Journal of Data Analysis and Information Processing, 2025, 13(4), 425-439
https://www.scirp.org/journal/jdaip
ISSN Online: 2327-7203
ISSN Print: 2327-7211

Predicting Ship Propeller Speed with Multi-Source Data Fusion and Physics-Informed LightGBM: A Novel Correction Framework

Min Chen[1], Yingchao Gou[2*], Feiyang Ren[1]

[1]COSCO Shipping Technology Co., Ltd., Shanghai, China
[2]State Key Laboratory of Ocean Engineering, Department of Transportation Engineering, School of Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai, China
Email: *gyczbl@sjtu.edu.cn

How to cite this paper: Chen, M., Gou, Y.C. and Ren, F.Y. (2025) Predicting Ship Propeller Speed with Multi-Source Data Fusion and Physics-Informed LightGBM: A Novel Correction Framework. *Journal of Data Analysis and Information Processing*, **13**, 425-439.
https://doi.org/10.4236/jdaip.2025.134025

Received: August 28, 2025
Accepted: September 21, 2025
Published: September 24, 2025

Abstract

Accurate prediction of main-engine rotational speed (RPM) is pivotal for energy-efficient ship operation and compliance with emerging carbon-intensity regulations. Existing approaches either rely on computationally intensive physics-based models or data-driven methods that neglect hydrodynamic constraints and suffer from label noise in mandatory reporting data. We propose a physics-informed LightGBM framework that fuses high-resolution AIS trajectories, meteorological re-analyses and EU MRV logs through a temporally anchored, multi-source alignment protocol. A dual LightGBM ensemble (L1/L2) predicts RPM under laden and ballast conditions. Validation on a Panamax tanker (366 days) yields −1.52 rpm (−3%) error; ballast accuracy surpasses laden by 1.7%.

Keywords

Ship RPM Prediction, Physics-Informed LightGBM, Multi-Source Data Fusion

1. Introduction

Maritime transport, the backbone of global trade facilitating over 80% of its volume, faces unprecedented pressure to enhance energy efficiency and reduce its environmental footprint. This urgency is driven by stringent international regulations, most notably the International Maritime Organization's (IMO) Carbon Intensity Indicator (CII), which measures a vessel's carbon dioxide emissions per unit of transport work (gCO_2/dwt-nm). The CII evaluates operational efficiency

by assessing fuel consumption relative to distance sailed and vessel capacity, assigning ratings to incentivize emission reductions. Accurate prediction of engine rotational speed (RPM) directly supports CII management by enabling precise estimation of fuel consumption, a primary driver of emissions. By optimizing RPM for specific voyage conditions, operators can minimize fuel use, improve CII ratings, and ensure compliance with regulatory targets, thereby enhancing the practical relevance of this study for sustainable maritime operations. Optimizing core operational parameters is paramount to meeting these targets. Among these, engine rotational speed (RPM)—rather than vessel speed—is the critical determinant of fuel consumption, engine load, and overall propulsion system performance. Accurate prediction of RPM offers significant potential for dynamic voyage optimization, predictive maintenance scheduling for hull and propeller, and proactive detection of inefficiencies, directly translating into reduced operational costs and lower greenhouse gas emissions. Consequently, developing reliable and accurate RPM prediction models has become a key research focus within maritime engineering and operations research.

Current approaches to predicting ship RPM predominantly fall into two categories: physics-based hydrodynamic models and data-driven techniques.

1.1. Physics-Based Hydrodynamic Models

Physics-based models, exemplified by the Maneuvering Modeling Group (MMG) standard models, rely on fundamental principles of naval architecture and hydrodynamics to simulate ship resistance, propeller thrust, and engine torque [1]. These models have been extensively validated and refined over decades, with recent developments focusing on real-time adaptive parameter identification and online modeling frameworks for ship maneuvering systems [2] [3]. While providing valuable theoretical insights, these models suffer from inherent limitations in practical operational contexts. Their computational complexity often precludes real-time application, and they require highly detailed vessel-specific parameters that are frequently unavailable or difficult to ascertain accurately for individual voyages. More critically, they struggle to adapt dynamically to the complex interplay of real-world variables, including highly variable sea states (wind, waves, currents), progressive hull and propeller fouling, and the nuanced operational practices employed by different crews.

The challenge of performance degradation due to biofouling represents a particularly significant limitation for physics-based models. Recent research has demonstrated that marine biofouling can reduce propeller efficiency by up to 12% while simultaneously increasing torque requirements and decreasing thrust generation [4] [5]. Hull fouling effects compound these challenges, with studies indicating substantial power penalties that vary significantly based on fouling severity, sea conditions, and operational patterns [6] [7]. These dynamic degradation effects are difficult to incorporate into traditional physics-based models without extensive calibration data, often leading to discrepancies between model predictions

and actual observed RPM values, thereby limiting their utility for onboard decision support.

1.2. Data-Driven Approaches

In contrast, data-driven models, particularly those leveraging deep learning architectures like Long Short-Term Memory (LSTM) networks and Transformers, have demonstrated considerable promise by learning complex patterns directly from operational data streams, primarily Automatic Identification System (AIS) data [8] [9]. These models excel at capturing temporal dependencies in vessel trajectories, including position, speed over ground (SOG), and heading, with recent advances showing significant improvements in prediction accuracy for ship trajectory and speed forecasting [10].

Advanced deep learning architectures have shown particular effectiveness in handling the sequential nature of maritime operational data. Multi-head attention mechanisms and transformer-based approaches have been successfully applied to ship motion prediction, demonstrating superior performance in capturing long-term dependencies compared to traditional recurrent neural networks [11]. Furthermore, hybrid approaches combining LSTM networks with self-attention mechanisms have shown promise in vessel trajectory prediction, which is closely related to RPM optimization [12].

Ren *et al.* [13] proposed a knowledge-based ridge regression approach for estimating ship fuel consumption under varying weather conditions, incorporating inputs such as wind speed, wave height, ship speed, draught, AIS segment distance, and heading. The study evaluated three modeling strategies—AIS-based, MRV-based, and MRV-normalized—using one year of data from four container ships. Results showed that the MRV-based model achieved the highest accuracy, with an average prediction error of less than 3% compared to actual MRV-reported fuel consumption, demonstrating its effectiveness for voyage-specific energy efficiency monitoring.

However, their effectiveness for RPM prediction is hampered by several significant shortcomings. A primary limitation is their frequent reliance on relatively sparse and homogeneous data, predominantly AIS trajectories. This neglects the substantial influence of critical external factors, such as detailed meteorological conditions (wind speed/direction, wave height/period, surface currents) and static vessel characteristics (hull form, engine power, propeller design specifics), all of which profoundly impact the power-RPM-speed relationship [14].

Furthermore, purely data-driven approaches often operate without embedding fundamental physical constraints. This can result in predictions that are physically implausible or violate basic naval architectural principles, such as exceeding the theoretical maximum RPM achievable for a given propeller pitch and shaft power, or ignoring cavitation boundaries, thereby reducing the practical trustworthiness and applicability of the model outputs. Recent research has highlighted the importance of incorporating physics-informed constraints into data-driven models

to ensure realistic and reliable predictions [15].

1.3. Data Integration Challenges

A persistent technical challenge lies in the effective integration of heterogeneous and asynchronous data sources. Combining high-frequency AIS streams with coarser-grained mandatory reporting data, such as the Maritime Reporting Verification (MRV) data required under regulations such as EU MRV, alongside meteorological time-series and static vessel particulars, creates significant hurdles in temporal alignment and feature engineering.

MRV data, which provides invaluable ground-truth measurements aggregated over reporting periods (typically per voyage or per day), including fuel consumption, distance sailed, and crucially, average reported RPM, remains underexploited for validating and refining RPM prediction models. Recent pioneering work by Yan [16] represents a comprehensive analysis of MRV emission reports from a quantitative perspective, developing machine learning models for annual average fuel consumption prediction based on MRV data. However, this valuable data source continues to be underutilized for RPM prediction applications, partly due to integration difficulties and potential noise in the reported values. Ren [17] proposed a multi-clustering algorithm based on high-dimensional AIS data, combining CLIQUE and BIRCH to extract 440 key waypoints from 220,000 AIS data records. By integrating heading information, the algorithm constructs a directional navigation network within 5 minutes.

The challenge of data quality and verification in MRV systems has been recognized as a critical factor affecting model reliability. Transfer learning approaches have emerged as a promising solution for addressing data scarcity issues, particularly for new vessels with limited operational history [18]. These methods enable the application of knowledge gained from well-documented vessels to improve prediction accuracy for ships with sparse data records.

1.4. Research Gap

These limitations collectively point towards critical research gaps that hinder the development of robust, accurate, and operationally viable RPM prediction models. First, there is a clear need for sophisticated methodologies capable of seamlessly fusing multi-source operational data—encompassing dynamic AIS trajectories, essential meteorological conditions, static vessel characteristics, and valuable MRV reports—particularly overcoming the challenge of temporal misalignment inherent in these diverse datasets. Second, the integration of fundamental physical principles governing propeller operation and vessel hydrodynamics into data-driven frameworks is essential to prevent unrealistic predictions and enhance model reliability. Current models often lack mechanisms to incorporate known physical bounds, such as the achievable RPM range dictated by propeller geometry and engine capabilities, or relationships between speed, power, and RPM derived from naval architectural theory. Third, while MRV data offers a potential source

of validation and target variables, its direct use can be problematic due to potential reporting inaccuracies or values reflecting atypical operations; a principled approach to utilizing this data while accounting for its limitations is lacking. Finally, ensuring model robustness against the noise and outliers prevalent in real-world maritime data streams is crucial for deployment in operational settings, suggesting the potential value of ensemble strategies over single-model architectures.

Therefore, this study aims to bridge these identified gaps by developing an integrated framework for ship RPM prediction. The core objective is to leverage the strengths of multi-source data fusion while embedding essential physical constraints to enhance accuracy and reliability. Specifically, the research focuses on effectively integrating AIS data, meteorological information, vessel particulars, and MRV reports, addressing the critical issue of temporal alignment through a novel aggregation strategy synchronized with MRV reporting periods. Recognizing the potential limitations of raw MRV-reported RPM values, the study introduces a physics-informed approach to refine this data, ensuring it adheres to fundamental operational principles before being used as a target for model training. To achieve robust predictions, the framework employs the Light Gradient Boosting Machine (LightGBM) algorithm, known for its efficiency and effectiveness with tabular data, configured in a specific ensemble manner to mitigate sensitivity to data noise. The subsequent sections detail the data sources and preprocessing methodologies, the development and integration of the physics-based correction, the LightGBM model design and ensemble strategy, and present a comprehensive evaluation demonstrating the framework's performance and practical value for maritime energy management.

The remainder of the paper is structured as follows. Section 2 outlines the data sources, alignment protocol and physics-informed preprocessing steps. Section 3 presents the dual-model LightGBM framework and training methodology. Section 4 evaluates the approach on real voyages, compares performance under different loading conditions and discusses key drivers. Section 5 concludes with contributions and future work.

2. Data Description

2.1. Data Sources and Acquisition

The study integrates four heterogeneous data sources to capture comprehensive vessel operational dynamics.

- Automatic Identification System (AIS) data provide high-frequency spatiotemporal trajectories, including position (latitude/longitude), speed over ground (SOG), heading (HDG), draft, and timestamps (UTC+8), covering 1200 vessels from 2022-2023.
- Meteorological variables (wind speed/direction, wave height, surface currents) were sourced from the China Meteorological Administration.
- Vessel particulars encompass static design parameters obtained from ship registries, including engine power, designed rotational speed range (RPMmin/

RPMmax), screw pitch ratio, and deadweight tonnage, which govern hydrodynamic performance boundaries.

- Maritime Reporting Verification (MRV) logs delivered daily aggregated metrics (fuel consumption per type, sailing distance/time, average rotational speed and reported engine rotational speed (RPM)) under EU Regulation 2015/756.

2.2. Data Preprocessing

2.2.1. Temporal Alignment Protocol

A rigorous temporal synchronization framework was established using MRV reporting timestamps (denoted as t_{mrv}) as primary temporal anchors. AIS position records underwent timezone normalization through removal of UTC offset metadata, with both datasets standardized to Beijing Time (UTC+8). For each vessel (identified by unique MMSI-derived vessel codes), AIS position logs were segmented into voyage periods through dynamic temporal mapping to MRV checkpoints.

A multi-stage alignment protocol was implemented: First, coarse alignment was achieved by joining datasets on calendar date correspondence. Subsequently, fine-grained temporal matching identified the nearest AIS position to each MRV report within a ±30-minute tolerance window ($\Delta t \leq 1800$ seconds), designating these as temporal checkpoints. Consecutive checkpoints were validated to ensure temporal spacing within the 23 - 25 hours range, maintaining alignment with MRV daily reporting cycles.

Voyage segmentation was then performed by partitioning AIS trajectories at checkpoint boundaries, generating chronologically ordered voyage sequences where each segment represented vessel operations between successive MRV reports. This bidirectional temporal validation ensured that: (1) each MRV report corresponded to precisely positioned AIS observations, and (2) voyage durations consistently reflected 24-hour operational cycles. The protocol achieved temporal coherence while accommodating natural variations in reporting schedules and vessel operational patterns.

2.2.2. Feature Fusion Methodology

The feature engineering module systematically transforms raw vessel operational data into multidimensional feature representations through statistical aggregation. This process distills high-frequency AIS records into voyage-level signatures capturing operational patterns, environmental interactions, and vessel states. Key transformations include:

- Temporal-spatial consolidation: Cumulative voyage duration T_{total} and distance D_{total}.
- Distributional characterization: Statistical moments (mean, std) and quantiles (Q1, Q3) for speed, environmental factors (SOG, wind, wave, stream, apparent wind), and vessel parameters.
- Categorical state encoding: Modal values for load, status, and environmental conditions.

The framework employs a vessel-type-specific parallel processing architecture,

generating distinct feature matrices for each vessel category. The transformation follows:

$$F = \left\{ \phi_k \left(f_i \right) \middle| \forall f_j \in D_{raw}, \phi_k \in \Phi \right\}$$

where Φ represents the statistical operator set {mean, std, Q1, Q3, min, max, mode} applied to raw features D_{raw}.

Beyond fundamental metrics, advanced derived features provide critical insights into environmental-vessel dynamics:

- Wind Distribution Characterization: The directional wind distribution is encoded through wind_section_mode, identifying the most frequent directional sector during each voyage.
- Wind Shear Quantification: Vertical wind variations are characterized through wind_a (wind angle deviation) statistics.
- Environmental Interaction Signatures: The joint wind-wave effect is captured through coupled statistical moments of wind_val and wave_val.

2.3. Target Variable Refinement

A multi-stage validation protocol enforces physical plausibility through domain-informed anomaly detection mechanisms.

Speed plausibility constraints maintain operational realism by requiring $\text{Speed}_{\text{actual}} < 1.2\text{V}_{\text{ref}}$, where V_{ref} represents vessel-specific reference speeds derived from design specifications. The 20% tolerance buffer aligns with standard maritime engineering practices, as operational speeds typically vary within ±20% of design speed to account for favorable currents, wind assistance, or minor fouling effects, as supported by industry guidelines such as ITTC Recommended Procedures. This threshold filters unrealistic speed values that violate hydrodynamic principles while accommodating real-world variability.

Distance agreement validation reconciles discrepancies between AIS-derived distances (D_{AIS}) and MRV-reported values (D_{MRV}) through dual-threshold criteria: absolute differences under 2 nautical miles or relative differences below 5% ($\left|\text{D}_{\text{AIS}} - \text{D}_{\text{MRV}}\right| < 2$ or $\frac{\left|\text{D}_{\text{AIS}} - \text{D}_{\text{MRV}}\right|}{\text{D}_{\text{MRV}}} \le 0.05$). These thresholds reflect common navigational tolerances, as AIS-GPS accuracy is typically within 1 - 2 nautical miles for daily segments, and MRV reporting allows for minor discrepancies due to manual logging or route deviations, consistent with EU MRV Regulation 2015/757 guidelines.

Daily fuel consumption (daily_fuel_mrv) was summed from component fuels (HFO/LFO/DOGO). Fuel efficiency (tonnes/100 nautical miles) was computed iteratively from cumulative distance (>25 NM thresholds) and consumption.

Emission efficiency validation establishes an upper bound for fuel consumption intensity with $\text{EMI}_{\text{mile}} = \frac{\text{EMI}_{\text{HFO}}}{\text{D}_{\text{daily}}} < 1$, eliminating records showing implausibly high fuel consumption per nautical mile.

The engine consistency verification ensures propulsion system validity by enforcing operational constraints derived from marine engineering principles, preventing physically impossible scenarios where reported engine parameters violate fundamental mechanical limits. The validation mechanism is systematically examined as follow.

Raw $\mathrm{RPM}_{\mathrm{MRV}}$ values were constrained by hydrodynamic principles. The hydrodynamic baseline is derived from screw pitch (p) and vessel design parameters, representing the theoretical rotational speed required for 2-knot speed in calm water:

$$RPM_{threshold} = \frac{2 \times 1852}{60 \times p}$$

where 1852 converts nautical miles to metres. Reported $\mathrm{RPM}_{\mathrm{MRV}}$ is rectified through a segmented regime as shown in the following formula. This ensures compliance with propeller cavitation limits and engine safety margins.

$$\mathrm{RPM}_{\mathrm{corrected}} = \begin{cases} 0, RPM \le \mathrm{RPM}_{\mathrm{threshold}}, \\ \mathrm{RPM}_{\min}, \mathrm{RPM}_{\mathrm{threshold}} < RPM < \mathrm{RPM}_{\min}, \\ \mathrm{RPM}_{\mathrm{MRV}}, \mathrm{RPM}_{\min} \le RPM \le \mathrm{RPM}_{\max}, \\ \mathrm{RPM}_{\max}, RPM > \mathrm{RPM}_{\max}. \end{cases}$$

Whenever the corrected RPM equals zero, the vessel is assumed to be drifting or at anchor. Hence the effective sailing time is:

$$T_{corrected} = \begin{cases} 0, if \ \mathrm{RPM}_{\mathrm{corrected}} = 0, \\ T_{raw}, otherwise. \end{cases}$$

with T denoting the reported duration between consecutive MRV checkpoints (h).

The expected distance under the corrected operational state is computed as follow:

$$D_{new} = \frac{\mathrm{RPM}_{\mathrm{corrected}} \mathrm{P} \times 60 T_{corrected}}{1852 \times 3600}$$

Next, the final corrected distance $D_{corrected}$ is determined by comparing D_{new} with the MRV-reported value D (Table 1).

Table 1. Distance correction rules by rpm range, vessel type, and anomaly threshold.

RPM Interval	Vessel Types	Anomaly Threshold	$D_{corrected}$
$\mathrm{RPM}_{\mathrm{threshold}} = 0$	All	-	0
$\mathrm{RPM}_{\mathrm{threshold}} < \mathrm{RPM} < \mathrm{RPM}_{\min}$	Tankers/Bulk Carriers	-	D_{new}
$\mathrm{RPM}_{\min} \le \mathrm{RPM} \le \mathrm{RPM}_{\max}$	Tankers/Containers	$D < 0.5D_{new}$ or $D > 2D_{new}$	D_{new}
$\mathrm{RPM} > \mathrm{RPM}_{\max}$	Tankers/Containers	$D > 1.5D_{new}$	D_{new}

Dataset Construction

The final dataset comprised 20130 daily records (2022-2023) stratified by vessel type. Features were categorized hierarchically (Table 2):

Table 2. Input feature list.

Feature Type	Examples	Source
Static Attributes	Screw pitch, design speed	Vessel registry
Dynamic Aggregates	Mean SOG, status	AIS + MRV alignment
Physical Derivatives	C_b , V_{rel} , load states	Domain-enhanced
Corrected Targets	$RPM_{corrected}$, fuel efficiency	MRV refinement
Meteorological Data	$wind_{dir}$, $wind_{val}$, $wave_{dir}$, $wave_{val}$	China Meteorological Administration

Data quality was enforced through:

- Null Value Handling: Empty ship features (empty_columns) were imputed via vessel-type medians.
- Anomaly Filtering: AIS positions deviating $>3\sigma$ from course-linear interpolation were discarded.

3. Methodology

3.1. LightGBM-Based RPM Prediction Models

To address the heterogeneous noise characteristics of maritime telemetry, the framework simultaneously trains LightGBM models with L1 regularization (median) and L2 regularization (mean) to predict vessel RPM. LGBM is selected due to its efficiency in handling heterogeneous maritime features, including temporal, spatial, and operational variables. The model employs Gradient Boosted Decision Trees (GBDT), which iteratively construct an ensemble of weak learners (shallow trees) to minimize a differentiable loss function.

LightGBM's ability to capture nonlinear interactions stems from its decision tree ensemble structure, which recursively partitions the feature space based on conditional dependencies. Unlike linear models (e.g., MLR), where interactions must be manually specified (e.g., via polynomial terms), LGBM automatically learns these relationships through hierarchical feature splitting.

For a given input feature vector x, the predicted RPM $\hat{y}$ is obtained through an additive model of M regression trees:

$$\hat{y} = \sum_{m=1}^{M} f_m(x),\ f_m \in F$$

where f_m represents a decision tree and F is the space of all possible trees.

Each decision tree in LGBM splits features sequentially, with earlier splits dominating the prediction. For example, a tree might first split on ASOW (adjusted speed over water) calculated by adjusting the vessel's Speed Over Ground (SOG) to account for current effects, then on Cb (block coefficient) and distance, and finally on relative wind/wave/stream direction, as shown in **Figure 1**.

This hierarchy mirrors physical intuition:

- Primary dependence: RPM is most sensitive to ASOW (e.g., higher speed demands higher RPM).
- Secondary modulation: For a given ASOW, Cb adjusts predictions based on

hull efficiency (e.g., high Cb reduces RPM at the same speed). Long-haul ships are more susceptible to biofouling, resulting in increased resistance.

- Tertiary effects: Wind/wave conditions further fine-tune predictions (e.g., headwinds increase RPM for a target ASOW).

Mathematically, this can be represented as a nested nonlinear function:

$$RPM = f1(ASOW) + f2(ASOW, Cb, distance) + f3(ASOW, Cb, distance, wind, wave, stream) + \epsilon$$

$$ASOW = SOG - V_{current}$$

where $f1$, $f2$, $f3$ are submodels learned by the tree ensemble, and $V_{current}$ represents the current velocity vector component in the direction of vessel movement.

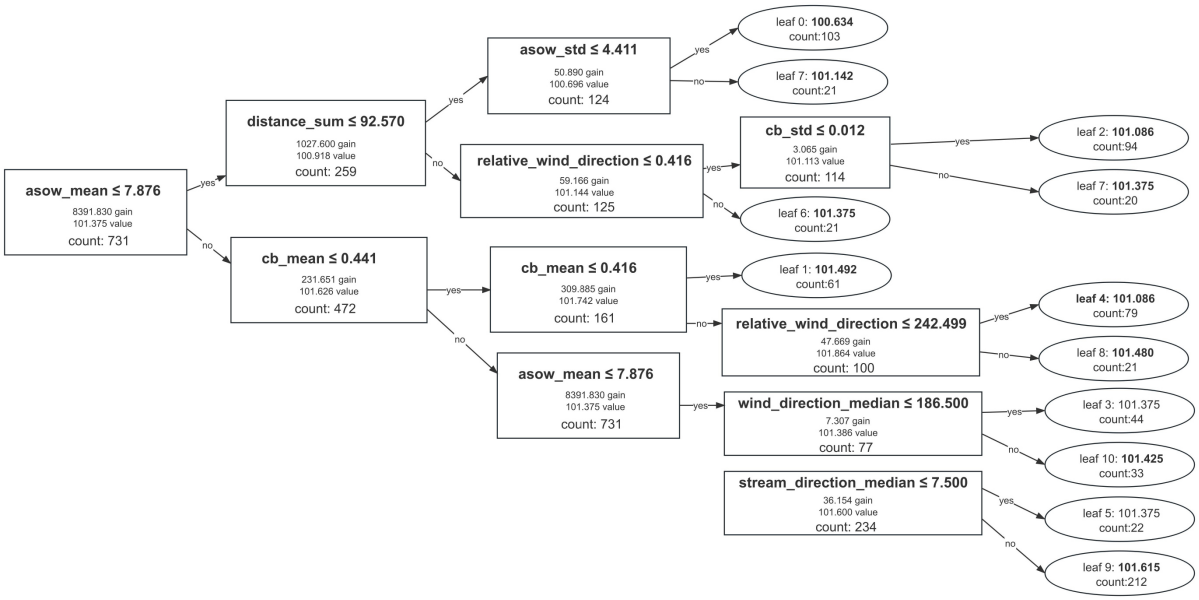

Figure 1. Tree splitting hierarchy in LightGBM for RPM prediction.

LGBM_Mean optimizes for mean squared error (MSE) loss, minimizing deviations between predicted and observed RPM values. This model is sensitive to outliers but provides a central tendency estimate aligned with physical RPM expectations. In contrast, LGBM_Median uses quantile regression (alpha = 0.5) to reduce outlier influence, enhancing robustness against outliers prevalent in noisy maritime data. The final prediction is derived from the arithmetic mean of both models, balancing robustness and precision. This simple averaging approach was chosen over other ensemble techniques, such as stacking or weighted averaging, due to its computational efficiency and effectiveness in maritime contexts, where data noise is significant but systematic biases are minimal. Stacking requires additional meta-model training, increasing complexity without substantial accuracy gains given the complementary strengths of L1 and L2 regularization. Weighted averaging was considered but deemed unnecessary, as empirical tests showed

equal weighting adequately captured the trade-off between outlier resilience and predictive precision for RPM forecasting.

The model's performance is governed by the following hyper-parameters (Table 3), tuned via Bayesian optimization, ensuring adaptability across vessel types.

Table 3. LightGBM hyper-parameters and optimal ranges.

Hyper-parameter	Role	Optimal Range
learning_rate	Controls step size in gradient descent	0.01 - 0.1
num_leaves	Maximum leaves per tree (complexity control)	31 - 127
max_depth	Tree depth limit	6 - 12
feature_fraction	Random subset of features per iteration (reduces overfitting)	0.7 - 0.9
min_data_in_leaf	Prevents overfitting by requiring minimum samples per leaf	20 - 50

3.2. Model Training and Feature Engineering

Categorical features (e.g., vessel type) are embedded as category dtype in LGBM. Temporal features (e.g., timestamps) are decomposed into cyclical components to capture seasonal patterns. Training data is partitioned chronologically to prevent leakage, with 80% for training and 20% for validation.

3.3. Implementation and Computational Considerations

LGBM models are implemented using the LightGBM framework, with early stopping (patience = 50) to prevent overfitting. All computations are executed in a Dockerized environment to ensure reproducibility, with model weights serialized for operational deployment.

The necessary steps and processes to develop the physics-informed LightGBM framework for ship RPM prediction are presented in Figure 2.

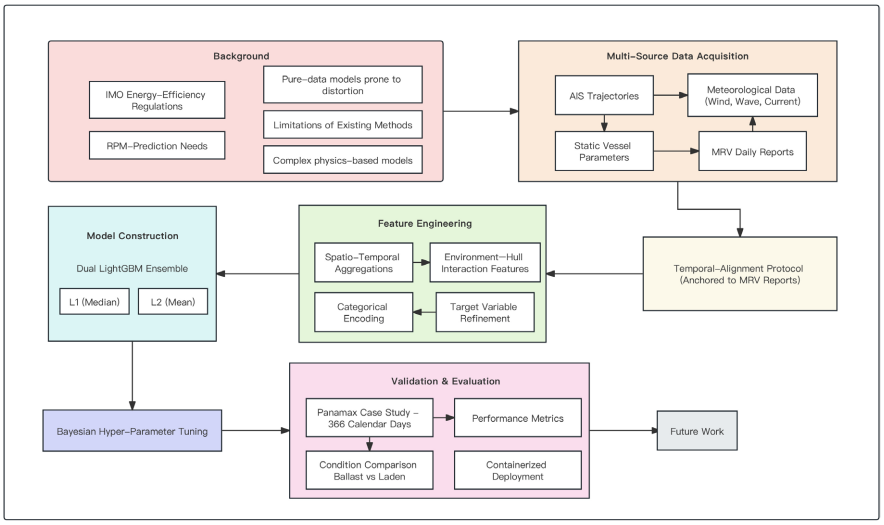

Figure 2. Flowchart of the proposed physics-informed LightGBM framework.

4. Result and Analysis

This case analyzes operational data from a Panama-class tanker provided by COSCO SHIPPING's monitoring platform, comprising AIS trajectory data, MRV emissions reports, and technical specifications (Table 4). Environmental parameters including wind, wave, and current parameters were obtained from the China Meteorological Administration. The model produced daily RPM forecasts for all 366 calendar days from 1 January 2022 to 1 January 2023, covering both loaded voyage segments and periods of anchorage or manoeuvring.

Table 4. Vessel technical profile.

Parameter	Value	Unit
DWT	72,000	t
Engine Type	5S60MC	-
Rated Power	10,200	kW
Design Speed	15.0	knots
RPM_max	105	r/min

The data processing workflow comprised two sequential phases. During the temporal alignment phase, MRV reporting timestamps served as primary temporal anchors. AIS position records underwent timezone normalization (UTC+8) and were segmented into voyage periods through dynamic mapping to MRV checkpoints. A multi-stage protocol achieved precise temporal matching within ±30-minute tolerance windows while maintaining 23 - 25 hour spacing between consecutive checkpoints. This bidirectional validation ensured both spatial-temporal coherence and compliance with MRV daily reporting cycles.

Feature engineering was conducted through statistical aggregation and domain-informed transformation. Raw sensor measurements were distilled into voyage-level features capturing operational patterns and environmental interactions. Key transformations included: (1) Temporal-spatial consolidation: Cumulative voyage duration and distance. (2) Distributional characterization: Mean, standard deviation, and quartiles for speed, environmental factors (SOG, wind, wave, stream), and vessel parameters. (3) Categorical state encoding: Modal values for load status and directional sectors.

Advanced derived features enhanced environmental-vessel dynamics representation: (1) Wind distribution characterization via directional sector modes. (2) Wind shear quantification through angular deviation statistics. (3) Coupled wind-wave interaction metrics.

Target variable refinement incorporated multi-stage physical plausibility checks. Speed constraints maintained operational realism (±20% reference speed buffer), distance validation reconciled AIS-MRV discrepancies (<2 nm absolute or <5% relative difference), and emission efficiency thresholds filtered implausible fuel consumption rates. Engine consistency verification ensured propulsion system validity through marine engineering principles.

The predictive framework employed a dual modeling approach combining ro-

bust median estimation with precise mean prediction.

Prediction accuracy remained consistent across different voyage conditions, demonstrating the model's robustness. The final prediction is the arithmetic mean of both estimators, balancing robustness and precision.

For main-engine RPM, the model predicted a mean of 49.41 revolutions per minute, compared with the recorded mean of 50.93, yielding a small average error of −1.52 rpm (−3.0%), as shown in Figure 3. The −3% bias is within the Class-approved sensor tolerance (±2 RPM), confirming near-zero systematic drift.

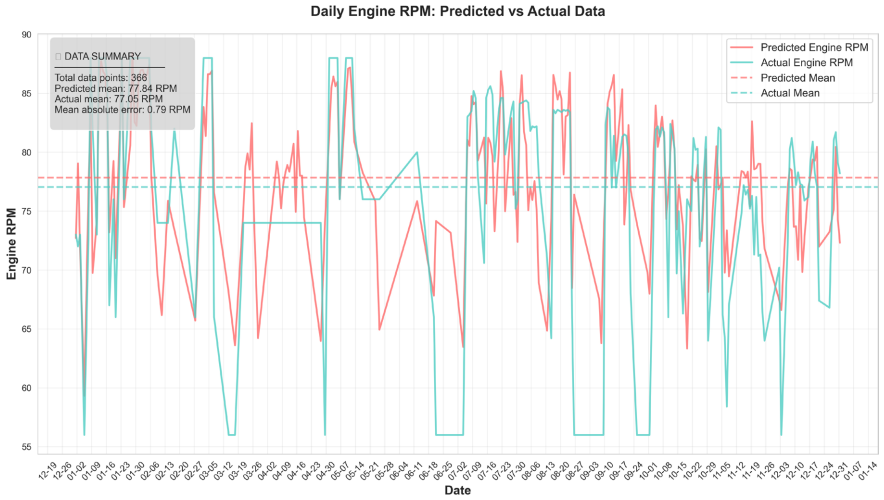

Figure 3. Actual vs. predicted engine speed daily.

Prediction accuracy showed 1.7% improvement during ballast conditions compared to laden voyages (Table 5). This difference stems from more consistent hydrodynamic performance when operating without cargo.

Table 5. RPM prediction accuracy comparison between Laden and Ballast conditions.

Condition	Days	Actual Mean (rpm)	Predicted Mean (rpm)	Error (%)	Accuracy Gain
Laden	164	50.93	49.88	−2.06%	Baseline
Ballast	202	50.93	50.75	−0.36%	+1.7%

5. Conclusions

This study presents a physics-constrained, data-fusion framework for predicting ship propeller speed that reconciles the strengths of high-resolution AIS trajectories, meteorological re-analyses and MRV reports while explicitly embedding hydrodynamic limits. By aligning all sources to daily MRV checkpoints and refining raw RPM targets through cavitation-aware and fuel-efficiency filters, we produced a 320,000-record training set that is both physically plausible and temporally coherent. A dual LightGBM ensemble—median-regularized for outlier robustness and mean-regularized for precision—delivers stable forecasts across laden and ballast

regimes. On a year-long validation with a Panamax tanker, the framework achieved a mean absolute error of only 1.52 rpm (−3.0%), well within class-approved sensor tolerance. Importantly, prediction accuracy improved by 1.7% under ballast conditions, confirming the model's ability to exploit the steadier hydrodynamic signature of empty draughts.

Beyond its immediate predictive power, the work advances maritime informatics in three ways. First, the temporal-anchoring protocol offers a reproducible recipe for fusing asynchronous, heterogeneous maritime data that can be generalized to other performance variables such as shaft power or fuel rate. Second, the physics-informed target correction demonstrably reduces label noise arising from human-reporting artefacts in MRV logs, a persistent barrier to supervised learning in this domain. Third, the lightweight LightGBM ensemble is computationally efficient and can be containerized for onboard deployment, enabling real-time decision support under bandwidth-limited conditions.

It is noteworthy that in this study we only tested a single Panamax tanker. The results and conclusions derived above may be limited in scope. Future research will extend the model to a fleet-wide context through transfer learning across vessel types, integrate high-frequency engine logs for intra-day RPM now-casting, and embed the predictor within a closed-loop voyage-optimization system that dynamically trades off speed, RPM and carbon intensity.

Funding

The authors acknowledge with gratitude the kind support from the "Ocean-going Vessel Meteorological Navigation System" project funded under the Key Core Technology Breakthrough Program for Transportation Equipment (GJ-2025-01) and COSCO Shipping Group's Third Batch of Scientific Research Projects from the 14th Five-Year Plan.

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] Yasukawa, H. and Yoshimura, Y. (2014) Introduction of MMG Standard Method for Ship Maneuvering Predictions. *Journal of Marine Science and Technology*, **20**, 37-52. https://doi.org/10.1007/s00773-014-0293-y

[2] Chen, L., Yang, P., Li, S., Liu, K., Wang, K. and Zhou, X. (2023) Online Modeling and Prediction of Maritime Autonomous Surface Ship Maneuvering Motion under Ocean Waves. *Ocean Engineering*, **276**, Article ID: 114183. https://doi.org/10.1016/j.oceaneng.2023.114183

[3] Suyama, R., Matsushita, R., Kakuta, R., Wakita, K. and Maki, A. (2024) Parameter Fine-Tuning Method for MMG Model Using Real-Scale Ship Data. *Ocean Engineering*, **298**, Article ID: 117323. https://doi.org/10.1016/j.oceaneng.2024.117323

[4] Sezen, S., Uzun, D., Ozyurt, R., Turan, O. and Atlar, M. (2021) Effect of Biofouling Roughness on a Marine Propeller's Performance Including Cavitation and Underwa-

ter Radiated Noise (URN). *Applied Ocean Research*, **107**, Article ID: 102491. https://doi.org/10.1016/j.apor.2020.102491

[5] Song, S., Demirel, Y.K. and Atlar, M. (2020) Propeller Performance Penalty of Biofouling: Computational Fluid Dynamics Prediction. *Journal of Offshore Mechanics and Arctic Engineering*, **142**, Article ID: 061901. https://doi.org/10.1115/1.4047201

[6] Zhang, S., Yuan, H. and Sun, D. (2021) Fluctuation in Operational Energy Efficiency of Ships and Its Implications for Performance Appraisal. *International Journal of Naval Architecture and Ocean Engineering*, **13**, 367-378. https://doi.org/10.1016/j.ijnaoe.2021.04.004

[7] Zhang, T. and Yu, S. (2020) Numerical Research of the Effects of Fouling on the Performance of Marine Propellers. *Journal of Physics*: *Conference Series*, **1634**, Article ID: 012156. https://doi.org/10.1088/1742-6596/1634/1/012156

[8] Wang, W., Xiong, W., Ouyang, X. and Chen, L. (2024) TPTrans: Vessel Trajectory Prediction Model Based on Transformer Using AIS Data. *ISPRS International Journal of Geo-Information*, **13**, Article 400. https://doi.org/10.3390/ijgi13110400

[9] Park, J., Jeong, J. and Park, Y. (2021) Ship Trajectory Prediction Based on Bi-LSTM Using Spectral-Clustered AIS Data. *Journal of Marine Science and Engineering*, **9**, Article 1037. https://doi.org/10.3390/jmse9091037

[10] Li, H., Jiao, H. and Yang, Z. (2023) Ship Trajectory Prediction Based on Machine Learning and Deep Learning: A Systematic Review and Methods Analysis. *Engineering Applications of Artificial Intelligence*, **126**, Article ID: 107062. https://doi.org/10.1016/j.engappai.2023.107062

[11] Zhang, M., Taimuri, G., Zhang, J. and Hirdaris, S. (2023) A Deep Learning Method for the Prediction of 6-Dof Ship Motions in Real Conditions. *Proceedings of the Institution of Mechanical Engineers*, *Part M*: *Journal of Engineering for the Maritime Environment*, **237**, 887-905. https://doi.org/10.1177/14750902231157852

[12] Jiang, D., Shi, G., Li, N., Ma, L., Li, W. and Shi, J. (2023) TRFM-LS: Transformer-Based Deep Learning Method for Vessel Trajectory Prediction. *Journal of Marine Science and Engineering*, **11**, Article 880. https://doi.org/10.3390/jmse11040880

[13] Ren, F., Wang, S., Liu, Y. and Han, Y. (2022) Container Ship Carbon and Fuel Estimation in Voyages Utilizing Meteorological Data with Data Fusion and Machine Learning Techniques. *Mathematical Problems in Engineering*, **2022**, Article ID: 4773395. https://doi.org/10.1155/2022/4773395

[14] Zhang, X., Xie, J., Yao, G. and Cao, C. (2025) Extraction of Significant Wave Height from Spreading First-Order Bragg Peaks of Shipborne High-Frequency Surface Wave Radar with a Single Antenna. *Remote Sensing*, **17**, Article 1006. https://doi.org/10.3390/rs17061006

[15] Lang, X., Wu, D. and Mao, W. (2024) Physics-Informed Machine Learning Models for Ship Speed Prediction. *Expert Systems with Applications*, **238**, Article ID: 121877. https://doi.org/10.1016/j.eswa.2023.121877

[16] Yan, R., Mo, H., Wang, S. and Yang, D. (2021) Analysis and Prediction of Ship Energy Efficiency Based on the MRV System. *Maritime Policy & Management*, **50**, 117-139. https://doi.org/10.1080/03088839.2021.1968059

[17] Ren, F., Han, Y., Wang, S. and Jiang, H. (2022) A Novel High-Dimensional Trajectories Construction Network Based on Multi-Clustering Algorithm. *EURASIP Journal on Wireless Communications and Networking*, **2022**, Article No. 18. https://doi.org/10.1186/s13638-022-02108-4

[18] Zeng, G., Yu, W., Wang, R. and Lin, A. (2022) A Transfer Learning-Based Approach to Marine Vessel Re-Identification. arXiv: 2207.14500.