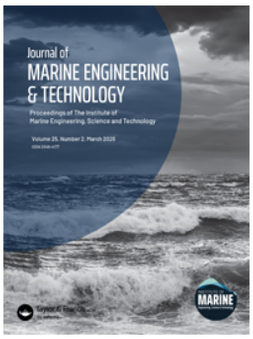

Taylor & Francis
Taylor & Francis Group

Journal of Marine Engineering & Technology

ISSN: 2046-4177 (Print) 2056-8487 (Online) Journal homepage: www.tandfonline.com/journals/tmar20

Machine learning for ship fuel consumption prediction from sensory data: a comparative analysis

Nabil Habib Zahmani, Onur Yuksel, Eduardo Blanco-Davis & Nikolaos Tsoulakos

To cite this article: Nabil Habib Zahmani, Onur Yuksel, Eduardo Blanco-Davis & Nikolaos Tsoulakos (14 May 2026): Machine learning for ship fuel consumption prediction from sensory data: a comparative analysis, Journal of Marine Engineering & Technology, DOI: 10.1080/20464177.2026.2673233

To link to this article: https://doi.org/10.1080/20464177.2026.2673233

Published online: 14 May 2026.

JOURNAL OF MARINE ENGINEERING & TECHNOLOGY
https://doi.org/10.1080/20464177.2026.2673233

Taylor & Francis
by informa

RESEARCH ARTICLE

OPEN ACCESS Check for updates

Machine learning for ship fuel consumption prediction from sensory data: a comparative analysis

Nabil Habib Zahmani[a], Onur Yuksel [a,b], Eduardo Blanco-Davis[a] and Nikolaos Tsoulakos[c]

[a]School of Engineering, Liverpool Logistics Offshore and Marine Research Institute (LOOM), Liverpool John Moores University, Liverpool, UK; [b]Marine Engineering Department, Maritime Faculty, Zonguldak Bülent Ecevit University, Ereğli/Zonguldak, Türkiye; [c]Laskaridis Shipping Co. Ltd., Kifissia, Greece

ABSTRACT

Machine learning models are increasingly used to predict ship fuel consumption from high-frequency operational data. However, the strong temporal autocorrelation of shipboard sensor measurements raises concerns regarding the validity of commonly adopted evaluation practices. Many existing studies rely on random data partitioning and cross-validation schemes that can introduce information leakage, resulting in overly optimistic performance estimates. This study examined the impact of the validation strategy on fuel consumption prediction using one-minute operational data collected from a bulk carrier over approximately 16 months. Gradient-boosted decision tree models (XGBoost and CatBoost) and deep learning architectures were evaluated using three validation schemes: random train–test splitting, blocked chronological hold-out, and rolling-window temporal evaluation. The results demonstrate that random splitting substantially inflates the predictive performance, with coefficients of determination exceeding 0.99, whereas temporally consistent validation reveals significantly reduced accuracy and performance degradation across unseen operating periods. Under realistic temporal testing, gradient-boosted models exhibit greater robustness than deep learning models, which exhibit higher sensitivity to distributional shifts. These findings highlight the critical importance of temporally aware validation for obtaining credible generalisation estimates in ship fuel consumption modelling.

ARTICLE HISTORY
Received 5 January 2026
Accepted 8 May 2026

KEYWORDS
Ship fuel consumption prediction; machine learning; temporal validation; time-series data; gradient boosting; shipboard sensor data

1. Introduction

International shipping remains a critical enabler of global trade; however, it is also a significant source of greenhouse gas emissions, which has intensified pressure on the sector to decarbonise (Kim et al. 2024; Yan et al. 2024; Nguyen et al. 2025). As maritime transport continues to expand, fuel consumption and carbon intensity remain central operational and policy concerns. Therefore, recent literature has increasingly emphasised predictive and operationally useful systems that can support fuel efficiency management, carbon accounting, and emissions reduction under real operating conditions (Kim et al. 2024; Li et al. 2024; Marijan et al. 2025).

To address these challenges, the International Maritime Organization (IMO) has implemented regulations such as the Energy Efficiency Design Index (EEDI), Energy Efficiency Existing Ship Index (EEXI), and Carbon Intensity Indicator (CII), all aimed at monitoring and reducing ship emissions (IMO 2021, 2023). In parallel, regional mechanisms have gained importance, most notably the European Union Emissions Trading System (EU ETS), which since 2024 has included maritime transport under its greenhouse-gas cap-and-trade framework. Therefore, large vessels calling at ports in the European Economic Area (EEA) are increasingly exposed to carbon pricing and compliance pressures. Considering these international and regional measures, accurate modelling and prediction of fuel consumption are essential not only for improving operational efficiency and reducing fuel costs, but also for supporting carbon accounting, voyage planning, and regulatory compliance (Ferlita et al. 2024; Lee et al. 2024; Li et al. 2024; Yan et al. 2024).

Traditional fuel consumption estimation relies heavily on empirical formulations or simplified physical models. Although these approaches remain useful, recent comparative and review studies have shown that they often struggle to capture the nonlinear, high-dimensional, and time-varying relationships among operational, environmental, and technical variables under real operating conditions (Luo et al. 2024; Marijan et al. 2025; Liang et al. 2025). With the growth of shipboard sensing, digitalisation, and high-frequency operational data, machine learning (ML) and Artificial Neural Network (ANN) methods offer powerful alternatives capable of modelling these complex interactions and delivering vessel-specific predictions (Kim et al. 2024; Nguyen et al. 2025). Recent studies have shown strong predictive performance for both tree-based ensemble methods and neural-network models when onboard sensor data are available, particularly where rich operational and environmental inputs are combined (Y. Chen et al. 2024; Fan et al. 2024; M. Wang et al. 2024; Zhang et al. 2024)

Simultaneously, recent studies have demonstrated that predictive performance alone is insufficient for maritime applications. Studies on application-oriented testing, interpretability, and operational deployment increasingly show that model usefulness depends on data quality, validation design, computational burden, and the ability to support real-world decision-making rather than only benchmark accuracy (Nguyen et al. 2023; H. Wang et al. 2023; Kim and Roh 2024). These issues are especially important in vessel-specific settings, where operational data are strongly temporally correlated, and models may be used for performance monitoring, energy management, and decarbonisation-related decision support.

CONTACT Onur Yuksel o.yuksel@ljmu.ac.uk School of Engineering, Liverpool Logistics Offshore and Marine Research Institute (LOOM), Liverpool John Moores University, Byrom Street, Liverpool L3 3AF, UK; Marine Engineering Department, Maritime Faculty, Zonguldak Bülent Ecevit University, Hacı Eyüp Street, No:1, Kdz, Ereğli/Zonguldak 67300, Türkiye

Despite this progress, there are important limitations in the current literature. First, many studies still rely on noon reports, voyage summaries, or other low-frequency datasets that do not adequately capture short-term variations in fuel consumption under real operating conditions. Recent work has shown that the diversity, quality, and temporal granularity of ship data strongly influence model behaviour and applicability, and that high-frequency sensor data can materially improve predictive fidelity and vessel-specific modelling (Cai et al. 2024; Gupta et al. 2024; Kim et al. 2025; Marijan et al. 2025). Second, although increasingly sophisticated models have been proposed, many studies continue to evaluate predictive performance using random train-test splitting or similar strategies that do not consider the temporal dependence of ship operational data. This can lead to information leakage and overly optimistic performance predictions. Recent methodological and benchmarking studies have reinforced the need for validation strategies that respect temporal structure, time granularity, and realistic operating conditions (Nguyen et al. 2023; Luo et al. 2024; Chen et al. 2025; Viga et al. 2025; Yan et al. 2025). Third, the literature often prioritises predictive accuracy without equivalent attention to interpretability, computational efficiency, extrapolation performance, and operational feasibility, all of which are important for implementation in shipboard or near-real-time decision-support systems (H. Wang et al. 2023; Kim and Roh 2024; Li et al. 2024; St-Pierre et al. 2024; Ruan et al. 2025).

Against this background, the present study develops and compares ML models for ship fuel consumption prediction using more than one year of high-frequency operational sensor data collected from a dry bulk carrier. The study focuses on eXtreme Gradient Boosting (XGBoost), Categorical Boosting (CatBoost), and ANN models, and evaluates them using both random validation and temporally consistent validation strategies. In doing so, it addresses a clear gap in the literature by examining how validation design affects the credibility of reported model performance in temporally correlated ship data, while also considering the trade-off between predictive accuracy and practical applicability. Therefore, this study contributes to the methodological assessment of ship fuel-consumption models and to the broader goal of data-driven maritime decarbonisation.

1.1. Problem statement

Traditional fuel consumption estimation relies heavily on empirical formulations or simplified physical models. Although useful, recent comparative studies have shown that these approaches often fail to capture the nonlinear and high-dimensional relationships among operational, environmental, and technical variables under real ship operating conditions (Luo et al. 2024; Marijan et al. 2025; Liang et al. 2025). With the rise of big data from shipboard sensors, ML and ANN methods present powerful alternatives capable of modelling complex interactions and offering vessel-specific predictions (Kim et al. 2025; Nguyen et al. 2025).

Existing approaches are limited in four key respects. First, many rely on noon reports or sparse datasets that lack the temporal resolution needed to capture short-term variations in fuel consumption. Second, although advanced models, such as gradient boosting and deep-learning architectures, have achieved high predictive accuracy, their practical deployment is often constrained by limited interpretability, higher computational cost, and dependence on specialised hardware. Third, many studies still use random validation strategies that ignore temporal correlation in ship operational data, which may overestimate predictive performance (Luo et al. 2024; Chen et al. 2025; Yan et al. 2025). Fourth, much of the literature emphasises theoretical accuracy without fully addressing operational feasibility in the context of regulatory frameworks such as CII, EEXI, and EU ETS (Ferlita et al. 2024; Li et al. 2024).

This study addresses these shortcomings by developing and systematically evaluating ML models and selected ANN architectures using high-frequency sensor data from a real bulk carrier. By situating the evaluation within a vessel-specific operational context and explicitly comparing random and temporally consistent validation strategies, the analysis provides practical insights into how predictive models can support regulatory compliance, operational optimisation, and cost reduction.

1.2. Research aim and objectives

This study aims to develop and compare ML models for ship fuel-consumption prediction using high-resolution operational sensor data, with particular emphasis on the impact of validation strategy on model performance and generalisation.

The specific objectives are as follows:

- Analyse and preprocess high-frequency shipboard sensor data, addressing missing values, noise, and anomalous measurements.
- Construct domain-informed features representing key operational and environmental drivers of fuel consumption.
- Implement representative machine learning models, including gradient-boosting methods and selected ANN architectures.
- Evaluate predictive performance using standard error metrics under different validation strategies, including random splitting and temporally consistent validation schemes.
- Quantify the extent to which random data partitioning overestimates model performance in temporally correlated ship operational data.
- Assess the relative robustness and practical suitability of different model classes when generalising to unseen operating periods.

1.3. Scope and assumptions

This study focuses on a single dry bulk carrier equipped with a comprehensive shipboard sensor system, covering more than one year of continuous operational data recorded at a high temporal resolution. The developed models are vessel-specific and are not intended to provide fleet-level generalisation or deployment-ready digital twin solutions. Instead, the scope emphasises methodological evaluation of predictive performance under realistic validation conditions.

The analysis assumes that the available sensor measurements are broadly reliable, while acknowledging that noise, missing data, and measurement uncertainty are inherent to shipboard data acquisition. The study does not attempt to directly translate prediction errors into regulatory compliance metrics or economic outcomes. Such extensions are proposed as potential trajectories for subsequent investigation.

2. Literature review

In recent years, ML and Artificial Intelligence (AI) methods have been increasingly employed to model and predict ship-level fuel consumption and carbon emissions. This development reflects both the rapid growth of shipboard sensing and the increasing need for operational tools that can support energy efficiency, carbon accounting, and regulatory compliance. Recent review studies indicate that maritime fuel consumption modelling has shifted from simplified empirical estimation and low-frequency reporting to data-driven prediction based on onboard measurements, multisource data integration, and digital performance-monitoring systems (Yan et al. 2024; Kim

et al. 2025; Nguyen et al. 2025). This shift is important because ship fuel consumption is shaped by nonlinear interactions among vessel speed, draft, trim, weather, machinery condition, hull state, and voyage context, which all vary over time and across operating regimes. Earlier synthesis studies also noted this transition, but more recent studies have placed much greater emphasis on high-frequency in-service data, temporally structured learning, and operational implementation rather than on model accuracy alone (Nguyen et al. 2023; Z. Wang et al. 2024).

From a broad methodological perspective, the literature can be grouped into empirical approaches, physics-based or grey-box models, and purely data-driven approaches. Empirical and simplified physical models remain valuable because they are transparent, computationally efficient, and closely linked to established engineering principles. They are useful for baseline estimation, early-stage assessment, and applications in which data availability is limited. However, recent comparative studies have shown that their simplifying assumptions can reduce accuracy under real operating conditions, particularly when the objective is to capture vessel-specific variability across changing routes, loading states, and environmental conditions (Luo et al. 2024; Marijan et al. 2025; Liang et al. 2025). This limitation has encouraged the wider adoption of ML approaches that can learn complex patterns directly from operational data. Recent work on simplified physical versus data-driven estimation and hybrid modelling confirms that the main advantage of data-driven methods lies in their ability to absorb operational heterogeneity which is difficult to encode explicitly in conventional engineering formulations (H. Wang et al. 2023; Ruan et al. 2025; Liang et al. 2025).

Among data-driven approaches, ensemble-based models such as Random Forest, Extra Trees, and Gradient Boosting Machines have attracted considerable attention because of their strong predictive accuracy and robustness on structured operational datasets. More recent comparative studies indicate that boosting models such as XGBoost and CatBoost remain among the most competitive methods for ship fuel-consumption prediction when high-quality onboard data are available (Fan et al. 2024; Luo et al. 2024). These models are effective in handling nonlinear relationships, variable interactions, and relatively large feature spaces. They are also attractive because they generally require less training effort than more complex deep learning architectures and can provide partial interpretability through feature-importance analysis or related explanatory tools. In maritime applications, this balance between predictive performance and transparency is particularly relevant because model outputs may need to support operational decisions, performance auditing, or compliance-related reporting rather than only academic benchmarking. Recent studies on vessel performance deterioration and fuel-use estimation continue to favour tree-based models when the goal is to obtain stable performance on noisy operational data with manageable computational effort (Kim and Roh 2024; Themelis et al. 2024; Mittendorf et al. 2025).

ANN models provide a different set of advantages. Recent work includes Feed-forward Neural Networks (FNN), Convolutional Neural Networks (CNN), Recurrent Neural Networks (RNN), Long Short-Term Memory models (LSTM), self-attention-enhanced architectures, dual-attention networks, digital-twin-oriented models, and broader time-series deep-learning for ship fuel-consumption prediction (Sanguino et al. 2024; Y. Chen et al. 2024; M. Li et al. 2024; M.Wang et al. 2024; Chen et al. 2025). Like fuel consumption prediction, ANNs are actively used in energy efficiency and management in ship systems (Yüksel and Köseoğlu 2020; Yüksel and Köseoğlu 2022; Tadros et al. 2025). These approaches are particularly relevant when ship operational data exhibit lagged effects, sequential dependencies, or complex temporal structures. Studies based on real ship operational datasets have shown that explicitly incorporating temporal context can improve forecasting performance, especially in short-term prediction tasks and under changing operating conditions (Zhang et al. 2024; Chen et al. 2025; Viga et al. 2025).

Recent bulk-carrier and smart-ship case studies further show that ANN models become more compelling when rich metocean, voyage-context, or onboard flowmeter data are available (Cho and Lee 2024; Lee et al. 2024; Zhang et al. 2024). However, these gains are often accompanied by greater training complexity, more demanding hyperparameter tuning, higher computational requirements, and lower interpretability than tree-based ensemble models. Therefore, recent literature increasingly treats model selection as a trade-off between temporal sensitivity, prediction accuracy, explainability, and computational burden rather than as a simple ranking of algorithms. This trade-off is also visible in recent frameworks 2026 work on graph-based and attention-based architectures, which further extend the modelling capacity but also reinforce the need to judge model quality beyond raw benchmark performance (Zhong et al. 2026).

A further development in the recent literature is the increasing use of hybrid and grey-box approaches. These models aim to combine the interpretability and extrapolation strengths of physical reasoning with the flexibility of machine learning. Related digital twin and surrogate modelling studies point in the same direction, seeking to embed richer operational structure into data-driven prediction while retaining engineering plausibility (Sanguino et al. 2024; Zhang et al. 2024). Recent studies suggest that such approaches may offer an attractive compromise in ship fuel-consumption modelling, particularly where engineers seek both predictive accuracy and a degree of consistency with known propulsion or hydrodynamic behaviour (Marijan et al. 2025; Liang et al. 2025). Related work on grey-box stacking and physics-guided learning argues that hybrid designs can improve extrapolation beyond the observed operating range, which is especially relevant in ship applications where route, draft, weather, and machinery state can shift over time (Yuksel et al. 2023; Karaçay et al. 2024; Ruan et al. 2025). Nevertheless, hybrid models remain less common than purely data-driven approaches, and their implementation often requires greater domain expertise, additional modelling assumptions, or more extensive calibration efforts. Although promising, they have not yet displaced ensemble or ANN models as the dominant tools in recent maritime ML literature.

Recent literature highlights the growing role of sensor-based ship performance monitoring and data preprocessing. High-frequency onboard datasets can significantly improve model accuracy and vessel-specific relevance; however, they also introduce practical challenges related to sensor reliability, missing values, noise, drift, and outlier detection. Current studies on ship performance analysis emphasise that careful preprocessing is essential for obtaining credible predictive models from in-service data and for avoiding misleading patterns caused by data quality problems (Gupta et al. 2024; Kim et al. 2025). Studies on sensor quality and missing-value compensation have reached similar conclusions, showing that prediction quality can deteriorate significantly when drift, aberrations, or missing measurements are not handled explicitly (Alexiou et al. 2023; Cai et al. 2024; Velasco-Gallego et al. 2026). Related studies on data integration and noise cleaning have also shown that preprocessing choices directly affect the stability of downstream fuel consumption models, particularly in high-frequency engine data streams (Chen et al. 2024). Similarly, research on real-ship fuel-consumption prediction shows that model performance is strongly affected by the diversity, quality, and quantity of the operational data used for training (Cai et al. 2024). These findings are particularly relevant for studies based on shipboard sensor streams, because data conditioning is

not merely a technical preparation step but a core methodological requirement that shapes the reliability of the final model.

Therefore, the question of data sources is central to the current state of knowledge. A substantial part of the earlier and even some recent literature still relies on noon reports, voyage summaries, Automatic Identification System (AIS)-derived variables, and other aggregated records because they are accessible and often available over long time horizons. Such data sources are useful for broad trend analyses and can support certain forms of retrospective assessments. However, they cannot fully capture short-term operational variability, manoeuvring behaviour, machinery response, or rapid changes in environmental loading. More recent analyses suggest that results obtained from sparse reporting data are not directly comparable with those based on continuous onboard measurements, and that high-frequency sensor data enable more reliable vessel-specific modelling and a more realistic representation of operational dynamics (Cai et al. 2024; Marijan et al. 2025; Sharma et al. 2025). This distinction is particularly important in studies intended to support operational decision-making, where the model must respond to variations that may be invisible in low-frequency data. This is also reflected in recent transfer-learning studies, which treat high-frequency sensor data and noon reports as related but not interchangeable sources (Sharma et al. 2025).

The literature also shows that multi-source data fusion is becoming increasingly important. Several recent studies have combined onboard operational measurements with environmental, hydrometeorological, or voyage-context variables to improve predictive performance and better represent the conditions under which fuel is consumed (Zhu et al. 2021; Fan et al. 2024; Li et al. 2024; Zhang et al. 2024; Mohamed et al. 2025). This development reflects a broader recognition that ship fuel consumption is not determined solely by propulsion variables. Instead, it emerges from the interaction between technical conditions, operational behaviour, route context, and environmental forcing. Recent deep learning studies based on multisource inputs, including dual-attention and feature-selection frameworks, have reached similar conclusions and shown that data fusion becomes more valuable as models move from laboratory settings to realistic operational environments (Li et al. 2024; Liu et al. 2024; Zhou et al. 2024). Therefore, models trained on richer operational and environmental inputs are generally better positioned to support real-world applications than models based only on a narrow subset of ship variables. Bulk-carrier case studies using in-service metocean and operational data further support this view, showing that environmental integration can materially improve long-horizon generalisation and vessel-specific realism (Zhang et al. 2024; Mohamed et al. 2025).

Another critical consideration involves the framing of model performance within the current state-of-art A significant proportion of the literature continues to prioritise conventional error metrics such as Root Mean Square Error (RMSE), Mean Average Percentage Error (MAPE) or coefficient of determination (R^2), which remain fundamental for comparative analysis. However, current studies have increasingly asserted that predictive accuracy alone is insufficient, particularly when models are intended for shipboard integration, operational decision support, or regulatory compliance (Kim and Roh 2024; Luo et al. 2024; Nguyen et al. 2025). In these contexts, factors such as interpretability, robustness to data stochasticity, computational efficiency, extrapolation behaviour, and practical operability emerge as vital criteria. This explains the sustained interest in ensemble models despite the prevalence of deep learning architectures. They often provide a more advantageous compromise between predictive precision and operational viability. Recent studies exploring the trade-off between interpretability and accuracy reinforce this perspective, suggesting that model selection should be dictated by the intended end-use rather than benchmark performance in isolation (S. Wang et al. 2023; Kim and Roh 2024).

A particularly important limitation in the literature concerns validation strategies. A significant number of studies continue to evaluate model performance using random train-test splitting or cross-validation procedures; however, these approaches often fail to account for the temporal autocorrelation inherent in ship operational data. Because nearby observations often share similar weather, route, loading, and machinery conditions, random partitioning may lead to information leakage and inflated performance estimates. Recent benchmarking and methodological studies have highlighted the need for testing regimes that respect temporal structure, time granularity, and realistic generalisation conditions (Nguyen et al. 2023; Luo et al. 2024; Chen et al. 2025; Viga et al. 2025; Yan et al. 2025). Current studies on adaptive prediction, transfer learning, and out-of-distribution forecasting reinforce this point by showing that model rankings can change materially once evaluation departs from simple in-distribution random splitting (Gao et al. 2025; Luo et al. 2025; Son and Kim 2026). This issue directly affects the credibility of reported model performance and the value of published algorithm comparisons. Despite growing awareness of this problem, temporally consistent validation remains insufficiently addressed in much of the existing maritime fuel consumption literature. Recent analyses on adaptive and online frameworks also suggest that evaluation under dynamic or out-of-distribution conditions will become increasingly important as prediction models move closer to operational deployment (Gao et al. 2025; Luo et al. 2025).

Furthermore, the literature demonstrates that predictive modelling is increasingly aligned with broader maritime digitalisation and decarbonisation objectives. Recent studies on near-real-time carbon accounting, rapid CII prediction, and 'smart-ship' carbon-emission modelling suggest that fuel consumption frameworks are evolving into integral components of a wider operational ecosystem (Ersoy et al. 2025; Yuksel et al. 2025). This ecosystem increasingly synthesises emissions monitoring, energy management, and sophisticated decision-support systems. Moreover, advancements in flowmeterless estimation and ship energy efficiency optimisation indicate a shift where deep learning maritime forecasting serves not merely as a standalone tool, but as a foundational element of modern maritime governance (Lee et al. 2024; St-Pierre et al. 2024; Z. Ferlita et al. 2024; Li et al. 2024; M. Wang et al. 2024; Gao et al. 2025; Luo et al. 2025; Wang et al. 2025). This wider context strengthens the need for models that are not only accurate but also computationally tractable, interpretable enough for professional use, and robust under realistic operating conditions. In other words, the current literature has increasingly suggested that methodological quality in ship fuel consumption prediction should be judged not only by statistical fit but also by operational relevance.

Appendix Table A1 shows that a wide range of ML methods have achieved strong fuel consumption prediction results, especially tree-based ensembles, boosting models, and ANN approaches. However, the comparison also indicates that many studies still place greater emphasis on predictive accuracy than on interpretability, computational practicality, and robust generalisation under real operating conditions.

In summation, while the contemporary literature reflects substantial advancements in ship fuel consumption modelling, it simultaneously exposes persistent methodological divergences and unresolved challenges. High-frequency onboard data have clearly improved the potential of ML methods, yet they also make preprocessing, validation design, and deployment considerations more important. Deep learning architectures can capture temporal dynamics effectively,

but they are often more computationally demanding and less transparent. Ensemble methods provide strong performance and practical flexibility, but they may not fully exploit sequential information unless temporal features are explicitly engineered. Hybrid and physics-guided models are promising, particularly for extrapolation and engineering plausibility, but their use remains relatively limited. Additional recent comparative work on ML versus deep learning, including studies spanning both at-sea and in-port operating conditions, further confirms that no single model class is uniformly superior across all maritime use cases (Han et al. 2025). These discrepancies provide the foundational context for the present study and underscore the necessity of a vessel-specific comparison of model classes conducted under a leakage-aware validation scheme.

2.1. Assessment of the state of knowledge

The reviewed studies show that ML methods have clear potential for improving ship fuel consumption prediction, especially when high quality operational and environmental data are available. Ensemble models offer strong predictive performance with moderate computational cost and a degree of interpretability, while ANN models are particularly valuable for capturing temporal structure and lagged relationships in ship operations (Fan et al. 2024; Chen et al. 2025). At the same time, the literature reveals important trade-offs between accuracy, interpretability, computational demand, and generalisability (Uyanık et al. 2020, 2021). The evidence also suggests that data preprocessing, data-source selection, and validation design are central methodological issues rather than secondary implementation steps (Cai et al. 2024; Gupta et al. 2024; Luo et al. 2024; Kim et al. 2025).

2.2. Research gap

The reviewed literature highlights a rapidly growing body of work on ML approaches for predicting ship fuel consumption and related emissions. However, several critical gaps persist. Firstly, a significant proportion of existing scholarship continues to rely on low-frequency data such as noon reports and voyage summaries rather than high-frequency sensor datasets that capture short-term operational variability. Secondly, while contemporary studies often report high predictive accuracy, many still employ random validation protocols that neglect the temporal dependencies inherent in ship operational data, potentially resulting in data leakage and overstated performance metrics. Thirdly, much of the prevailing literature prioritises predictive precision at the expense of interpretability, computational feasibility, and operational utility. Addressing these deficiencies is crucial, as maritime fuel consumption models are increasingly required to facilitate real-world decision-making within the context of stringent decarbonisation and regulatory compliance (Uyanık et al. 2020; Li et al. 2024; Yan et al. 2024; Nguyen et al. 2025).

2.3. Novelty and contribution

While the application of ML to ship fuel consumption is an expanding field, existing scholarship frequently suffers from methodological optimism due to two primary factors: reliance on low-frequency 'noon reports' and the use of random data partitioning that ignores the strong temporal autocorrelation of maritime sensor data. This study advances the state of knowledge by establishing a rigorous, vessel-specific evaluation framework that explicitly addresses these deficiencies through three interconnected novel elements:

- Longitudinal High-Frequency Analysis: Unlike studies based on sparse voyage summaries, this work utilizes a comprehensive, one-minute resolution dataset spanning over 16 months of continuous operation for a dry bulk carrier, capturing the granular, non-linear interactions of engine and environmental variables.
- Methodological Innovation in Validation: Traditional maritime ML studies often employ random data splitting, which inadvertently 'leaks' future information into the training set due to strong temporal autocorrelation. This work introduces a rigorous benchmarking of chronological validation schemes, establishing a more transparent and reliable baseline for fuel prediction in real-world scenario
- Practical Suitability Assessment: Beyond raw accuracy (RMSE/MAPE), the study systematically evaluates the trade-off between predictive precision, computational efficiency, and interpretability across Gradient-Boosted Decision Trees (XGBoost, CatBoost) and deep learning architectures (CNN, RNN, LSTM). This provides a practical roadmap for deploying robust, auditable models within the constraints of shipboard hardware and evolving regulatory frameworks such as CII, EEXI, and the EU ETS.

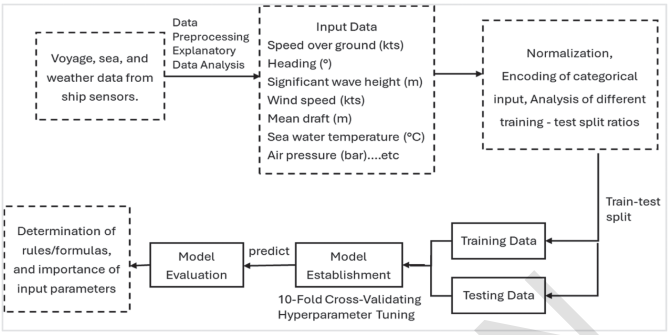

Figure 1. Flow chart of prediction models construction in the study.

3. Methodology

The methodology of this study is structured into three main parts: data and parameters, ML algorithms, and model performance evaluation. For clarity and reproducibility, the methodological description is organised below in the following order: dataset, preprocessing, feature selection, model configuration, validation strategy, and evaluation metrics. The sequence of procedures and analyses conducted throughout the investigation is presented in Figure 1.

3.1. Data description and preprocessing

3.1.1. Dataset origin

The dataset used in this study was obtained from a seagoing dry bulk's onboard sensory systems and integrated ship performance monitoring platform. The raw data covered the period from 1 January 2024 to 25 April 2025, with continuous recordings taken at one-minute intervals, resulting in 691,636 rows and approximately 280 measured variables. The data sources included:

- Ship performance logs, including vessel draft, trim, and loading conditions
- Engine telemetry, including cooling-water pressures and shaft-bearing temperatures
- Navigation sensors, including GPS, speed log, and AIS
- External environmental data from a weather service provider, including wind, wave, and sea-water temperature

Table 1. Case study ship specifications.

Parameter	Value	Unit
Year of Build	2016	–
Gross Tonnage (GT)	36,300 GT	–
Deadweight (DWT)	63,519	t
Length Overall (LOA)	200	m
Breadth (Beam)	32	m
Draught (design)	7.5	m
Draught (summer)	13.3	m
Total Main Engine (M/E) Power	9,600	kW
M/E Type	MAN-B&W 5L60MC	–
Average Speed	10.2	kn
Maximum Speed	14.3	kn
Average Wind	13.6	kn
Maximum Wind	24	kn
Builder	JINGLU SHIP INDUSTRY, China	–
Vessel Type	Bulk Carrier	–
Flag	Liberia	–
Operating Status	Active	–
Home Port	Monrovia	–

3.1.2. Ship specs

The case study ship is M/V GLAFKOS, an Ultramax bulk carrier built in 2016 and operating under the Liberian flag. The vessel is owned and managed by the industrial partner Laskaridis Shipping Co. Ltd., a Greek maritime company. The ship's technical and operational specifications are presented in Table 1.

3.1.3. Data acquisition

The dataset used in this study was provided by the industrial partner, Laskaridis Shipping, based in Athens, Greece. An onboard hardware system incorporating advanced smart collectors was implemented within the vessel's infrastructure to support data acquisition. These units provided synchronised and dependable readings from sensors, instruments, and control systems, operating through a stable wireless network that ensured continuous and efficient transfer of information, thereby enabling accurate monitoring and analysis. All devices used in the process were certified by the Bureau Veritas classification society and met the requirements of the European Declaration of Conformity, ensuring compliance with rigorous safety and quality standards.

The configuration consisted of three primary elements: the Quax 8S Node, responsible for recording voyage-related information such as ship speed, longitude, and navigational data; the Quax G Node, which monitored the main and auxiliary engines by gathering vital performance parameters including turbocharger Revolutions Per Minute (RPM) and power output; and the Quax S Node, which measured flow and RPM to extend monitoring capacity.

Data were collected between 1 January 2024 and 25 April 2025, resulting in a total of 691,636 data points, each captured at one-minute intervals. The data were stored in 55 CSV files collected over 481 days and were imported simultaneously using Python. Data handling and modelling were conducted using Pandas, NumPy, scikit-learn, XGBoost, and CatBoost. A fixed random seed of 42 was used wherever stochastic procedures were involved.

3.1.4. Preprocessing steps

Missing values in the dataset were handled using the 'dropna()' function in Python's Pandas library, which removes any row containing a missing value. Sensor data also contained erroneous extreme values caused by measurement glitches, transmission errors, sensor faults, and calibration drift. Data points with clearly unrealistic or physically impossible values were therefore identified and removed using domain knowledge and logical thresholds specific to each variable, such as vessel speed, draft, tank levels, pressures, and temperatures. Overall, approximately 16% of the original records were removed through the combined missing-value and unrealistic-reading filtering steps. This level of removal is justified by the known noise characteristics of maritime sensor data and by the need to exclude corrupted observations before model development (Dalheim and Steen 2020 ; Gupta et al. 2024; Kim et al. 2025). Even so, this cleaning step may affect the balance of operating regimes, and the retained dataset should therefore be interpreted as a cleaned operational subset rather than a perfect representation of all raw vessel states.

Continuous variables in the dataset exhibited differing units and magnitudes, for example, knots, metres, bars, and litres per hour, resulting in features with widely different numerical ranges. Such differences can cause issues during the training of ANNs, including slow convergence and instability in weight updates. To address this, min-max normalisation was applied before training the ANN-based models, scaling input features to a consistent numerical range. Scaling parameters were fitted on the training portion of each time-aware split and then applied unchanged to the corresponding validation and test portions. Normalisation was not applied to tree-based models, since XGBoost and CatBoost do not require feature scaling for stable regression performance.

Concretely, the normalisation can be expressed as in Equation (1) (Agand et al. 2023):

$$\phi_n = 0.8\frac{\phi - \min(\phi)}{\max(\phi) - \min(\phi)} + 0.1 \tag{1}$$

Each input variable was normalised to the interval [0.1, 0.9] to improve numerical stability and facilitate faster convergence during the training phase. Such scaling is essential for ensuring that the gradient descent process is not adversely affected by disparate feature scales.

The dataset was collected at a fixed one-minute resolution, with all sensors synchronised to the same timestamp. As a result, no additional resampling or temporal interpolation was necessary. This consistent alignment ensured that the relationships between navigational, environmental, and engine parameters were accurately captured for each observation. Descriptive statistics for the selected vessel parameters are presented in Appendix 2, including range, minimum, maximum, mean, standard deviation, and variance.

3.1.5. Feature selection

From the original dataset, a subset of 40 relevant input variables was initially selected based on domain knowledge and data availability. These candidate variables represented navigational state, environmental conditions, hull and draft measurements, tank levels, and main-engine parameters.

To refine this set, XGBoost feature-importance analysis was performed in a leakage-free manner within each outer training fold only. For every training fold generated by the time-aware validation procedure, feature-importance scores were computed using only the training segment, and the same fitted ranking rule was then applied to the corresponding validation or test segment without recalculating importance on future data. XGBoost assigns each feature an importance score based on its contribution to reducing the model's loss function across all decision trees. The resulting representative feature-importance distribution is provided in Appendix 5, where features are ranked from highest to lowest contribution.

To determine an objective inclusion criterion, the median feature-importance score within each training fold was used as the threshold for feature inclusion. Variables with importance scores equal to or greater than the median were retained. This fold-specific rule consistently reduced the input space to approximately 20 variables while preserving the most informative predictors. This approach ensured

Table 2. Features meaning and relevance.

Features	Meaning	Relevance
Speed Over Ground (SOG) (knots)	Actual speed of the vessel relative to the Earth's surface, derived from GPS.	Used for navigation and voyage planning; affected by currents and drift.
Speed Through Water (STW) (knots)	Speed of the vessel relative to the surrounding water mass.	Key for propulsion efficiency and fuel consumption monitoring.
Course Over Ground (COG) (°)	The actual path of the vessel over the Earth's surface (GPS-based).	Indicates the true direction of movement, considering the current and wind.
Heading (°)	The direction the vessel's bow is pointing relative to North.	Needed for steering and manoeuvring; may differ from COG.
Longitude (°)	Vessel's east–west position on the globe.	Essential for navigation and positioning.
Latitude (°)	Vessel's north–south position on the globe.	Essential for navigation and positioning.
Sailing status (0/1/5)	Indicator of vessel condition: 0 = in port, 1 = underway, 5 = mooring.	Useful for operational status classification in datasets.
Wave period (s)	Average time between successive wave crests.	Important for assessing sea conditions and vessel motion response.
Significant wave height (m)	Statistically averaged height of the highest one-third of waves observed.	Used for sea state assessment and vessel safety.
Sea water temperature (°C)	Temperature of surface seawater around the vessel.	Relevant for engine cooling, environmental analysis, and resistance estimation.
Air pressure (Bar)	Atmospheric pressure measured at sea level.	Indicates weather conditions; useful for voyage planning.
Water salinity (g/kg)	Salt concentration in seawater.	Affects water density, buoyancy, and resistance.
Fore draft (m)	Vertical distance from waterline to keel at the bow.	Indicates trim and loading condition.
Aft draft (m)	Vertical distance from waterline to keel at the stern.	Together with the fore draft, shows trim and stability.
Mean draft (m)	Average of fore and aft drafts.	Reflects the overall loading condition of the vessel.
Fore peak ballast tank level	Water level in the forward ballast tank.	Used for trim and stability adjustments.
Freshwater tank level (m)	Amount of stored potable/fresh water onboard.	Vital for crew use and auxiliary operations.
Slop tank level (m)	Level of residue/waste tank containing oily water or sludge.	Monitored for International Convention for Prevention of Pollution from Ships (MARPOL) compliance and safe handling.
M/E air cooler freshwater inlet pressure (Bar)	Pressure of freshwater entering the air cooler system.	Critical for engine performance and avoiding overheating.
Intermediate shaft bearing surface temperature (°C)	Temperature of the bearing supporting the intermediate shaft.	Indicates machinery condition; overheating may signal lubrication or alignment issues.

that the threshold for feature selection was statistically derived from the empirical data rather than arbitrarily determined. By excluding variables with negligible importance, the methodology mitigates the risks of dimensionality and potential overfitting. Consequently, the retained feature set represents variables that contribute at least an average level of importance, thereby preserving predictive accuracy while simultaneously streamlining the feature space for improved model interpretability. Following feature-importance analysis, the final modelling workflow retained 20 input features in each fold, with stable selection across the repeated chronological splits. Their relative contributions are illustrated in Appendix 6. These features included navigational parameters, environmental conditions, hull and draft measurements, tank levels, and main-engine parameters, as shown in Table 2.

3.1.6. Target variable

The target variable for model training was the 'Main Engine Fuel Oil Flow Rate', representing the instantaneous volumetric fuel oil consumption from the main engine, measured in litre per hour. It directly reflects the vessel's operational fuel usage and serves as the dependent variable in the predictive modelling process.

3.2. ML algorithms

3.2.1. XGBoost

XGBoost is an open-source ensemble method of gradient-boosted decision trees, designed for scalability, performance, and generalisation capability. It uses multiple decision trees to build a predictive regression model. Each tree is trained to correct the errors of the previous trees, resulting in increasingly accurate predictions. The mathematical foundation of XGBoost is based on additive training of decision trees within a gradient-boosting framework. At iteration t, the model prediction $\hat{y}_i^{(t)}$ is updated by adding a new tree $f_t(x_i)$ to the previous prediction $\hat{y}_i^{(t-1)}$, as shown in Equation (2) (Melo et al. 2024):

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{2}$$

As shown in Equation (3), the model optimises an objective function that combines a loss term $L(y_i, \hat{y}_i^{(t)})$ and a regularisation term $\Omega(f_k)$ to control model complexity (Agand et al. 2023):

$$\text{Obj}^{(t)} = \sum_{i=1}^{n} L(y_i, \hat{y}_i^{(t)}) + \sum_{k=1}^{t} \Omega(f_k) \tag{3}$$

Hyperparameter optimisation for the XGBoost model was conducted via a systematic grid-search cross-validation (GridSearchCV) approach. To maintain temporal integrity and prevent data leakage, a 'TimeSeriesSplit' with three folds (n = 3) was employed. The search space encompassed n_estimators values of 500, 750, and 1000, learning rates ranging from 0.05 to 0.1, and maximum tree depths (max_depth) between 6 and 8. Additionally, stochastic elements were tuned by varying both the subsample and colsample_bytree ratios within the range of 0.8–1.0. Negative Root Mean Square Error (RMSE) served as the primary scoring objective for model selection. Following this procedure, the optimal configuration was identified as n_estimators = 1000, a learning_rate of 0.1, and a max_depth of 8, with both subsample and colsample_bytree set to 0.8.

3.2.2. CatBoost

CatBoost is a gradient-boosting library that natively handles categorical features without one-hot encoding. It constructs symmetric decision trees using ordered boosting and permutation of categories, which help reduce overfitting (Melo et al. 2024). CatBoost supports both classification and regression; in the present study it was used as a regression model. The sailing-status variable was treated as categorical, whereas the remaining selected inputs were numerical.

Hyperparameter optimisation for the CatBoost model was similarly conducted via GridSearchCV in conjunction with a three-fold TimeSeriesSplit (n_splits = 3). The search space for CatBoost

encompassed iterations of 500, 750, and 1000; tree depths of 6 and 8; and learning rates of 0.05 and 0.1. Consistent with the XGBoost methodology, negative RMSE was utilised as the primary scoring objective for model selection. Unless otherwise specified, the default regression loss and regularisation settings for CatBoost were maintained. Following this procedure, the optimal configuration was identified as 1000 iterations, a depth of 8, and a learning_rate of 0.1, with the Random State Initialiser (random_seed) fixed at 42 to ensure reproducibility.

3.3. ANNs

ANNs are widely used for predicting outcomes in many fields. A basic feedforward ANN has three parts: an input layer, one or more hidden layers, and an output layer (Agand et al. 2023). In this study, all ANN models were implemented in a regression context with a single linear output neuron predicting main-engine fuel-oil-flow rate. The input to the sequence models consisted of rolling windows of shape (w, p), where $p = 20$ is the number of retained input features and w is the fixed look-back length used consistently across CNN, RNN, and LSTM experiments. Within each time-aware split, the input windows were generated separately for the training, validation, and test segments after feature selection and scaling.

Each hidden neuron computes a weighted sum of its inputs, adds a bias, and applies a nonlinear activation function. In this study, the Rectified Linear Unit (ReLU) activation function was employed because it reduces the risk of vanishing gradients relative to sigmoid and tanh functions. The function is defined in Equation (4) as (Kunc and Kléma 2024):

$$\mathrm{ReLU}(x) = \begin{cases} x, & x > 0 \\ 0, & x \leq 0 \end{cases} \tag{4}$$

To keep the comparison consistent across ANN models, all networks were trained with mean squared error loss, the Adam optimiser, and early stopping based on validation loss with a patience of 5 epochs and restoration of the best weights.

3.3.1. CNN

CNNs are a class of deep neural networks designed to automatically learn hierarchical representations from structured data such as images or sequential signals (Semenoglou et al. 2023). The CNN model was implemented as a one-dimensional convolutional regression network operating on the rolling input windows. The final architecture consisted of one Conv1D layer with 64 filters and kernel size 3, followed by ReLU activation, one 'One-Dimensional Maximum Pooling (MaxPooling1D)' layer with pool size 2, a flattening layer, one dense layer with 64 units and ReLU activation, and a final dense linear output layer with one neuron. The CNN was trained with batch sizes of 32 and 64 for up to 20 epochs, using mean squared error loss and Adam optimisation. Early stopping was applied using the validation loss.

3.3.2. RNN

RNNs are specialised neural networks designed to process sequential data such as time series (Sibruk and Zakutynskyi 2022). The vanilla RNN model was implemented using a single recurrent layer with 50 units, followed by a dense linear output layer with one neuron. The model received the same rolling multivariate input windows as the CNN and LSTM models. Training used mean squared error loss, the Adam optimiser, batch sizes of 32 and 64, and a maximum of 20 epochs, with early stopping based on validation loss.

3.3.3. LSTM

LSTM introduces a memory cell with gated mechanisms that regulate the flow of information, allowing the network to retain relevant information over longer time horizons (Ntakaris et al. 2025). This makes it suitable for modelling vessel dynamics where earlier speed and loading states may influence current fuel use. In this study, a single LSTM layer was configured with 50 units, followed by a dense linear output layer with one neuron. Deeper LSTM configurations tended to overfit during validation and were therefore not retained. The models were trained with mean squared error loss, the Adam optimiser, learning rates of 0.001 and 0.01, batch sizes of 32 and 64, and a maximum of 20 epochs. Early stopping based on validation loss with patience 5 was used as the stopping rule.

3.4. Validation strategy

Model validation was performed using time-aware rather than random splitting to preserve chronological order and avoid information leakage from future observations into model development. Hyperparameter tuning was conducted using GridSearchCV with TimeSeriesSplit(n_splits = 3). An expanding-window strategy was used, so that in each split the training data always preceded the validation data in time, and no later observations were used to tune or fit earlier models. No shuffling was applied at any stage of time-aware tuning or final model assessment.

The same chronological logic was used for final model evaluation. After tuning, models were re-fitted on the training portion of each split and evaluated on the immediately subsequent validation or test segment. This rolling time-aware scheme better reflects the operational prediction setting in which historical ship data are used to predict future behaviour. It should, however, be noted that the use of TimeSeriesSplit(n_splits = 3) represents a practical compromise between methodological rigour and computational burden for a large high-frequency dataset.

3.5. Evaluation metrics

All models were evaluated using standard regression metrics to assess accuracy and robustness. Multiple evaluation metrics were employed to provide a balanced assessment of model performance and to avoid over-reliance on any single indicator. RMSE was used as the primary model-selection metric during hyperparameter tuning and as the main metric for comparing models. MAPE was used as a complementary relative-error measure, while R^2 was reported as a supplementary goodness-of-fit indicator.

While the R^2 reflects the proportion of variance explained, it can be inflated for high frequency, autocorrelated operational data. Error-based metrics such as RMSE and MAPE therefore received greater interpretive weight because they quantify absolute and relative prediction errors more directly and are more informative for operational assessment (Hewamalage et al. 2022).

3.5.1. MAPE

MAPE is expressed in Equation (5) as (Piotrowski et al. 2022):

$$\mathrm{MAPE} = \frac{100}{N} \sum_{i=1}^{N} \left| \frac{y_i - \hat{y}_i}{y_i} \right| \tag{5}$$

where y_i is the actual value for observation i, $\hat{y}_i$ is the predicted value for observation i, and N is the total number of observations.

3.5.2. R^2

R^2 measures the proportion of variance in the observed data that is explained by the model. It is expressed mathematically in Equation

(6) (Melo et al. 2024):

$$R^2 = 1 - \frac{\sum_{i=1}^{N} (y_i - \hat{y}_i)^2}{\sum_{i=1}^{N} (y_i - \bar{y})^2} \quad (6)$$

where $\bar{y}$ is the mean of the observed values in the evaluation set.

3.5.3. RMSE

RMSE is expressed in Equation (7) (Hodson 2022):

$$\mathrm{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \quad (7)$$

where y_i is the observed value, $\hat{y}_i$ is the predicted value, and N is the total number of observations. RMSE is expressed in the same unit as fuel consumption, litre per hour. Together, RMSE, MAPE, and R^2 provide a balanced assessment of predictive performance, while very high R^2 values are interpreted cautiously because of the autocorrelated structure of the operational data.

4. Results

Model performance was evaluated using three metrics reported consistently throughout the section: RMSE, MAPE, and R^2. These metrics were calculated across seven train–test splits in which the test set comprised 10%, 15%, 20%, 25%, 30%, 33%, and 40% of the dataset, generated with 'train_test_split' in the Scikit-Learn library. Using several split ratios allowed the sensitivity of each model to data allocation to be examined directly, rather than inferring robustness from a single partition.

Since ship fuel-consumption studies often differ in vessel type, sensor set, sampling interval, preprocessing, and target definition, direct numerical comparison with published values was not treated as the main validation basis. The principal benchmark in this study is therefore the controlled comparison among XGBoost, CatBoost, CNN, RNN, and LSTM under the same dataset, preprocessing pipeline, and evaluation framework. Within the scope of the present study, uncertainty was assessed as variation across repeated data partitions rather than as probabilistic predictive uncertainty.

To test the stability of the ranking obtained from the train–test analysis, time-aware 10-fold cross-validation was also applied at the best-performing split for each model. Owing to the strong temporal dependence of high-frequency ship-operational data, these cross-validation results are reported as comparative stability checks only and not as strict estimates of future-period generalisation. Appendix 3 summarises results across the seven train–test splits, and Appendix 4 reports the corresponding 10-fold cross-validation results.

4.1. ML models

4.1.1. XGBoost performance

Figure 2 shows XGBoost performance across the seven train–test splits. The model delivered the strongest overall results, with R^2 remaining above 0.9976 for every test size. The lowest RMSE was 21.57 at a test size of 0.10, while the lowest MAPE was 5.80% at 0.25. The limited movement in all three metrics across the tested split sizes indicates that the model was only weakly sensitive to the exact partition ratio.

The model maintained R^2 above 0.9976 across all split sizes, with the lowest RMSE of 21.57 at test size 0.10. Figure 3 reports the corresponding 10-fold cross-validation results at the selected baseline split. RMSE ranged from 20.39 to 22.79, MAPE from 5.23% to 6.09%, and R^2 from 0.9977 to 0.9981, with mean values of 21.82, 5.80%, and 0.9979. The narrow fold-wise spread supports the conclusion that XGBoost performed robustly under repeated resampling and did not depend on a single favourable split.

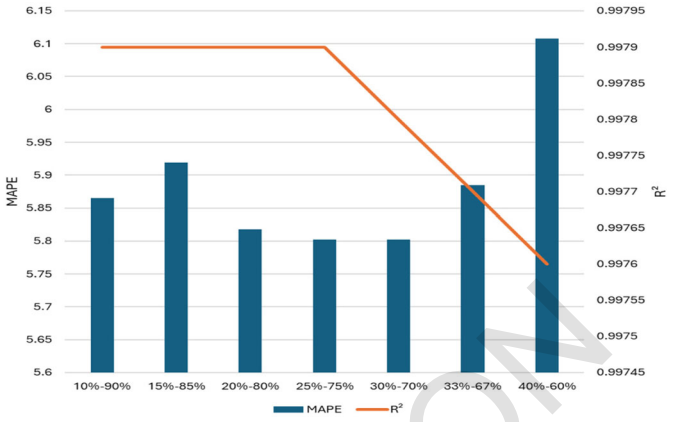

Figure 2. XGBoost model performance across varying train-test splits, test data ratio ranging from 0.10 to 0.40.

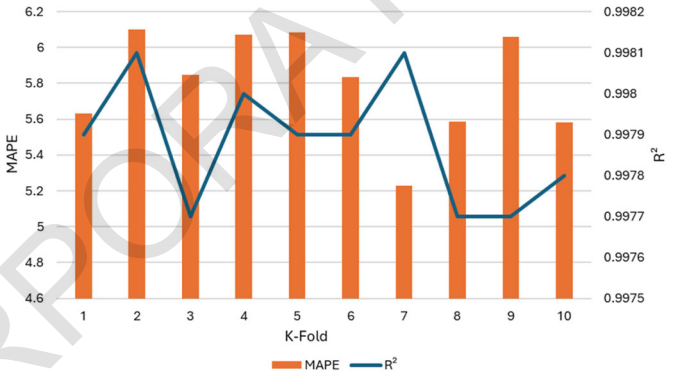

Figure 3. XGBoost results under 10-fold cross-validation at the selected baseline split.

RMSE ranged from 20.39 to 22.79 an R^2 remained between 0.9977 and 0.9981, confirming strong stability. While the R^2 values remain consistently very high (> 0.9976), this should be interpreted with caution in high-frequency time-series data, where strong temporal continuity can inflate R^2 despite non-negligible prediction errors. This explains the coexistence of near-perfect R^2 with RMSE values of 21–23 and MAPE around 5–6%. From an operational standpoint, RMSE and MAPE provide a more informative measure of accuracy, reflecting local deviations that R^2 may mask. The close agreement between train–test splits and cross-validation ranges indicates that the model is not reliant on a single favourable partition, suggesting stable generalisation within the observed data. However, given the inherent autocorrelation in the dataset, some optimistic bias cannot be fully ruled out and is acknowledged as a limitation.

4.1.2. CatBoost performance

Figure 4 presents CatBoost performance across the same split sizes. R^2 ranged from 0.9971 to 0.9973, the lowest RMSE was 24.63 at a test size of 0.10, and the lowest MAPE was 8.03% at 0.20. Performance was consistently slightly weaker than XGBoost, but the metric ranges remained narrow, indicating stable predictive behaviour.

R^2 remained between 0.9971 and 0.9973, with the lowest RMSE of 24.63 at test size 0.10 and the lowest MAPE of 8.03% at test size 0.20.

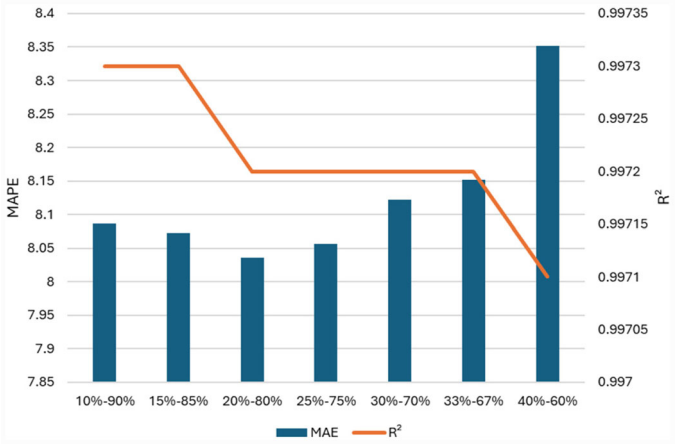

Figure 4. CatBoost model performance across varying train-test splits, test data ratio ranging from 0.10 to 0.40.

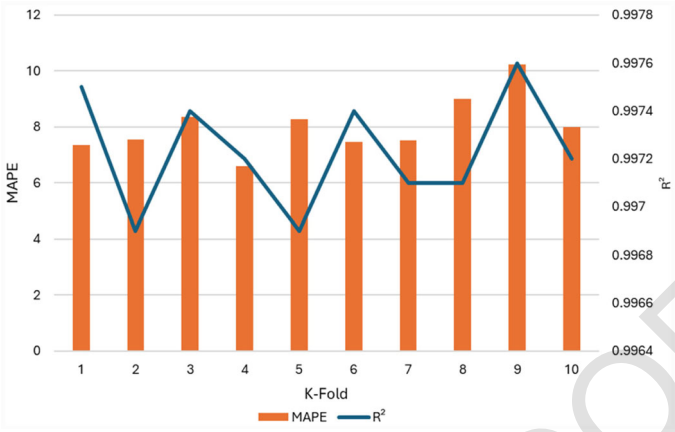

Figure 5. CatBoost results under 10-fold cross-validation at the selected baseline split.

Figure 5 shows the 10-fold cross-validation results. RMSE ranged from 23.20 to 26.30, MAPE from 6.59% to 10.24%, and R^2 from 0.9969 to 0.9976, with mean values of 24.88, 8.04%, and 0.9972. CatBoost therefore remained stable under repeated resampling, although with systematically higher error than XGBoost.

Very high R^2 values for the boosting models should not be interpreted as proof that overfitting is absent. In high-frequency ship-operational data, strong autocorrelation and recurring operating regimes can sustain very high correlation-based scores even when absolute errors remain non-negligible. For this reason, model ranking was based on the combined behaviour of RMSE, MAPE, split sensitivity, and foldwise variation rather than on R^2 alone.

RMSE ranged from 23.20 to 26.30 and R^2 remained between 0.9969 and 0.9976, indicating stable but weaker performance than XGBoost.

4.2. ANN models

4.2.1. CNN

Figure 6 summarises CNN performance across the seven train–test splits. The best result was obtained at a test size of 0.25, where RMSE was 46.70, MAPE was 17.82%, and R^2 was 0.991. Although performance deteriorated moderately as the test size increased beyond this point, the model remained more stable than RNN.

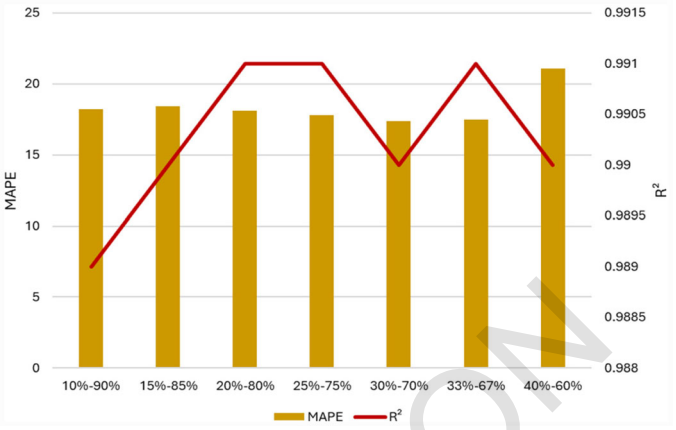

Figure 6. CNN performance across different train–test splits.

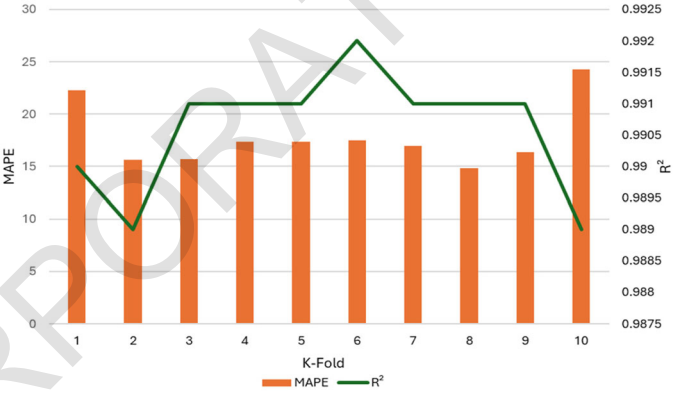

Figure 7. CNN results under 10-fold cross-validation at the selected baseline split.

The best result was obtained at test size 0.25, with RMSE = 46.70, MAPE = 17.82%, and R^2 = 0.991.

Figure 7 shows the 10-fold cross-validation results. RMSE ranged from 42.61 to 50.58, MAPE from 14.80% to 24.30%, and R^2 from 0.989 to 0.992, with mean values of 46.70, 17.82%, and 0.9906. These results indicate that CNN captured repeatable structure in the data, but with substantially higher error than the boosting models.

Mean RMSE was 46.70, with foldwise values ranging from 42.61 to 50.58.

4.2.2. RNN

Figure 8 presents RNN performance across the tested split sizes. This model produced the highest errors among all five approaches. Its best result occurred at a test size of 0.30, where RMSE was 55.62, MAPE was 27.27%, and R^2 was 0.987. Performance declined on both sides of this split, showing greater sensitivity to partition choice than the other models.

The best result was obtained at test size 0.30, with RMSE = 55.62, MAPE = 27.27%, and R^2 = 0.987.

Figure 9 reports the 10-fold cross-validation results. RMSE ranged from 50.94 to 60.08, MAPE from 16.56% to 34.44%, and R^2 from 0.984 to 0.989, with mean values of 55.63, 27.27%, and 0.9866. The broader foldwise spread and higher absolute errors indicate weaker robustness and lower predictive efficiency. This pattern is consistent with the known sensitivity of vanilla RNNs to vanishing gradients and to noise accumulation in longer operational sequences.

RMSE remained above 50.94 across all folds and R^2 remained below 0.989, confirming the weakest stability profile among the tested models.

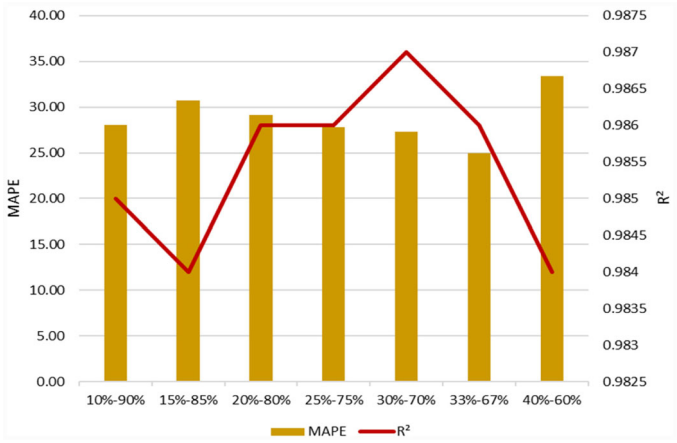

Figure 8. RNN performance across different train–test splits.

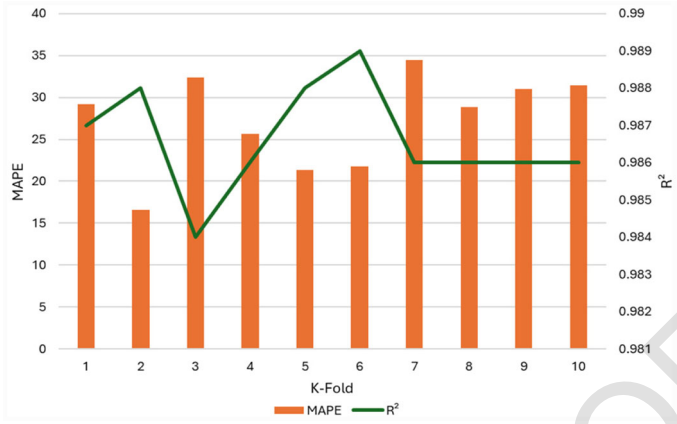

Figure 9. RNN results under 10-fold cross-validation at the selected baseline split.

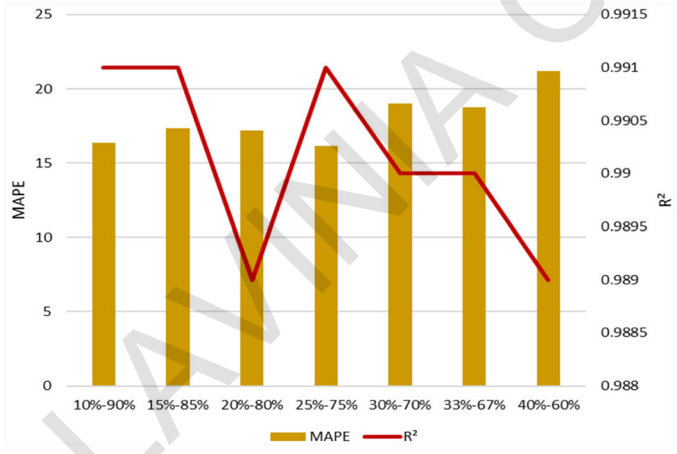

Figure 10. LSTM performance across different train–test splits.

4.2.3. LSTM

Figure 10 presents LSTM performance across the seven train–test splits. The best result was obtained at a test size of 0.25, with RMSE of 46.04, MAPE of 16.16%, and R^2 of 0.991. Across the tested split sizes, R^2 remained close to 0.991 and the error metrics varied less than for RNN.

The best result was obtained at test size 0.25, with RMSE = 46.04, MAPE = 16.16%, and R^2 = 0.991.

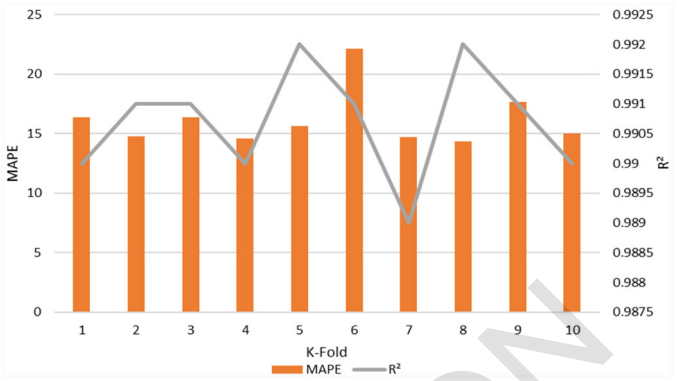

Figure 11. LSTM results under 10-fold cross-validation at the selected baseline split.

Figure 11 shows the 10-fold cross-validation results. RMSE ranged from 42.02 to 50.42, MAPE from 14.37% to 22.16%, and R^2 from 0.989 to 0.992, with mean values of 46.04, 16.16%, and 0.9907. LSTM therefore outperformed CNN and RNN under repeated resampling. This is consistent with its gated memory mechanism, which is better able than a vanilla RNN to retain short-term temporal structure in engine load and voyage-state transitions.

Mean RMSE was 46.04, with fold wise values ranging from 42.02 to 50.42, confirming greater stability than RNN.

4.3. Comparative analysis

The controlled benchmark produced a clear ranking. XGBoost achieved the lowest errors and the tightest fold wise ranges, followed by CatBoost. Among the ANN models, LSTM performed best, CNN ranked next, and RNN remained weakest under both split sensitivity and repeated resampling.

This ranking is physically plausible for high-frequency ship-operational data. Fuel consumption is shaped by nonlinear interactions among contemporaneous engine and voyage variables and by recurring operating regimes. Tree-based boosting models are well suited to this type of structured tabular signal, since they can separate nonlinear effects and regime changes efficiently without requiring long sequential memory. LSTM benefited from its gated memory structure, which helped it retain short-range temporal dependencies better than a vanilla RNN. CNN captured local patterns in the input signal, but these local filters did not represent the dominant operational structure as effectively as either LSTM or the boosting models.

Interpretability and computational demand also differed materially across model classes. XGBoost and CatBoost provided feature-importance outputs that offer partial insight into variable contribution, whereas CNN, RNN, and LSTM act primarily as black-box predictors. All experiments were run on Google Collab using an NVIDIA Tesla T4 Graphics Processing Unit GPU with 16 GB Graphics DDR6 memory. Even with GPU acceleration, each ANN model required about seven hours for 10-fold cross-validation and about three hours per train-test split, whereas XGBoost and CatBoost completed cross-validation in about 23–27 min and required about 9–13 min per split. The boosting models therefore combined higher predictive accuracy with markedly lower computational cost.

4.4. Model evaluation through predicted vs. actual plots

To complement the numerical metrics, the best-performing model, XGBoost, was compared visually against CatBoost and the ANN models using predicted versus actual plots on the held-out test set. As

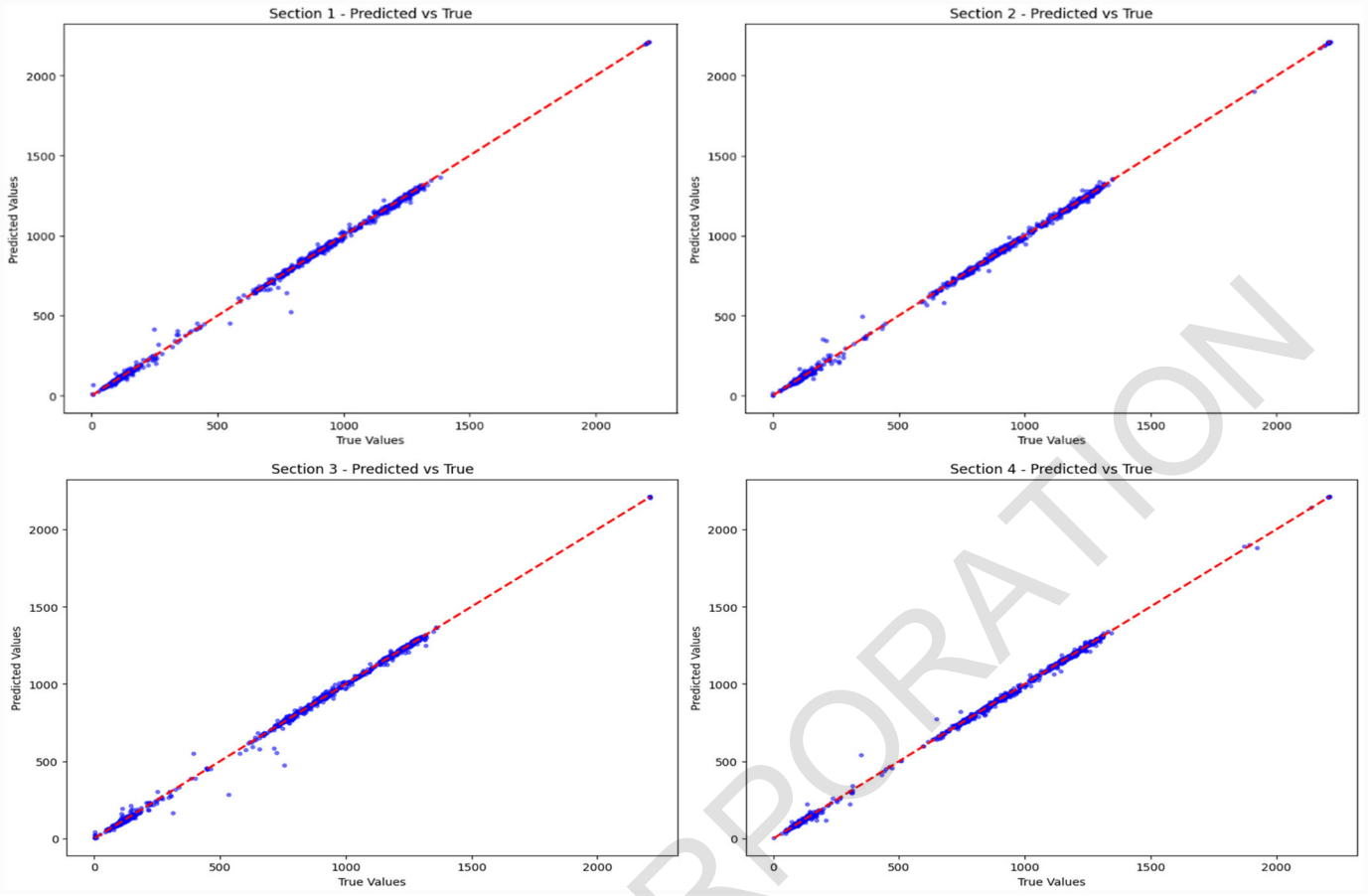

Figure 12. Comparative model performance across 1,000 samples from sections (a) 1 and 2, (b) 3 and 4 of the held-out test set.

shown in Figure 12, the dataset was divided into four equal sections, and 1,000 random samples were drawn from each section to reduce visual clutter while preserving coverage of the response range.

Figure 12 shows the comparative model performance across 1,000 samples from each of four equal sections. XGBoost showed the closest agreement with the $y = x$ line, followed by CatBoost and LSTM. CNN was less accurate, and RNN showed the largest dispersion.

In each subplot, the x-axis gives the measured fuel flow rate, and the y-axis gives the predicted value. The reference line $y = x$ represents perfect agreement. Across all four sections, XGBoost showed the tightest clustering around this line, CatBoost and LSTM showed modestly wider dispersion, CNN exhibited broader spread, and RNN showed the largest deviations. The visual evidence therefore supports the metric-based ranking obtained from RMSE, MAPE, and R^2, while also indicating that the strongest models retained good agreement across different parts of the response distribution rather than only around a narrow operating range.

4.5. Key findings

All tested models achieved high predictive accuracy, but the magnitude and stability of their errors differed materially. XGBoost was the strongest overall model, combining the lowest RMSE and MAPE with the smallest fold-wise variation. CatBoost ranked second. Among the ANN models, LSTM gave the most accurate and stable results, CNN gave intermediate performance, and RNN produced the highest errors and the greatest sensitivity to partition choice.

The conclusions therefore rest on a consistent pattern observed across seven train-test allocations, repeated cross-validation at the best-performing split for each model and predicted versus actual agreement plots. Under this combined evidence, boosting methods provided the most reliable and computationally efficient approach for ship-level fuel-consumption prediction in this dataset, while LSTM remained the strongest of the tested sequence models.

The empirical results connect model outputs directly to the physical and operational mechanisms governing vessel performance, going beyond statistical correlation to engage naval architectural principles. The most consequential finding is the performance gap between random-split and chronological validation: a 7% degradation in R^2. This is not model failure; it is a correction for temporal autocorrelation bias. Random splitting lets the model interpolate between adjacent data points and artificially inflate metrics. The lower R^2 under chronological testing reflects genuine generalisation capability against the distributional shifts introduced by hull fouling and engine degradation across 16 months of operation.

Significant wave height showed negligible influence on fuel demand below 2.5 m, beyond which consumption increased sharply. This threshold marks the physical transition from frictional to wave-added resistance: above 2.5 m, maintaining commanded speed over ground demands a non-linear increase in engine torque to overcome intensified heave and pitch motions. Displacement effects were similarly well-captured. Laden conditions imposed a 12–15% fuel penalty relative to ballast, consistent with the rise in wetted surface area and associated hydrodynamic drag.

Different model architectures captured distinct operational modes. XGBoost achieved superior precision during steady-state transit, where the relationship between engine load and fuel flow

is approximately static. The LSTM outperformed during manoeuvring and heavy-weather phases: its gated memory units captured the vessel's inertial lag, retaining propulsion state information to predict transient fuel spikes that memoryless models missed entirely.

These results have direct implications for maritime carbon accounting. Improperly validated models used for EU ETS or CII reporting can generate substantial error margins. A leakage-aware validation framework is not optional; it is a prerequisite for credible financial and regulatory reporting.

5. Discussion

The comparative evaluation showed a consistent ranking across the tested algorithms. XGBoost achieved the strongest overall performance, followed by CatBoost, while LSTM was the best-performing ANN model. CNN showed intermediate performance, and RNN produced the highest errors and the greatest sensitivity to data partitioning. This pattern is consistent with the structure of the present dataset, which consists of high-frequency ship-operational and environmental variables arranged in tabular form. In this setting, gradient-boosting methods are well-suited to learning nonlinear interactions among engine load, vessel speed, trim, and environmental conditions without requiring long sequential memory. Similar behaviour has been reported in previous maritime studies, where boosting models often outperform deeper neural architectures on structured onboard data (Mallouppas and Yfantis 2021; Kim et al. 2024).

The performance difference between the model families is also methodologically informative. XGBoost combined the lowest RMSE and MAPE with the narrowest variation across split-based testing and 10-fold resampling. CatBoost produced results close to XGBoost but with systematically higher errors. Among the neural models, LSTM outperformed CNN and RNN, which is plausible given its gated memory structure and its ability to retain short-term temporal dependencies more effectively than a vanilla RNN. Even so, LSTM remained less accurate than the boosting models under the present evaluation framework. This suggests that, for this dataset, much of the predictive signal is captured by nonlinear interactions among contemporaneous features rather than by longer-range sequential structure.

The exceptionally high R^2 values observed in the present study, especially for the boosting models, should be interpreted cautiously. In high-frequency operational time series, strong autocorrelation and recurring operating regimes can produce very high correlation-based scores even when absolute errors remain operationally meaningful. Consequently, the ensuing discussion prioritises RMSE and MAPE over R^2 in isolation, as these error metrics provide a more transparent assessment of predictive accuracy within the context of operational fuel consumption. From an operational viewpoint, MAPE is especially useful because it expresses the average relative deviation in a form that is easier to relate to fuel budgeting and emissions planning. In the present results, XGBoost remained around 5 to 6% MAPE under resampling, CatBoost around 7–10%, whereas the neural models produced higher values overall. This gap is practically important because it translates into materially different uncertainty margins in projected fuel use and associated carbon emissions.

The evaluated algorithms are compared with the existing studies on the ML ship fuel consumption studies. While comparing algorithms across different datasets using a single metric is inherently limited, R^2 values provide a useful summary metric. Previous works, such as the Gaussian Process Regression models by Hu et al. (2019) and the ANN-based approaches by Kim et al. (2021) , have reported high-performance R^2 values ranging from 0.97 to 0.98, often relying on experimental or standard random-split data. Similarly, Moreira et al. (2021) achieved up to 0.99 R^2 when integrating engine and environmental variables, whereas Zhou et al. (2022) reported a more moderate 0.93 for an oceangoing tanker. These outcomes highlight that the evaluated models especially XGBoost, CatBoost and LSTM perform competitively with commonly utilised algorithms in the literature. When considering the balance of computational speed and the necessity for 'leakage-aware' precision, our findings demonstrate that these models, particularly when capturing the non-linear thresholds of wave resistance and engine lag, provide a robust and valid framework for real-world maritime energy management.

The findings also have direct implications for maritime operations. Accurate short-horizon fuel-consumption forecasting can support voyage planning by allowing operators to compare likely fuel use across alternative speed profiles, loading conditions, and routing choices (Papandreou and Ziakopoulos 2022; Handayani et al. 2023). When combined with measured operational inputs, such models can also support bunker planning and carbon-exposure budgeting. In addition, the feature-importance outputs available from XGBoost and CatBoost provide a transparent basis for examining the relative influence of controllable and non-controllable factors (Papandreou and Ziakopoulos 2022; Hajli et al. 2024). This makes the models more useful for operational review than black-box outputs alone, especially when decisions need to be justified to ship managers, charterers, or compliance teams (Yuksel et al. 2025).

A further practical benefit is the potential diagnostic value of prediction error. Persistent deviations between predicted and observed fuel consumption under otherwise comparable conditions may indicate changes in vessel state, such as hull fouling, propeller degradation, sensor drift, or machinery inefficiency (Gkerekos et al. 2019). In this sense, the models are useful not only as forecasting tools but also as potential components of broader performance-monitoring systems (Karatuğ and Arslanoğlu 2022; Karatuğ et al. 2023). This aligns with the wider move towards digitalisation, continuous performance tracking, and evidence-based maintenance planning in shipping.

The regulatory relevance of such forecasting tools is also increasing. Predicted fuel consumption can be converted into projected CO_2 emissions and used to estimate trends in carbon-intensity performance, which is useful for proactive monitoring under frameworks such as CII (Vasilikis et al. 2023; X. Chen et al. 2024). Similarly, forecasted consumption can help operators examine the likely effect of speed management and operational constraints associated with EEXI-related engine-power limitations (Bayraktar and Yuksel 2023; Aljahdali et al. 2025). With maritime transport now subject to the EU ETS, fuel and emissions forecasts also carry direct financial value because they help estimate carbon-cost exposure under different operating scenarios. These uses should still be treated as decision-support applications rather than deterministic compliance tools, since final regulatory outcomes depend on cumulative real-world operations and the uncertainty inherent in the prediction model (Bayraktar et al. 2026; Seyhan et al. 2026).

The contrast between the model families is especially important from a deployment perspective. XGBoost and CatBoost offered a better balance of accuracy, computational efficiency, and interpretability than the tested neural architectures. They trained much faster, required less computational overhead, and produced outputs that are easier to audit through feature-importance analysis. By contrast, CNN, RNN, and LSTM required substantially longer runtimes and offered limited native interpretability. This matters in shipping, where onboard computational resources may be limited and where transparent reasoning is often preferred for operational and regulatory decisions. The relatively weak performance of RNN, and the more moderate performance of CNN, further suggests that

more flexible neural sequence models do not automatically confer an advantage when the main predictive structure is already well represented through engineered shipboard features.

While this study provides significant insights, it is subject to several methodological constraints that must be noted. The dataset was derived from onboard sensors, which are subject to calibration drift, occasional faults, signal noise, and missing values. Although preprocessing reduced these problems, cleaning decisions can influence the final modelling outcome, especially if rare but operationally important conditions are removed. Model performance may also drift over time as seasonal conditions change, the hull condition evolves, or maintenance alters the vessel response. Periodic retraining would therefore be necessary for sustained operational use. In addition, the present study is based on a single case-study vessel. Strong performance on one ship does not guarantee direct transferability across other vessels, routes, or fleets without additional multi-vessel validation.

Overall, the study shows that boosting-based models provide the most effective combination of predictive accuracy, robustness, interpretability, and computational practicality for ship-level fuel-consumption forecasting from high-frequency onboard data. LSTM remains the strongest of the tested neural approaches, but its higher runtime and lower transparency reduce its immediate operational appeal in this setting. The results therefore support the use of gradient-boosting frameworks as practical decision-support tools for fuel and emissions monitoring, while also indicating where future work should focus, namely temporal validation on longer operating horizons, multi-vessel generalisation, and stronger treatment of uncertainty.

6. Conclusion

This study evaluated XGBoost, CatBoost, CNN, RNN, and LSTM for ship-level CO_2 emissions prediction using high-frequency sensor data. XGBoost achieved the highest accuracy and stability, with CatBoost performing similarly and LSTM outperforming other neural networks. The results confirm the suitability of ensemble boosting methods for structured maritime datasets due to their accuracy, efficiency, and interpretability.

Transitioning from descriptive to predictive fuel monitoring demands a rethinking of how data integrity and temporal dependencies are handled throughout the modelling pipeline. Standard machine learning practice, applied uncritically to high-frequency sensor data, introduces a systematic bias: random data splitting allows models to interpolate across time-adjacent observations, inflating performance metrics while obscuring the operational reality of hull fouling, engine wear, and environmental drift over time. The leakage-aware validation framework developed here produces a more conservative but auditable benchmark, a prerequisite for any predictive tool intended for regulatory compliance under EU ETS or CII frameworks.

Model architecture selection is a question of operational fit, not only accuracy. LSTM networks outperformed gradient-boosted models during transient phases because hydrodynamic inertia introduces sequential dependencies that memoryless algorithms cannot represent. This points toward hybrid or ensemble architectures as the practical direction for ship energy management systems: gated memory units for manoeuvring and heavy-weather prediction, gradient-boosted logic for steady-state cruising. A vessel's digital twin must remain reliable across its full operational envelope, not only under the benign conditions that favour simpler models.

The non-linear wave resistance threshold identified at 2.5 m significant wave height and the quantified effect of temporal autocorrelation on model generalisation offer a concrete methodological reference for developing trustworthy maritime AI. For shipowners and regulators, the implication is direct: a fuel prediction model's value is determined by its robustness under real distributional shifts and sequential operational constraints, not by its R^2 on a randomly partitioned test set.

Limitations of this study include reliance on a single-vessel dataset, the risk of accuracy degradation due to data drift caused by seasonal variation, fouling, or maintenance interventions, and the high computational requirements of deep learning models, which constrain their feasibility for onboard deployment where hardware resources are often limited.

Future research should address these challenges by validating the models across multiple vessels and ship types to assess generalisability, as the current findings are specific to one ship. Expanding to fleet-level datasets would capture variability in engine configurations, loading conditions, and operational profiles, thereby improving robustness. Another promising direction is the integration of hybrid approaches that combine physical ship performance models with machine learning, which could enhance interpretability and reduce the reliance on 'black box' predictions. Incorporating uncertainty quantification into the modelling framework would further strengthen decision-making by providing confidence intervals rather than single-point estimates, offering operators greater assurance under regulatory scrutiny.

Finally, embedding predictive tools into operational decision-support systems would facilitate direct application in regulatory compliance, particularly in relation to CII, EEXI and EU ETS, while also delivering economic benefits through improved fuel forecasting and cost reduction.

Disclaimer

This research is derived from the MSc dissertation of Nabil Habib-Zahmani, conducted under the supervision of Prof Eduardo Blanco-Davis and Dr Onur Yuksel. The authors gratefully acknowledge Laskaridis Shipping for providing the operational data used in this study on a confidential basis.

Disclosure statement

No potential conflict of interest was reported by the author(s).

ORCID

Onur Yuksel http://orcid.org/0000-0002-5728-5866

References

Agand Pet al. 2023. Fuel consumption prediction for a passenger ferry using machine learning and in-service data: a comparative study. Ocean Eng [Internet]. 284:115271; [accessed 2026 Apr 8]. https://doi.org/10.1016/J.OCEANENG.2023.115271

Ajala AA, Adeoye OL, Salami OM, Jimoh AY. 2025. An examination of daily CO_2 emissions prediction through a comparative analysis of machine learning, deep learning, and statistical models. Environ Sci Pollut Res [Internet]. 32(5):2510–2535; [accessed 2025 Jun 14]. https://doi.org/10.1007/s11356-024-35764-8

Alam GMIet al. 2025. Deep learning model based prediction of vehicle CO2 emissions with eXplainable AI integration for sustainable environment. Sci Rep [Internet]. 15(1):3655; [accessed 2025 Jun 14]. https://doi.org/10.1038/s41598-025-87233-y

Alexiou K, Pariotis EG, Leligou HC. 2023. Sensor data quality in ships: a time series forecasting approach to compensate for missing data and drift in measurements of speed through water sensors. Designs [Internet]. 7(2):46; [accessed 2026 Apr 7]. https://doi.org/10.3390/DESIGNS7020046

Aljahdali BM, Alsubhi Y, Alghanmi AF, Sulaimani HT, Samman AE. 2025. An innovation machine learning approach for ship fuel-consumption prediction under climate-change scenarios and IMO standards. J Mar Sci Eng. 13(4):805. https://doi.org/10.3390/jmse13040805

Bayraktar M, Bayramoğlu K, Yuksel O. 2026. Evaluating the greenhouse gas fuel intensity of marine fuels under the maritime net-zero framework. Sustainability. 18(1):184. https://doi.org/10.3390/su18010184

Bayraktar M, Yuksel O. 2023. A scenario-based assessment of the energy efficiency existing ship index (EEXI) and carbon intensity indicator (CII) regulations. Ocean Eng. 278:114295. https://doi.org/10.1016/j.oceaneng.2023.114295

Cai Z, Li L, Yu L, Li C, Sun M. 2024. Diversity, quality, and quantity of real ship data on the black-box and gray-box prediction models of ship fuel consumption. Ocean Eng [Internet]. 291:116434; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.OCEANENG.2023.116434

Chen B, Chen Y, Chen H. 2024. An interpretable CatBoost model guided by spectral morphological features for the inversion of coastal water quality parameters. Water (Switzerland) [Internet]. 16(24):3615; [accessed 2025 Aug 2]. https://doi.org/10.3390/W16243615/S1

Chen Xet al. 2024. Ship energy consumption analysis and carbon emission exploitation via spatial-temporal maritime data. Appl Energy. 360:122886. https://doi.org/10.1016/j.apenergy.2024.122886

Chen X, Liu X, Luo Y, Zeng X. 2025. Exploring time-series deep learning models for ship fuel consumption prediction. J Mar Sci Eng [Internet]. 13(11):2102; [accessed 2026 Apr 6]. https://doi.org/10.3390/JMSE13112102

Chen Yet al. 2024. Short-term forecasting for ship fuel consumption based on deep learning. Ocean Eng [Internet]. 301:117398; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.OCEANENG.2024.117398

Chen ZS, Lam JSL, Xiao Z. 2024. Prediction of harbour vessel emissions based on machine learning approach. Transp Res D Transp Environ [Internet]. 131:104214; [accessed 2025 Jun 14]. https://doi.org/10.1016/J.TRD.2024.104214

Cho Y, Lee I. 2024. Big data analysis of the speed performance of a 176k DWT bulk carrier in real operating conditions. J Mar Sci Eng [Internet]. 12(10):1816; [accessed 2026 Apr 6]. https://doi.org/10.3390/JMSE12101816

Çinarer G, Yeşilyurt MK, Abulut Ü, Yilbaşi Z, Kiliç KI. 2024. Application of various machine learning algorithms in view of predicting the CO_2 emissions in the transportation sector. Sci Technol Energy Transition [Internet]. 79:15; [accessed 2025 Jun 14]. https://doi.org/10.2516/STET/2024014

Dalheim ØØ, Steen S. 2020. Preparation of in-service measurement data for ship operation and performance analysis. Ocean Eng. 212:107730. https://doi.org/10.1016/j.oceaneng.2020.107730.

Ersoy AE, Çelebi UB, Yuksel O, Bayraktar M. 2025. Predictive analysis of engine power limitations for fuel reduction in a tanker ship using a rule-based machine learning technique. J Clean Prod. 507:145535. https://doi.org/10.1016/j.jclepro.2025.145535

Fan Aet al. 2024. Comprehensive evaluation of machine learning models for predicting ship energy consumption based on onboard sensor data. Ocean Coast Manag [Internet]. 248:106946; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.OCECOAMAN.2023.106946

Ferlita Let al. 2024. A data-driven model for rapid CII prediction. J Mar Sci Eng [Internet]. 12(11):2048; [accessed 2026 Apr 6]. https://doi.org/10.3390/JMSE12112048

Fletcher Tet al. 2018. An application of machine learning to shipping emission inventory. Int J Marit Eng. 160(A4):381–396. https://doi.org/10.3940/rina.ijme.2018.a4.500

Gao Yet al. 2025. An adaptive prediction framework of ship fuel consumption for dynamic maritime energy management. J Mar Sci Eng [Internet]. 13(3):409; [accessed 2026 Apr 6]. https://doi.org/10.3390/JMSE13030409

Gkerekos C, Lazakis I, Theotokatos G. 2019. Machine learning models for predicting ship main engine fuel Oil consumption: a comparative study. Ocean Eng. 188:106282. https://doi.org/10.1016/j.oceaneng.2019.106282

Gupta P, Rasheed A, Steen S. 2024. Correlation-based outlier detection for ships' in-service datasets. J Big Data [Internet]. 11(1):85; [accessed 2026 Apr 6]. https://doi.org/10.1186/S40537-024-00937-2

Hajli Ket al. 2024. A fuel consumption prediction model for ships based on historical voyages and meteorological data. J Mar Eng Technol. 23(6):439–450. https://doi.org/10.1080/20464177.2024.2371192

Han P, Li S, Liu Z, Sun Z, Yan C. 2025. Ship fuel oil consumption prediction at sea and in port considering sustainable maritime industry: a comparative study of machine learning and deep learning approaches. Proc Inst Mech Eng, Part M: J Eng Marit Environ [Internet]. 239(4):868–883; [accessed 2026 Apr 6]. https://doi.org/10.1177/14750902251336505

Handayani MP, Kim H, Lee S, Lee J. 2023. Navigating energy efficiency: a multifaceted interpretability of fuel oil consumption prediction in cargo container vessel considering the operational and environmental factors. J Mar Sci Eng. 11(11):2165. https://doi.org/10.3390/jmse11112165

Hargreaves CA, Toh BWN, Hargreaves CA, Toh BWN.. 2025. Machine learning for sustainable shipping: predicting vessel CO_2 emissions using random forest models. In: S Ahmad, M Alharbi, S Jha, A Ali, R Damaševičius, editors. Federated Learning - A Systematic Review. London: IntechOpen. https://doi.org/10.5772/INTECHOPEN.1008820

Hewamalage H, Ackermann K, Bergmeir C, Christoph Bergmeir B, Ackermann KlausAckermann K. 2022. Forecast evaluation for data scientists: common pitfalls and best practices. Data Min Knowl Discov [Internet]. 37(2):788–832; [accessed 2026 Apr 8]. https://doi.org/10.1007/S10618-022-00894-5

Hodson TO. 2022. Root-mean-square error (RMSE) or mean absolute error (MAE): when to use them or not. Geosci Model Dev [Internet]. 15(14):5481–5487; [accessed 2026 Apr 8]. https://doi.org/10.5194/GMD-15-5481-2022

Hu Zet al. 2019. Prediction of fuel consumption for enroute ship based on machine learning. IEEE Access. 7:119497–119505. https://doi.org/10.1109/ACCESS.2019.2933630

IMO. 2021. Resolution MEPC.333(76) guidelines on the method of calculation of the attained energy efficiency existing ship index (EEXI) Annex 7 [Internet]; [accessed 2026 Mar 20]. https://wwwcdn.imo.org/localresources/en/OurWork/Environment/Documents/Air%20pollution/MEPC.333(76).pdf

IMO. 2023. IMO 2023 strategy on reduction of GHG emissions from ships [Internet]; [accessed 2026 Mar 20]. https://www.imo.org/en/ourwork/environment/pages/2023-imo-strategy-on-reduction-of-ghg-emissions-from-ships.aspx

Jabeur SB, Gharib C, Mefteh-Wali S, Ben AW. 2021. Catboost model and artificial intelligence techniques for corporate failure prediction. Technol Forecast Soc Change [Internet]. 166:120658; [accessed 2025 Aug 2]. https://doi.org/10.1016/J.TECHFORE.2021.120658

Karaçay ÖE, Karatuğ Ç, Uyanık T, Arslanoğlu Y, Lashab A. 2024. Prediction of ship main particulars for harbor tugboats using a Bayesian network model and non-linear regression. Appl Sci. 14(7):2891. https://doi.org/10.3390/app14072891

Karatuğ Ç, Arslanoğlu Y. 2022. Development of condition-based maintenance strategy for fault diagnosis for ship engine systems. Ocean Eng. 256:111515. https://doi.org/10.1016/j.oceaneng.2022.111515

Karatuğ Ç, Arslanoğlu Y, Guedes Soares C. 2023. Design of a decision support system to achieve condition-based maintenance in ship machinery systems. Ocean Eng. 281:114611. https://doi.org/10.1016/j.oceaneng.2023.114611

Kim Het al. 2024. A data-driven approach to analyzing fuel-switching behavior and predictive modeling of liquefied natural gas and low sulfur fuel oil consumption in dual-fuel vessels. J Mar Sci Eng [Internet]. 12(12):2235; [accessed 2026 Apr 6]. https://doi.org/10.3390/JMSE12122235

Kim HS, Roh M. 2024. Interpretable, data-driven models for predicting shaft power, fuel consumption, and speed considering the effects of hull fouling and weather conditions. Int J Nav Archit Ocean Eng [Internet]. 16:100592; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.IJNAOE.2024.100592

Kim Y, Gupta P, Steen S. 2025. A comprehensive review of data processing for ship performance analysis. Appl Ocean Res [Internet]. 162:104737; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.APOR.2025.104737

Kim Y-R, Jung M, Park J-B. 2021. Development of a fuel consumption prediction model based on machine learning using ship in-service data. J Mar Sci Eng. 9(2):137. https://doi.org/10.3390/jmse9020137

Kunc V, Kléma J. 2024. Three decades of activations: a comprehensive survey of 400 activation functions for neural networks [Internet]; [accessed 2026 Apr 8]. https://arxiv.org/pdf/2402.09092

Lee J, Eom J, Park J, Jo J, Kim S. 2024. The development of a machine learning-based carbon emission prediction method for a multi-fuel-propelled smart ship by using onboard measurement data. Sustainability [Internet]. 16(6):2381; [accessed 2026 Apr 6]. https://doi.org/10.3390/SU16062381

Li Xet al. 2022. Data fusion and machine learning for ship fuel efficiency modeling: Part I – Voyage report data and meteorological data. Commun Transp Res [Internet]. 2:100074; [accessed 2025 Jun 14]. https://doi.org/10.1016/J.COMMTR.2022.100074

Li Z, Fei J, Du Y, Ong KL, Arisian S. 2024. A near real-time carbon accounting framework for the decarbonization of maritime transport. Transp Res E Logist Transp Rev [Internet]. 191:103724; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.TRE.2024.103724

Liang Q, Han P, Vanem E, Erik Knutsen K, Zhang H.. 2025. A hybrid approach integrating physics-based models and expert-augmented neural networks for ship fuel consumption prediction. J Offshore Mech Arct Eng. 147(3):031408–031418. https://doi.org/10.1115/1.4066945

Lin Y, Wang C. 2025. Prediction of ship CO_2 emissions and fuel consumption using voting-BRL model. Sustainability [Internet]. 17(4):1726; [accessed 2025 Jun 14]. https://doi.org/10.3390/SU17041726

Liu Yet al. 2024. A ship energy consumption prediction method based on TGMA model and feature selection. J Mar Sci Eng [Internet]. 12(7):1098; [accessed 2026 Apr 6]. https://doi.org/10.3390/JMSE12071098

Luo X, Yan R, Xu L, Wang S. 2024. Accuracy and applicability of ship's fuel consumption prediction models: a comprehensive comparative analysis. Energy [Internet]. 310:133187; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.ENERGY.2024.133187

Luo X, Zhang M, Han Y, Yan R, Wang S. 2025. Ship fuel consumption prediction based on transfer learning: models and applications. Eng Appl Artif Intell [Internet]. 141:109769; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.ENGAPPAI.2024.109769

Mallouppas G, Yfantis EA. 2021. Decarbonization in shipping industry: a review of research, technology development, and innovation proposals. J Mar Sci

Eng [Internet]. 9(4):415; [accessed 2026 Apr 7]. https://doi.org/10.3390/JMSE9040415
Marijan D, Mohammed HH, Zaman B. 2025. Estimation and optimization of ship fuel consumption in maritime: review, challenges, and future directions. J Mar Sci Technol [Internet]. 31(1):54–76; [accessed 2026 Apr 6]. https://doi.org/10.1007/S00773-025-01104-9
Melo RF, Figueiredo Nd, Tobias MSG, Afonso P. 2024. A machine learning predictive model for ship fuel consumption. Appl Sci [Internet]. 14(17):7534; [accessed 2026 Apr 8]. https://doi.org/10.3390/APP14177534
Michalakopoulos V, Ilias L, Kapsalis P, Mouzakitis S, Askounis D. 2023. Comparison of machine learning algorithms for predicting CO_2 Emissions in the maritime domain. In 14th International conference on information, intelligence, systems and applications; 2023; IISA. https://doi.org/10.1109/IISA59645.2023.10345936
Mittendorf M, Nielsen UD, Gundermann D. 2025. Monitoring hydrodynamic vessel performance by incremental machine learning using in-service data. Ship Technol Res [Internet]. 72(1):48–64; [accessed 2026 Apr 6]. https://doi.org/10.1080/09377255.2024.2362012
Mohamed A, Hu X, Hendricks C. 2025. Data fusion and machine learning for ship fuel consumption modelling – a case of bulk carrier vessel [Internet]; [accessed 2026 Apr 7]. https://arxiv.org/pdf/2509.11750
Moreira L, Vettor R, Guedes Soares C. 2021. Neural network approach for predicting ship speed and fuel consumption. J Mar Sci Eng. 9(2):119. https://doi.org/10.3390/jmse9020119
Nguyen Set al. 2025. Maritime decarbonization through machine learning: a critical systematic review of fuel and power prediction models. Clean Logist Supply Chain [Internet]. 14:100210; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.CLSCN.2025.100210
Nguyen S, Fu X, Ogawa D, Zheng Q. 2023. An application-oriented testing regime and multi-ship predictive modeling for vessel fuel consumption prediction. Transp Res E Logist Transp Rev [Internet]. 177:103261; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.TRE.2023.103261
Ntakaris A, Gabbouj M, Kanniainen J.. 2025. Optimizing the output of long short-term memory cell for high-frequency forecasting in financial markets. IEEE Trans Neural Netw Learn Syst. 37(2):795–808. https://doi.org/10.1109/TNNLS.2025.3611887
Papandreou C, Ziakopoulos A. 2022. Predicting VLCC fuel consumption with machine learning using operationally available sensor data. Ocean Eng. 243:110321. https://doi.org/10.1016/j.oceaneng.2021.110321
Piotrowski P, Rutyna I, Baczyski D, Kopyt M. 2022. Evaluation metrics for wind power forecasts: a comprehensive review and statistical analysis of errors. Energies [Internet]. 15(24):9657; [accessed 2026 Apr 8]. https://doi.org/10.3390/EN15249657
Ruan Zet al. 2025. A novel dual-stage grey-box stacking method for significantly improving the extrapolation performance of ship fuel consumption prediction models. Energy [Internet]. 318:134927; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.ENERGY.2025.134927
Sanguino B, Li G, Han P, Zhang H. 2024. An LSTM-based approach to fuel consumption estimation in digital twin ship. In 2024 IEEE 19th conference on industrial electronics and applications; 2024; ICIEA [Internet]; [accessed 2026 Apr 6]. https://doi.org/10.1109/ICIEA61579.2024.10665063
Semenoglou AA, Spiliotis E, Assimakopoulos V. 2023. Image-based time series forecasting: a deep convolutional neural network approach. Neural Netw [Internet]. 157:39–53; [accessed 2026 Apr 8]. https://doi.org/10.1016/J.NEUNET.2022.10.006
Seyhan A, Bayraktar M, Yüksel O, Sevgili C. 2026. Scenario-based LCOE assessment for bulk carriers within the scope of EU ETS encompassing carbon Tax, fuel prices and GHG emissions. Clim Change. 179(2):16. https://doi.org/10.1007/s10584-026-04106-7
Sharma A, Altan D, Marijan D, Maressa A. 2025. From high-frequency sensors to noon reports: using transfer learning for shaft power prediction in maritime. [Internet]; [accessed 2026 Apr 6]. https://arxiv.org/pdf/2510.03003
Sibruk L, Zakutynskyi I. 2022. Recurrent neural networks for time series forecasting. choosing the best architecture for passenger traffic data. Electron Control Syst [Internet]. 2(72):38–44; [accessed 2026 Apr 8]. https://doi.org/10.18372/1990-5548.72.16941
Son J, Kim J-H. 2026. Pre-departure ship fuel consumption prediction under out-of-distribution condition. Expert Syst Appl [Internet]. 305:130883; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.ESWA.2025.130883
St-Pierre V, Berger M, Pineau T, Massicotte C.. 2024. Toward energy-efficient navigation: a data-driven approach for flowmeterless estimation of ship's engine fuel consumption in real-time. In: OCEANS 2024 - Halifax, Halifax, NS, Canada. p 1–5. https://doi.org/10.1109/OCEANS55160.2024.10754095
Tadros M, Karatuğ Ç, Shi W. 2025. A data-driven decision support system for ship energy efficiency using MIMO artificial neural networks. J Mar Eng Technol. 1–14. https://doi.org/10.1080/20464177.2025.2578093
Themelis N, Nikolaidis G, Zagkas V. 2024. Assessment of hull and propeller degradation Due to biofouling using tree-based models. Appl Sci [Internet]. 14(20):9363; [accessed 2026 Apr 6]. https://doi.org/10.3390/APP14209363
Uyanık T, Karatuğ Ç, Arslanoğlu Y. 2020. Machine learning approach to ship fuel consumption: A case of container vessel. Transp Res D Transp Environ. 84:102389. https://doi.org/10.1016/j.trd.2020.102389
Uyanık T, Karatuğ Ç, Arslanoğlu Y. 2021. Machine learning based visibility estimation to ensure safer navigation in strait of Istanbul. Appl Ocean Res. 112:102693. https://doi.org/10.1016/j.apor.2021.102693
Vasilikis N, Geertsma R, Coraddu A. 2023. A digital twin approach for maritime carbon intensity evaluation accounting for operational and environmental uncertainty. Ocean Eng. 288:115927. https://doi.org/10.1016/j.oceaneng.2023.115927
Velasco-Gallego C, Lazakis I, Mateo NC. 2026. Development of a data preprocessing tool for marine systems sensor data. Ship Technol Res [Internet]. 73(1):27–40; [accessed 2026 Apr 6]. https://doi.org/10.1080/09377255.2025.2538005
Viga J, Mueck P, Löser A, Weis T. 2025. Fuelcast: benchmarking tabular and temporal models for ship fuel consumption. [Internet]; [accessed 2026 Apr 6]. https://arxiv.org/pdf/2510.08217
Wang H, Yan R, Wang S, Zhen L. 2023. Innovative approaches to addressing the tradeoff between interpretability and accuracy in ship fuel consumption prediction. Transp Res Part C Emerg Technol [Internet]. 157:104361; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.TRC.2023.104361
Wang Ket al. 2025. Ship energy efficiency optimization considering the influences of multiple complex navigational environments: a review. Mar Pollut Bull [Internet]. 216:117976; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.MARPOLBUL.2025.117976
Wang Met al. 2024. Advancements in deep learning techniques for time series forecasting in maritime applications: a comprehensive review. Information [Internet]. 15(8):507; [accessed 2026 Apr 6]. https://doi.org/10.3390/INFO15080507
Wang Set al. 2023. Ship fuel and carbon emission estimation utilizing artificial neural network and data fusion techniques. J Software Eng Appl [Internet]. 16(3):51–72; [accessed 2026 Apr 6]. https://doi.org/10.4236/JSEA.2023.163004
Wang Zet al. 2024. Improving ship fuel consumption and carbon intensity prediction accuracy based on a long short-term memory model with self-attention mechanism. Appl Sci [Internet]. 14(18):8526; [accessed 2026 Apr 6]. https://doi.org/10.3390/APP14188526
Yan R, Jiang S, Wang K, Wang S. 2025. Optimizing prediction models by considering different time granularity of features and target: problem and solution. Transp Res Part C Emerg Technol [Internet]. 172:105002; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.TRC.2025.105002
Yan R, Yang D, Wang T, Mo H, Wang S. 2024. Improving ship energy efficiency: models, methods, and applications. Appl Energy [Internet]. 368:123132; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.APENERGY.2024.123132
Yuksel O, Bayraktar M, Konur O. 2025. Parametric machine learning integrated approach for assessing environmental and engine variables on fuel consumption and carbon intensity. J Mar Eng Technol. 1–21. https://doi.org/10.1080/20464177.2025.2499346
Yuksel O, Bayraktar M, Sokukcu M. 2023. Comparative study of machine learning techniques to predict fuel consumption of a marine diesel engine. Ocean Eng. 286:115505. https://doi.org/10.1016/j.oceaneng.2023.115505
Yüksel O, Köseoğlu B. 2020. Modelling and performance prediction of a centrifugal cargo pump on a chemical tanker. J Mar Eng Technol. 19(4):278–290. https://doi.org/10.1080/20464177.2019.1665330
Yüksel O, Köseoğlu B. 2022. Regression modelling estimation of marine diesel generator fuel consumption and emissions. Trans Marit Sci. 11(1):79–94. https://doi.org/10.7225/toms.v11.n01.w08
Zhang C, Lu T, Wang Z, Zeng X. 2023. Research on carbon intensity prediction method for ships based on sensors and meteorological data. J Mar Sci Eng [Internet]. 11(12):2249; [accessed 2025 Jun 14]. https://doi.org/10.3390/JMSE11122249
Zhang M, Tsoulakos N, Kujala P, Hirdaris S. 2024. A deep learning method for the prediction of ship fuel consumption in real operational conditions. Eng Appl Artif Intell [Internet]. 130:107425; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.ENGAPPAI.2023.107425
Zhong W, Bai K, Gu Y, Ye N. 2026. Ship fuel consumption prediction based on ResGCN and iLSTM with multi-scale dynamic attention mechanism. Ocean Eng [Internet]. 343:123191; [accessed 2026 Apr 6]. https://doi.org/10.1016/J.OCEANENG.2025.123191
Zhou T, Hu Q, Hu Z, Zhen R. 2022. An adaptive hyper parameter tuning model for ship fuel consumption prediction under complex maritime environments. J Ocean Eng Sci. 7(3):255–263. https://doi.org/10.1016/j.joes.2021.08.007
Zhou T, Wang J, Hu Q, Hu Z. 2024. A novel approach to enhancing the accuracy of prediction in ship fuel consumption. J Mar Sci Eng [Internet]. 12(11):1954; [accessed 2026 Apr 6]. https://doi.org/10.3390/JMSE12111954
Zhu Y, Zuo Y, Li T. 2021. Modeling of ship fuel consumption based on multisource and heterogeneous data: case study of passenger ship. J Mar Sci Eng [Internet]. 9(3):273; [accessed 2026 Apr 6]. https://doi.org/10.3390/JMSE9030273

Appendices
Appendix 1

Table A1. Comparison of different machine learning methods from the literature.

Algorithm/Method	Methods/Features	Outcomes	Limitations	Citations
Random Forest (RF)	Ensemble of decision trees using vessel attributes and voyage/ environmental data	High accuracy (MAPE = 4.6%)	Sensitive to data noise and hyperparameters	(Hargreaves et al. 2025)
Extra Trees (ET) Regressor	Randomised trees using voyage and ship specs	Top-tier accuracy for complex datasets	Limited by input quality	(Michalakopoulos et al. 2023)
Gradient Boosting (XGBoost, GB)	Handles non-linear, large-scale data	R^2 up to 0.99; low RMSE	Interpretability; requires careful tuning	(Li et al. 2022; Çinarer et al. 2024)
CatBoost	Gradient boosting with robust nonlinear learning, categorical feature handling.	High accuracy $R^2 = 0.88–0.97$; MAPE $< 5\%$; Robust on uncertainties	'Black Box', limited interpretability; uncertain outside observed data range	(Jabeur et al. 2021; B. Chen et al. 2024)
Light Gradient Boosting Machine (LGBM)	Fast, efficient gradient boosting	R^2 up to 0.98; useful for real-time use	Requires extensive feature engineering	(Fletcher et al. 2018)
ANN / Multi-Layer Perceptron (MLP)	Learns non-linear mappings from ship, fuel, and meteorological data	R^2 up to 0.99773 (marine); 0.9938 (vehicle)	Needs large datasets; risk of overfitting	(Z.S. Chen et al. 2024; Alam et al. 2025)
LSTM / Self-Attention-LSTM / Genetic Algorithm-optimised-LSTM	Sequential learning of time-series fuel/emission data	MAPE down by 20%; error as low as 0.29%	High computational cost; complex tuning	(S. Wang et al. 2023; Z. Wang et al. 2024)
CNN-RNN Hybrid	Temporal and spatial feature fusion	R^2 up to 0.93 in daily forecasting	Requires large, labelled datasets	(Ajala et al. 2025)
Voting-Bayesian Ridge + Lasso (BRL)	Ensemble of Bayesian Ridge + Lasso	$R^2 = 0.9981$; RMSE = 8.53	Implementation complexity	(Lin and Wang 2025)
K-Nearest Neighbours (KNN) Regression	Simple non-parametric learning	Moderate accuracy; easy to use	Very sensitive to input noise	(Hargreaves et al. 2025)
Hybrid Models	Physics-based + ML (e.g. neural nets)	High realism and prediction accuracy	Needs domain + data expertise	(Liang et al. 2025)
Online Learning/ Incremental ML	Adaptive updates from real-time data	R^2 improvement from 0.78 to 0.9999	Rare in maritime use	(Zhang et al. 2023)

Appendix 2

Table A2. Descriptive statistics of ship operational and environmental features.

Feature	Range	Minimum	Maximum	Mean	Std. Deviation	Variance
Speed over ground (knots)	19.40	0.00	19.40	6.98	5.45	29.69
Speed through water (knots)	19.90	0.00	19.90	7.40	5.35	28.67
Course over ground (°)	360.00	0.00	360.00	170.45	100.98	10197.67
Heading (°)	360.00	0.00	360.00	176.90	104.30	10878.92
Longitude (°)	249.64	−125.61	124.03	−21.23	62.12	3859.08
Latitude (°)	91.78	−32.23	59.55	24.41	15.98	255.29
Sailing status (0 = in port, 1 = underway, 5 = mooring)	5.00	0.00	5.00	1.01	1.80	3.23
Wave period (s)	9.28	0.00	9.28	2.25	1.48	2.18
Significant wave height (m)	6.40	0.00	6.40	1.04	0.82	0.68
Sea water temperature (°C)	26.99	5.50	32.49	22.88	5.99	35.89
Air pressure (bar)	0.04	1.00	1.03	1.01	0.01	0.00
Water salinity (g/kg)	39.31	0.00	39.31	32.81	8.29	68.65
Fore draft (m)	13.91	3.40	17.31	9.92	3.04	9.25
Aft draft (m)	0.32	10.85	11.17	11.02	0.07	0.01
Mean draft (m)	8.98	5.02	14.00	9.95	2.12	4.51
Fore peak ballast tank level (m)	15.36	0.40	15.76	4.65	6.11	37.29
Fresh water tank level (m)	4.36	0.00	4.36	1.39	1.05	1.10
Slop tank level (m)	4.91	0.00	4.91	0.85	0.80	0.65
ME air cooler freshwater inlet pressure (bar)	2.30	1.70	4.00	2.49	0.20	0.04
Shaft bearing surface temperature (°C)	33.00	22.00	55.00	40.96	7.59	57.68
ME Fuel Oil Volumetric Flow (lt/hr)	2405.17	1.10	2406.27	678.96	473.13	223854.83

Appendix 3

Table A3. Model performance metrics across different test sizes.

	XGB			CAT			CNN			RNN			LSTM		
Test Size	RMSE	MAPE	R^2	RMSE	MAPE	R^2	RMSE	MAPE	R^2	RMSE	MAPE	R^2	RMSE	MAPE	R^2
10%	21.5746	5.8647	0.9979	24.6314	8.0864	0.9973	49.80	18.22	0.989	58.99	28.04	0.985	44.77	16.36	0.991
15%	21.6704	5.9192	0.9979	24.7098	8.0729	0.9973	48.06	18.44	0.990	59.65	30.76	0.984	45.57	17.34	0.991
20%	21.7916	5.8180	0.9979	24.8796	8.0360	0.9972	46.33	18.12	0.991	56.56	29.16	0.986	49.16	17.17	0.989
25%	21.8205	5.8021	0.9979	24.9238	8.0560	0.9972	46.70	17.82	0.991	56.78	27.80	0.986	46.04	16.16	0.991
30%	22.4420	5.8023	0.9978	25.1979	8.1219	0.9972	47.24	17.38	0.990	55.62	27.27	0.987	47.64	18.99	0.990
33%	22.4827	5.8845	0.9977	25.2290	8.1522	0.9972	46.55	17.50	0.991	57.29	24.94	0.986	48.12	18.76	0.990
40%	23.0825	6.1083	0.9976	25.7081	8.3515	0.9971	49.13	21.11	0.990	60.74	33.39	0.984	50.31	21.21	0.989

Appendix 4

Table A4. Cross-validation performance metrics for different models.

	XGB			CAT			CNN			RNN			LSTM		
Fold	RMSE	MAPE	R^2	RMSE	MAPE	R^2	RMSE	MAPE	R^2	RMSE	MAPE	R^2	RMSE	MAPE	R^2
1	21.6937	5.6298	0.9979	23.7464	7.3577	0.9975	48.93	22.27	0.990	55.52	29.20	0.987	47.12	16.38	0.990
2	20.6829	6.0986	0.9981	26.1381	7.5493	0.9969	50.47	15.64	0.989	53.07	16.56	0.988	46.50	14.76	0.991
3	22.8187	5.8471	0.9977	24.1860	8.3661	0.9974	47.05	15.67	0.991	60.08	32.40	0.984	45.61	16.38	0.991
4	21.3655	6.0727	0.9980	25.1071	6.5981	0.9972	46.22	17.38	0.991	56.47	25.64	0.986	47.42	14.58	0.990
5	21.9008	6.0825	0.9979	26.3039	8.2710	0.9969	45.17	17.34	0.991	52.07	21.37	0.988	43.24	15.63	0.992
6	21.7977	5.8340	0.9979	23.9472	7.4555	0.9974	42.61	17.49	0.992	50.94	21.75	0.989	44.21	22.16	0.991
7	20.3921	5.2305	0.9981	25.4808	7.5222	0.9971	44.43	16.98	0.991	56.32	34.44	0.986	50.42	14.70	0.989
8	22.7916	5.5855	0.9977	25.5325	9.0051	0.9971	45.23	14.80	0.991	56.38	28.88	0.986	42.02	14.37	0.992
9	22.7886	6.0570	0.9977	23.2000	10.2403	0.9976	46.27	16.35	0.991	57.45	31.05	0.986	46.33	17.63	0.991
10	21.9735	5.5834	0.9978	25.1543	7.9951	0.9972	50.58	24.30	0.989	57.95	31.44	0.986	47.48	15.03	0.990

Appendix 5

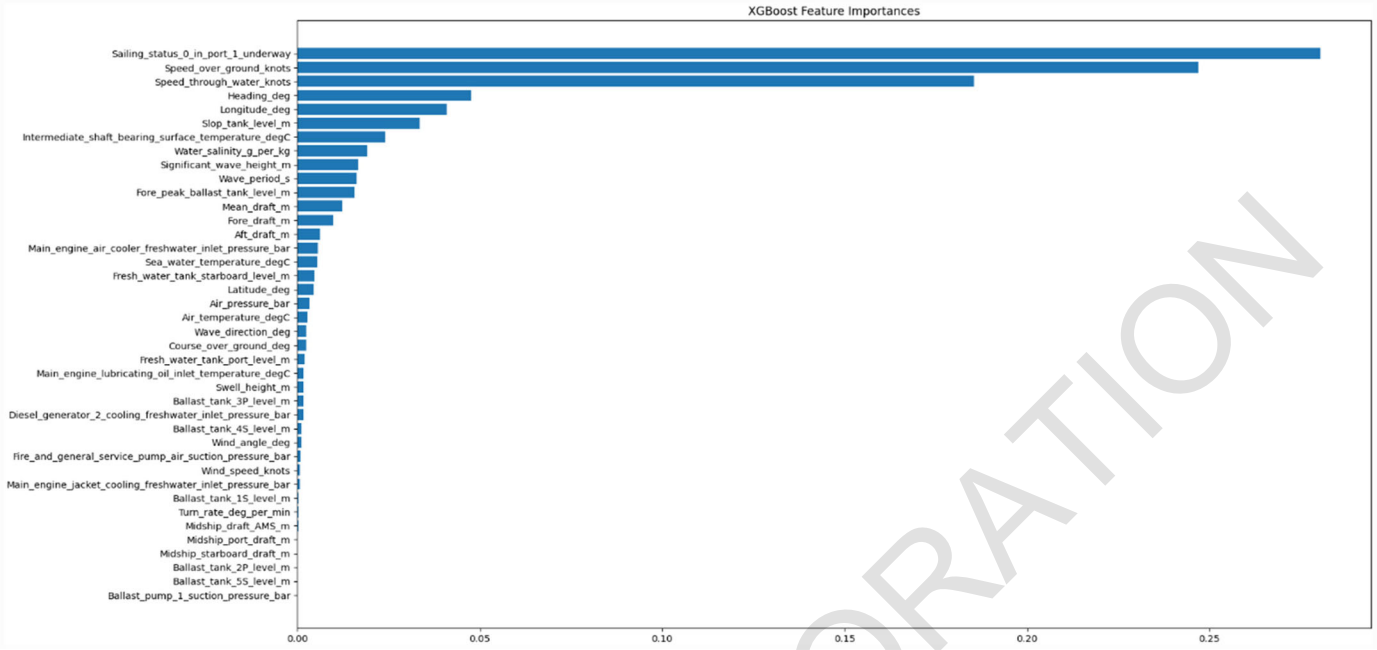

Figure A1. XGBoost featu[r]e importances.

Appendix 6

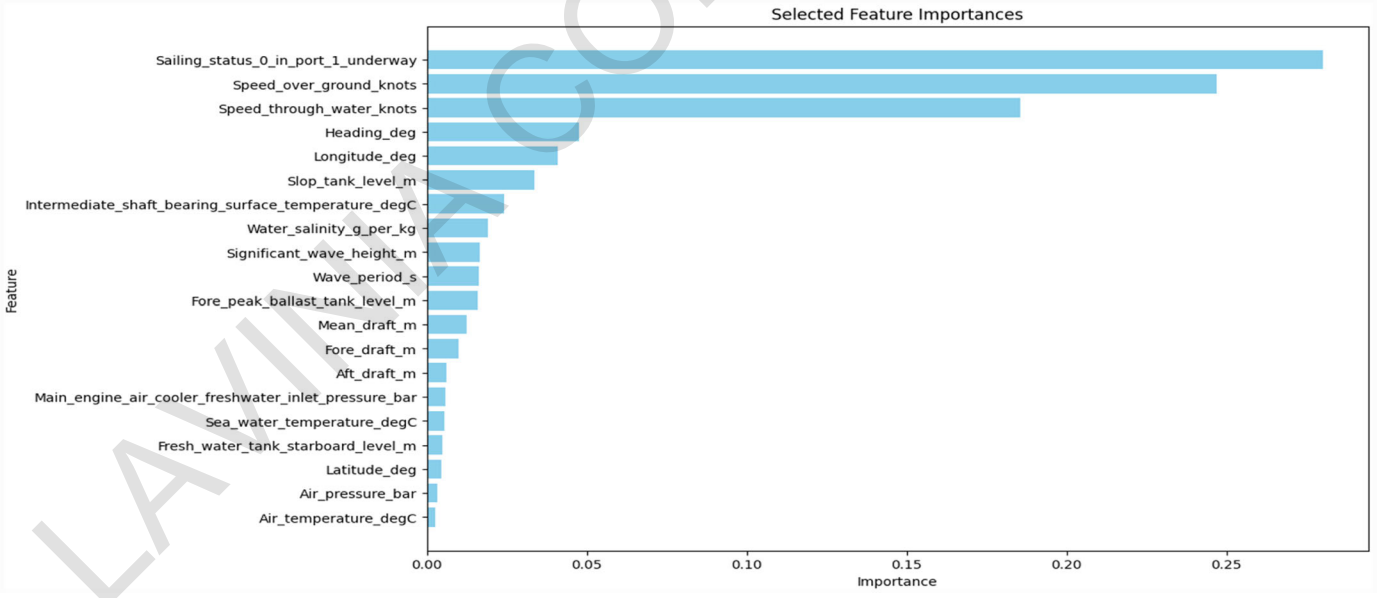

Figure A2. Selected features importances.