MOB/MSLD 972..., EPIRB-AIS 974..., AIS AtoN 99 + MID + XXXX (generally mapped physical/virtual/mobile), dan craft associated 918 + MID + XXXX.All of the above patterns are implemented in the `validate_mmsi()` function as a rule set so that each mmsi value can be labeled with a category and reason.
4. Input/output files:
 - GFW (Excel): `..._mmsi_validation_GFW.xlsx` berisi `mmsi_valid`, `mmsi_category`, `mmsi_note` + highlight invalid.
 - NOAA (CSV + txt): `_valid.csv`, `_invalid.csv`, dan `_mmsi_summary.txt` (total, valid, invalid, percentage valid).

3.5 Data Analysis

The data analysis for this study are:

1. Validation and labeling (the `validate_mmsi()` method).

We process each mmsi in the following ways: (1) check the format (9 digits, all numbers); (2) check the category pattern based on the prefix (ship/group/coast/handheld/9xx); (3) check the MID policy within the valid range; and (4) arrange the output:

 - `mmsi_valid` $\in \{\text{True}, \text{False}\}$;
 - `mmsi_category` $\in \{\text{ship_station}, \text{group}, \text{coast}, \text{handheld}, \text{ais_sart}, \text{mob_msld}, \text{epirb_ais}, \text{aton}, \text{craft}\}$;
 - `mmsi_note` (brief reason, e.g.: "length not 9", "first digit not {2-7}", "prefix 9 does not match 970/972/974/98/99+MID", "MID 503", etc.).
2. Heuristic for detecting identity anomalies.

From the three output columns above, we get four indicators:

 - Length $\neq 9$ or non-digit.
 - Misclassification/misuse, e.g., prefix 8 (handheld) used consistently like an operational vessel.

- Policy/MID outside 201–775 or inconsistent with administration/area of operation.
 - Potential identity collision due to MMSI cloning, indicated by concurrent occurrences of a single MMSI in far apart areas (needs time-overlap testing of position data, prepared as a subsequent stage).
3. Summary in numbers.
To describe data quality and the rule-based screening framework, the findings are summarized as follows:
- Validity distribution: the percentage of records that are marked as valid compared to those that are not;
 - Category distribution: the breakdown of `mmsi_category` classes to show how items are used and stored.
 - The most common `mmsi_note` strings that point up systematic problems (such as prefix violations that happen again and again or MIDs that are out of range).
- A concise summary table is provided to streamline the presentation of findings in the Results and Discussion section and to enable quicker, more transparent interpretation.
4. Export and look over: The full output is sent to Excel, where faulty rows are automatically highlighted to make manual auditing faster. In the Results & Discussion section, the summary table is used. All exports have consistent column definitions and provenance metadata (timestamp, rule version, and dataset source) to make sure they can be traced and reproduced.

3.6 Flowchart

The flowchart summarizes the rule-based MMSI validation and identity-anomaly screening workflow used in this study. The process starts with a focused literature review (ITU-R M.585-9, NAVCEN, AMSA) to ground the problem statement and the proposed solution. Three streams then proceed in parallel: (i) specifying anomaly-detection requirements, (ii) implementing the script (`validate_mmsi()` and related routines), and (iii) integrating domain rules (format, prefix/class, MID policy). These streams converge into a concise theoretical basis that formalizes the rule set.

A decision point is used to ensure that the rule set performs as intended. If the rules do not meet the expected criteria, the workflow iteratively applies targeted tests and quantitative evaluations to refine thresholds, pattern definitions, and diagnostic checks, followed by re-validation. This loop continues until the rules satisfy the required performance standards, at which point the validation is finalized. While the pipeline is executed on two AIS data streams (GFW and NOAA), it relies on a single reusable core of MMSI rules and generates consistent outputs (validity, category, and note). To ensure reproducibility, dataset provenance (source, access date, and input scope) is documented so that the entire screening process can be repeated and independently verified.

The flowchart aims to make the MMSI validation pipeline transparent and repeatable for other users, in addition to summarizing the steps. Each box makes clear what data is needed, what presumptions are being used, and how dataset-specific decisions fit into the workflow. This facilitates tracking the classification of specific identities, documenting any deviations, and transferring the same rule set to new AIS archives. Practically speaking, other maritime data projects can use the diagram as a checklist to implement and audit MMSI-based screening.

Lastly, the visual structure aids in distinguishing between dataset-specific and generic and reusable elements. The pipeline's fixed components (the MMSI rules) and adaptable components (data sources, thresholds, or local conventions) are easily visible to readers.

Because of this, the flowchart is a helpful tool for expanding the approach or contrasting it with other MMSI validation techniques.

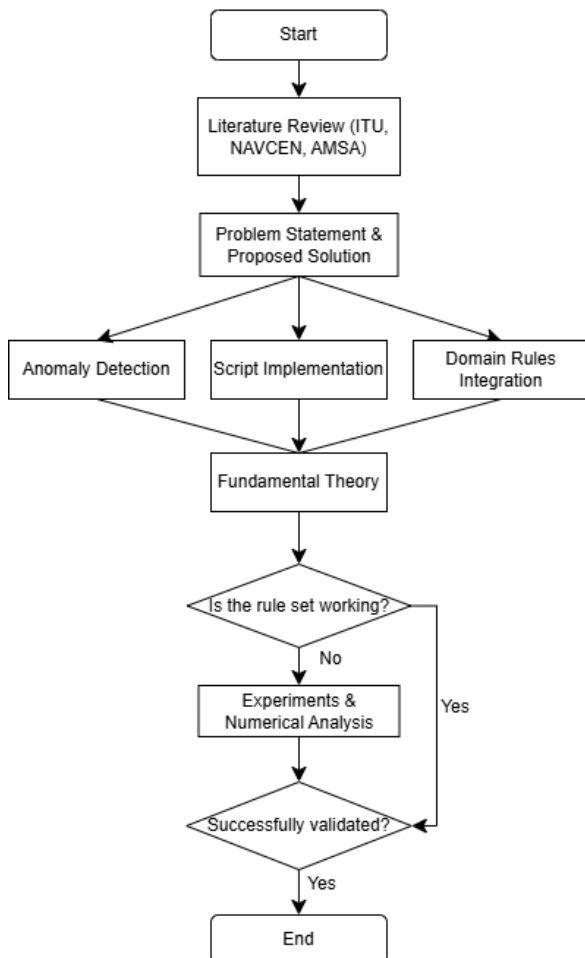


Fig. 3.1 Flowchart of the Rule-Based MMSI Validation and Identity-Anomaly Screening Workflow

4 Results and Discussion

4.1 Data Analysis Dataset Sources

The datasets used in result and discussion were obtained from two sources: Sentinel-1 SAR-based vessel detections from the Global Fishing Watch (GFW) Vessel Detections product and public AIS messages from the USCG Nationwide AIS (NAIS) redistributed via MarineCadastre NOAA AccessAIS. The GFW product supplies georeferenced detections (polygons/centroids) with confidence attributes derived from standardized processing of Sentinel-1 GRD scenes, while AccessAIS provides time-stamped AIS reports aggregated from terrestrial receivers across U.S. waters. For analysis, detection time, location, detection score (where available), and scene metadata (satellite, orbit, pass) were retained from GFW; the minimum required attribute from AccessAIS was MMSI, with additional fields (BaseDateTime, LAT/LON, SOG, COG, VesselType, CallSign, IMO) preserved for

descriptive statistics. Both sources were accessed in October 2025, harmonized to UTC, and filtered to overlapping time windows; MMSI values were trimmed, non-digits removed, and 9-digit length enforced prior to identity screening [6], [7].

The MMSI validation pipeline was implemented in Python. The source code for the GFW AIS data is available in the MMSI Based GFW repository, and the code for the NOAA AIS data is available in the MMSI Based NOAA repository [9], [10].

4.2 Validity Summary

Rule-based filtering of the MMSI column produces two main labels: valid and invalid. The invalid category is mainly caused by (i) length \neq 9 digits, (ii) non-digit characters, and (iii) prefix pattern mismatches (e.g., prefix 9 but not 970/972/974 or 98/99+MID). The complete distribution is shown in Table 4.1 for GFW and Table 4.2 for NOAA. The proportion of invalid entries confirms the high level of identity noise in field AIS and reinforces the importance of identity based pre filtering prior to trajectory analysis.

Table 4.1 Distribution of MMSI Validity in the GFW Dataset.

No	Status	Number (n)	Percentage (%)
1	True	156067	76.58
2	False	47741	23.42

Table 4.2 Distribution of MMSI Validity in the NOAA Dataset.

No	Status	Number (n)	Percentage (%)
1	True	8143389	99.84
2	False	13120	0.16

The MMSI validity screening results in Tables 4.1 and 4.2 reveal clear differences in identifier quality across the two AIS sources. For the GFW dataset (Table 4.1), 156,067 records (76.58%) conform to the rule base and are labeled as valid MMSI, while 47,741 records (23.42%) are flagged as invalid, indicating that a substantial portion of entries do not satisfy the enforced constraints, such as not being exactly 9 digits after cleaning, containing residual non digit characters, or failing expected prefix and MID patterns. This finding supports the need for a rules first quality gate, as nearly one in four records could otherwise propagate unreliable identity information into downstream analyses such as trajectory based anomaly detection or AIS SAR fusion. In contrast, the NOAA dataset (Table 4.2) shows near complete conformity to the same rules, with 8,143,389 records (99.84%) classified as valid and only 13,120 records (0.16%) invalid, suggesting a more consistent encoding of MMSI in NOAA while still reflecting a small number of corrupted or out of standard entries typical of real world archives. Taken together, the tables demonstrate that the proposed screening rules can deterministically separate conforming versus non conforming identifiers and that the extent of early stage identity cleaning is strongly source dependent, with GFW requiring substantially more filtering than NOAA before reliable identity driven analytics can be performed.

4.3 MMSI Category Distribution

Mapping of MMSI patterns shows the dominance of ship_station, consistent with the character of the AIS subset driven by ship activity. Note that ‘X’ is used as a wildcard for numeric digits (0–9) in the pattern notation (e.g., MID+XXXX), not as a literal character in MMSI values. The appearance of handheld (8...) and 9xx devices (e.g., AIS-SART/EPIRB-AIS) signals caution regarding mis-categorization/misuse (e.g., handhelds behaving like operational ships). Group/coast entries help identify non-vessel station records that sometimes enter the detection pipeline. Numerical summaries are shown in Table (see Table 4.3 for GFW and Table 4.4 for NOAA category derivatives from validation reasons).

Table 4.3 MMSI Category Distribution of GFW Dataset

No	Category	Number (n)	Percentage (%)
1	ship_station	156262	76.67
2	Unknown	45672	22.41
3	free_form	785	0.39
4	handheld_vhf_dsc	719	0.35
5	auxiliary_craft	358	0.18
6	ais_sart	8	0.0
7	epirb-ais	2	0.0
8	mob_msid	2	0.0

Table 4.4 Category Distribution of NOAA Dataset.

No	Category	Number (n)	Percentage (%)
1	unknown	10690	81.48
2	ship_station	2059	15.69
3	auxiliary_craft	371	2.83

Table 4.3 and Table 4.4 provide a breakdown of MMSI categories derived from the rule based pattern mapping, which helps distinguish vessel station identities from non vessel or special device identities that may appear in AIS streams. In the GFW dataset (Table 4.3), the distribution is dominated by ship_station with 156,262 records (76.67%), indicating that the subset largely represents conventional vessel transmissions, consistent with a ship activity oriented AIS sample. The second largest class is unknown with 45,672 records (22.41%), which suggests that a substantial portion of MMSI entries do not match the expected structural patterns used for category assignment, and therefore require caution in downstream analyses because identity type cannot be reliably inferred from MMSI alone. Smaller categories appear at low frequencies, including free_form (785; 0.39%), handheld_vhf_dsc (719; 0.35%), and auxiliary_craft (358; 0.18%), while special emergency and safety related devices such as AIS SART (8), EPIRB AIS (2), and MOB MSLD (2) occur only rarely (each

≈0.0% by proportion). Although these minor classes are numerically small, their presence is operationally important because handheld and 9xx device types can be misused or misconfigured and may produce tracks that resemble vessels if not filtered or handled separately.

In contrast, the NOAA dataset (Table 4.4) shows a markedly different profile, with unknown as the dominant category at 10,690 records (81.48%), followed by ship_station at 2,059 records (15.69%) and auxiliary_craft at 371 records (2.83%). This shift implies that, for the specific NOAA subset processed in this study, many MMSI values either fall outside the implemented category patterns or reflect data characteristics that make pattern based classification less straightforward, reinforcing the need to document dataset scope and pre processing choices when interpreting results. Taken together, the two tables show that MMSI category composition is strongly source dependent: GFW is primarily vessel station traffic, while NOAA contains a much higher proportion of identifiers that are not cleanly classifiable under the applied MMSI pattern rules. This insight is important for downstream pipelines, because category aware filtering can reduce false alarms by excluding non vessel stations and handling special devices separately before trajectory analysis or AIS SAR fusion.

4.4 Dominant Reasons for Invalid MMSI

The collection of reasons in the mmsi_note/mmsi_validation_reason column facilitates mass auditing and data cleansing prioritization. The most common reasons include: length not 9, non-digit, prefix 9 not compliant (not 970/972/974/98/99+MID), and MID outside 201–775. A summary of the top reasons is presented in Table (see Table 4.5 for GFW and Table 4.6 for NOAA).

Table 4.5 Dominant Reasons for Invalid MMSI in the GFW Dataset

No	Category	Number (n)	Percentage (%)
1	missing MMSI	41926	35.17
2	MID 412	23996	20.13
3	MID 413	9478	7.95
4	MID 525	8813	7.39
5	MID 636	8583	7.2
6	MID 538	6116	5.13
7	MID 352	4610	3.87
8	MID 574	3993	3.35
9	MID 431	3425	2.87
10	MID 257	3133	2.63
11	MID 563	2706	2.27
12	MID 477	2418	2.03

Table 4.6 Distribution of MMSI Validity in the NOAA Dataset.

No	Category	Number (n)	Percentage (%)
1	length not 9	8160	62.20
2	MID out of range	2240	17.07
3	invalid starting digit	2059	15.69
4	auxiliary MID out of range	371	2.83
5	invalid prefix 9 pattern	290	2.21

4.5 Dominant Reasons for Invalid MMSI

The foregoing results indicate that MMSI based screening functions effectively as a rules-first baseline, it attenuates noise, prioritizes invalid or suspicious records, and yields a cleaner corpus for downstream behavioral modeling. Detection of cloned identities can be incorporated by applying a spatiotemporal overlap test to records sharing the same MMSI, this indicator augments identity controls while imposing minimal computational overhead in early processing stages.

5 Results and Discussion

5.1 Summary of Findings

The analysis results demonstrate that the MMSI validation framework can be operationalized as a deterministic, auditable front end for AIS data quality control. Using rule references aligned with ITU R M.585 9 and NAVCEN and AMSA operational summaries, the implementation assigns each AIS record a structured set of outputs consisting of a validity label, an MMSI category, and an explicit rule based reason for any rejection or caution flag. Because the decision logic is expressed as transparent pattern and policy checks, the screening remains computationally lightweight and easy to audit, and it can be reproduced on different AIS sources using the same script. Importantly, the parameterizations that govern the screening behavior, such as MID bounds and prefix based patterns, can be versioned and documented, which supports faithful replication across experiments and facilitates controlled updates when numbering standards or operational interpretations evolve.

Beyond binary validity, the rule outputs also function as interpretable indicators of identity related risk. The observed error modes can be organized into four practical indicators: identity format violations, misclassification or misuse signals reflected by unexpected station patterns, MID policy violations, and potential identity collision or cloning cues. The first three indicators are derived directly from MMSI only checks and therefore can be applied even when positional fields are unavailable, while the cloning or collision cue is naturally treated as an extension that becomes testable once time and position attributes are included through a spatio temporal overlap analysis. This layered structure allows the pipeline to

preserve the same core rules while expanding diagnostic depth as additional attributes become available, which is valuable for incremental deployment in real AIS processing environments.

When applied to the two datasets, GFW and NOAA, the screening behavior remains consistent in its logic while revealing source dependent quality characteristics and complementary reporting strengths. On the GFW stream, the rule base produces a concise profile of valid versus invalid identifiers, the composition of MMSI categories, and the most frequent rule based reasons, which directly supports audit triage by highlighting dominant error patterns and reducing identity noise prior to trajectory analytics or machine learning. On the NOAA and NAIS stream, the same rules remain stable at larger scale and generate separated valid and invalid outputs along with summary reports that are well suited for routine quality control and cleanup prioritization, particularly when executed reproducibly through a CLI workflow. Taken together, the category oriented view and the granular reason codes can be harmonized through simple standardization steps, such as consistent column naming and source tagging, enabling cross source comparisons of validity levels, mapping of error profiles, and preparation of cleaned datasets that are ready for subsequent trajectory based or multi sensor modules.

5.2 Recommendations

Based on the findings of this study, several recommendations are proposed to strengthen MMSI based identity screening and its downstream use in maritime analytics. First, the MID reference data should be enhanced by expanding and regularly synchronizing the MID dictionary against authoritative administrative lists and relevant national archives. Where appropriate, the screening can be reinforced with regionalization logic, such as fisheries zones, ALKI corridors, and EEZ boundaries, to improve policy conformance checks. Second, external registration cross checks should be automated whenever access is available, by linking screened MMSI records to official registration or assignment sources to validate MMSI ownership and class specific patterns, including identifiers associated with handheld devices, SAR aircraft, and AtoN units. Third, identity screening should be integrated upstream in kinematics and machine learning pipelines by deploying the rules first validation as an initial gating stage and benchmarking downstream model performance under ablation settings with and without pre filtering, using metrics such as precision and recall to quantify its operational benefit.

References

1. ITU-R, Recommendation ITU-R M.585-9, “Assignment and use of Maritime Mobile Service Identities (MMSI),” Geneva, Switzerland: International Telecommunication Union, 2023.
2. U.S. Coast Guard Navigation Center (NAVCEN), “Automatic Identification System (AIS): Overview,” n.d. [Online]. Available: <https://navcen.uscg.gov/automatic-identification-system-overview> (accessed Oct. 2025).
3. U.S. Coast Guard Navigation Center (NAVCEN), “MMSI formats (including summary tables; MID 201–775),” n.d. [Online]. Available: <https://www.navcen.uscg.gov/mmsi-formats> (accessed Oct. 2025).

4. Australian Maritime Safety Authority (AMSA), “About maritime mobile service identity (MMSI) information,” n.d. [Online]. Available: <https://www.amsa.gov.au/mmsi> (accessed Oct. 2025).
5. Australian Maritime Safety Authority (AMSA), “AIS aids to navigation,” n.d. [Online]. Available: <https://www.amsa.gov.au/safety-navigation/navigating-coastal-waters/ais-aids-navigation> (accessed Oct. 2025).
6. Global Fishing Watch, “Global static dataset of AIS-based apparent fishing effort v3: Format changes (FAQ),” 2025. [Online]. Available: <https://globalfishingwatch.org/faqs/2025-march-static-apparent-fishing-effort-v3-difference-with-map/> (accessed Oct. 2025).
7. NOAA Office for Coastal Management, “AIS data and tools — Frequently asked questions,” 2023 (ver. Nov. 2023). [Online]. Available: <https://coast.noaa.gov/data/marinecadastre/ais/faq.pdf> (accessed Oct. 2025).
8. K. Wolsing, L. Roepert, J. Bauer and K. Wehrle, “Anomaly detection in maritime AIS tracks: A review of recent approaches,” *J. Mar. Sci. Eng.* **10**, 112 (2022).
9. D. A. Pramudhita, “MMSI-Based-GFW: Rules-first MMSI validation for GFW AIS data,” 2025. [Online]. Available: <https://github.com/dhyaaqilla15-phy/MMSI-Based-GFW>
10. D. A. Pramudhita, “MMSI-Based-NOAA: Rules-first MMSI validation for NOAA AIS data,” 2025. [Online]. Available: <https://github.com/dhyaaqilla15-phy/MMSI-Based-NOAA>