

Machine Learning-Based Prediction and Optimization of Ship Design Particulars for CO₂ Emission Reduction

S M Rashidul Hasan

Bangladesh Maritime University

Md. Shariful Islam

sharif.naoe@gmail.com

Bangladesh Maritime University

Zobair Ibn Awal

Bangladesh University of Engineering and Technology

Khandakar Akhter Hossain

Bangladesh Maritime University

Article

Keywords: Ship design, Machine learning, Neural Network, Random Forest, Gaussian Process Regression, Support Vector Regression, optimisation

Posted Date: August 11th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10544825/v1>

License:   This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Machine Learning-Based Prediction and Optimization of Ship Design Particulars for CO₂ Emission Reduction

S M Rashidul Hasan^{1*}, Md. Shariful Islam^{1*}, Zobair Ibn Awal², Khandakar Akhter Hossain³

¹Department of Naval Architecture and Offshore Engineering, Bangladesh Maritime University, Pallabi, Mirpur-12, Dhaka, Bangladesh

²Department of Naval Architecture and Marine Engineering, Bangladesh University of Engineering and Technology, Palashi, Dhaka, Bangladesh

³Vice-Chancellor, Bangladesh Maritime University, Mirpur-12, Dhaka, Bangladesh and Rear Admiral at Bangladesh Navy

*Corresponding Authors: sharif.naoe@gmail.com and rashidul.naoe@bmu.edu.bd

Abstract

This study develops a data-driven framework for predicting principal ship particulars and supporting low-emission preliminary design for inland and selected coastal vessels. Using datasets from cargo vessels, oil tankers, and passenger vessels, the model considers key geometric and operational variables, including length of waterline, breadth, draft, block coefficient, deadweight tonnage/gross tonnage, service speed, and engine power. Four supervised-learning methods—Artificial Neural Networks, Random Forests, Support Vector Regression, and Gaussian Process Regression—were compared using preprocessing, outlier treatment, cross-validation, and hyperparameter tuning. The validated models were then embedded in constrained MATLAB-based optimization to reduce engine power under practical design limits. Results show that engine power is the most predictable target, while block coefficient is more difficult to estimate directly. Among the tested methods, Gaussian Process Regression showed the most consistent performance overall, Support Vector Regression performed strongly for oil tankers, and Random Forests were stable for passenger vessels. A displacement-based reformulation improved the robustness and physical consistency of block-coefficient estimation.

Keywords: Ship design, Machine learning, Neural Network, Random Forest, Gaussian Process Regression, Support Vector Regression, optimisation

1 Introduction

Shipping is a significant contributor to global greenhouse-gas emissions and is increasingly targeted by international mitigation policies. The International Maritime Organization has strengthened its greenhouse-gas strategy to reduce carbon intensity and accelerate the adoption of low-emission technologies (IMO, 2023). In this context, improving ship design at the early stage is critical, as even incremental enhancements in hull form and propulsion efficiency can result in substantial lifetime emission reductions.

Traditional naval architectural methods rely on first-principles hydrodynamics, empirical correlations, and reference-ship analogies to estimate principal particulars and propulsion requirements. While these approaches are essential, they are often limited in capturing complex nonlinear interactions among hull geometry, loading conditions, and operational parameters, particularly in inland and shallow-water environments (Bertram, 2012 and Schneekluth & Bertram, 1998). These limitations restrict their effectiveness in heterogeneous vessel fleets such as those found in inland waterways.

Recent studies highlight that no single technological intervention can fully address maritime emissions; instead, incremental improvements in design and operation remain highly valuable (Bouman, et. al., 2017). Data-driven modelling provides a complementary approach by extracting patterns from historical vessel data and enabling rapid evaluation of design trade-offs. Machine learning methods are particularly suitable for predicting ship design parameters and propulsion requirements from operational inputs.

The present study builds on the authors' earlier phase of research, in which supervised machine-learning models were applied to predict principal ship particulars and support energy-efficient preliminary design (Hasan, 2025). While that earlier work established the feasibility of using data-driven models for ship-design prediction, its scope was limited in terms of vessel coverage, algorithmic diversity, and uncertainty-aware interpretation. The present phase extends that foundation by expanding the dataset to include cargo vessels, oil tankers, and passenger vessels; comparing ANN, RF, SVR, and GPR models; improving the treatment of difficult-to-predict parameters such as block coefficient; and linking the prediction framework with constrained optimisation for reduced engine-power demand and CO₂-emission mitigation.

The contributions of this paper are fourfold:

- (i) development of a harmonised inland-vessel dataset,
- (ii) comparative evaluation of ANN, RF, SVR, and GPR models,
- (iii) incorporation of uncertainty-aware and physically consistent modelling, and
- (iv) integration of predictive models into an optimisation framework for CO₂ reduction.

2 Literature Review

Recent advances in maritime digitalization have made machine learning an important tool for ship design, operation, maintenance, and optimisation. Prior studies show that data-driven models can complement conventional naval-architectural methods by learning nonlinear relationships among vessel geometry, operating conditions, and performance outcomes. Panda (Panda, 2023) reviewed machine learning applications across naval architecture, ocean engineering, and marine engineering, and identified the need for better datasets, feature selection, and optimisation strategies. Rawson and Brito (Rawson & Brito, 2023) highlighted the importance of methodological refinement for improving robustness in maritime safety assessment, while Kim et al. (Kim et al., 2021) emphasized accurate fuel-consumption prediction as a basis for operational efficiency. Gupta et al. (Gupta et al., 2022) showed that machine learning models can estimate hydrodynamic performance from onboard data and capture degradation due to fouling effects.

Several studies have demonstrated the value of machine learning in specific ship-performance problems. Sun et al. (Sun et al., 2022) developed an ANN-based model for ship resistance in ice-covered waters and reported superior performance relative to semi-empirical approaches. Tadros et al. (Tadros et al., 2020) applied machine-learning-supported strategies to marine diesel engines and showed potential for reducing fuel consumption and emissions. Mekkaoui et al. (Mekkaoui et al., 2023) used deep learning on voyage data to predict ship speed for improved traffic management. Gypa et al. (Gypa et al., 2023) combined human-computer interaction with support vector machine models in marine propeller optimisation, illustrating the utility of machine learning in design-oriented decision support.

Broader maritime-machine-learning studies also indicate that different model families serve different purposes. CNNs have been used for hull-form analysis and drag reduction, while RNNs and LSTMs support predictive maintenance and time-series forecasting (Kim et al., 2021). Ensemble methods such as random forests and gradient boosting improve predictive accuracy and interpretability for fuel, emissions, and structural-response prediction (Hassan, 2023 and Bentéjac & Martínez, 2021). Reinforcement learning has also been explored for route optimisation and adaptive hull-shape control under changing conditions (Tomljenovic et al., 2024). In addition, surrogate modelling and genetic algorithms combined with ANNs have been used to reduce resistance and optimize propeller performance (Panda, 2023, Tadros et al., 2021, 2023). Despite these advances, current studies still face issues related to scalability, generalisation, uncertainty handling, interpretability, and data quality (Freiesleben & Grote, 2023, Parisineni & Pal, 2023, Durlík, 2024). Hybrid approaches that combine machine learning with physics-based methods are increasingly recommended to improve reliability and extrapolation capability (Dotse, 2024).

2.1 Research gap

Existing machine-learning studies in maritime engineering largely target ocean-going or canonical vessel classes and typically report point predictions without principled uncertainty quantification or transparent interpretability diagnostics; as a result, transferable, uncertainty-aware, and interpretable ML frameworks for predicting principal design particulars of heterogeneous inland fleets (notably oil tankers) remain underdeveloped (Hasan et al., 2025, Durlík et al., 2024, Lundberg & Lee, 2017). This paper fills that gap by assembling a harmonised inland-vessel dataset and by systematically evaluating probabilistic, kernel and ensemble methods (including GPR and SVM/SVR), together with interpretability analyses and a prototype decision-support interface tailored to Bangladesh's inland shipping context (Hasan, 2020, Raissi et al., 2019).

3 Experimental Design and Model Development

This study was designed as a supervised learning and constrained optimisation framework for predicting principal ship design particulars and main engine power from a compact set of operational inputs. The central modelling objective was to estimate ship principal particulars using deadweight capacity, draft, and service speed as the primary predictors. The overall workflow is shown in Figure 1. In line with the report structure, the methodology proceeds through data assembly, preprocessing, exploratory analysis, outlier treatment, algorithm selection, model training, validation, and optimisation.

3.1 Research design and data structure

The foundation of the study comprises three vessel classes: cargo vessels, oil tankers, and passenger vessels. For each class, the dataset includes length of waterline (L_{WL}), breadth (B), draft (T), block coefficient (C_B),

deadweight tonnage or gross tonnage (DWT/GRT), service speed (V_s), and required engine power (EME). The descriptive characteristics of the datasets are summarised in tables 1.

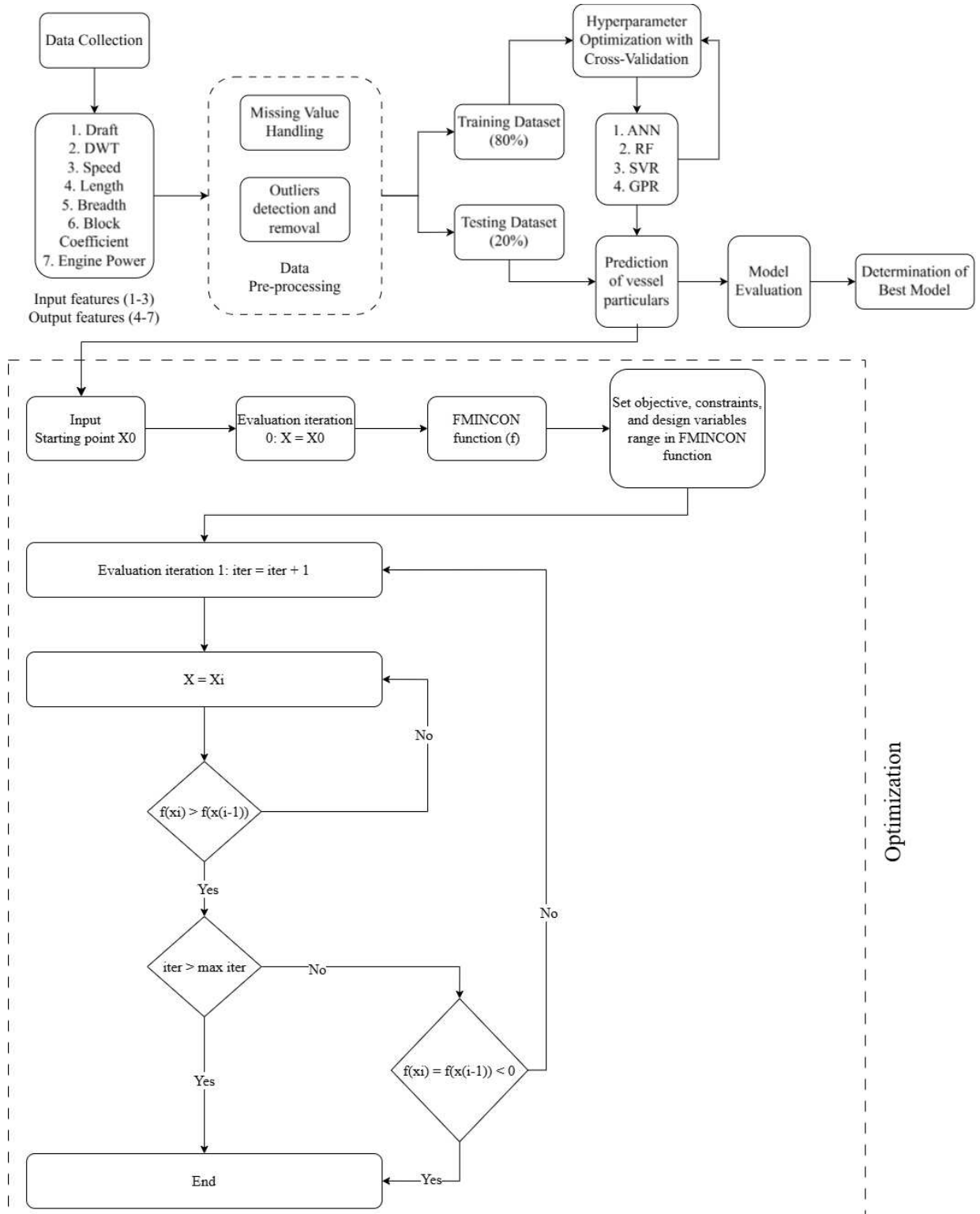


Figure 1: Methodology Flowchart

Table 1: Attributes of data set

Attribute	Unit	Cargo Vessel	Oil Tanker	Passenger Ships
-----------	------	--------------	------------	-----------------

		Min	Max	Average	Min	Max	Average	Min	Max	Average
LWL	meters	25	83.53	50.54	17.3	90.08	60.13	20	91.46	35.96
B	meters	4.88	16	9.24	11	18	12.02	3.15	15.85	7.09
T	meters	1.55	4.76	3.22	2.29	4.84	4.56	0.81	2.6	1.41
C_B	-	0.62	0.83	0.73	0.77	0.82	0.81	0.55	0.74	0.63
DWT (GT for Passenger ships)	tonnes	125	3560	1307.10	2073	2225	2130.05	25	3704.6	330.64
Vs	Knots	6.25	13.00	9.15	6	14	10.10	6	14	10.10
EME	kW	102	2364	595.88	708	783.3	726.07	49	2900	294.91

3.2 Data preprocessing and quality control

Prior to modelling, the data were examined for consistency, range plausibility, and distributional irregularities. The study emphasises that raw naval-architecture data are rarely directly usable because they may contain inconsistent records, sparse categories, and anomalous values. Data preprocessing therefore served two purposes: first, to improve model stability; and second, to ensure that the resulting predictive relationships reflected the structure of the underlying ship classes rather than artefacts of noisy observations.

Outlier detection was implemented using both the Z-score (Figure2) and interquartile range (IQR) approaches. In case of Z-score method, any point with a Z-score of 3 or more and any point with a Z-score of -3 or less is considered to be an outlier. The IQR method identifies the outliers as values that fall below $Q1 - 1.5 * IQR$ or above $Q3 + 1.5 * IQR$. The workflow and detected outliers are documented in figures 2. The corresponding summary statistics are reported in tables 2.

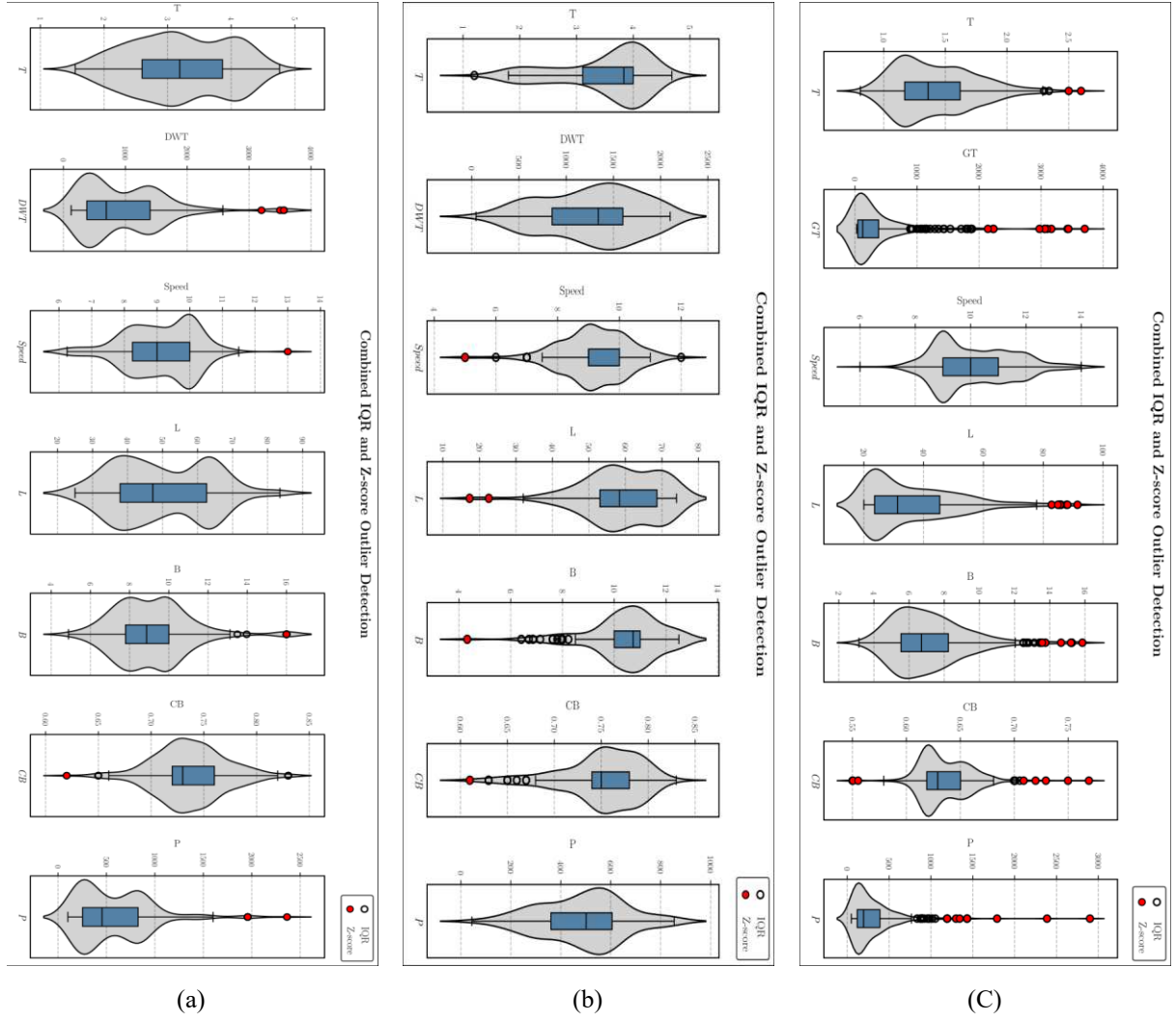


Figure 2: Outliers detection by Z-Score method for (a) Cargo Vessel (b) Oil Tanker (c) Passenger Vessel

Table 2: Summary of Outlier Detection

Attributes	Outliers (Cargo ships)	Outliers (Oil Tanker)	Outliers (Passenger Ships)
T	0 (Z)	1 (IQR)	4 (IQR)
DWT (GT for Passenger ships)	7 (IQR)	0 (Z)	37 (IQR)
Speed	1 (Z)	4 (IQR)	0 (IQR)
L	0 (IQR)	2 (IQR)	9 (IQR)
B	9 (IQR)	12 (IQR)	14 (IQR)
CB	8 (IQR)	10 (IQR)	13 (IQR)
P	9 (IQR)	0 (Z)	23 (IQR)

After removing outliers, the refined datasets contained 262 cargo vessels, 281 oil tankers, and 434 passenger vessels. This cleaning step is methodologically important because the predictive models used in the study—particularly SVR and GPR—are sensitive to atypical observations. The resulting dataset structure supports regression modelling under realistic inland-vessel operating conditions rather than artificial full-range extrapolation.

3.3 Model selection

A broad range of machine-learning methods has been applied successfully to engineering problems, including Neural Networks, Random Forests, Linear Regression, Logistic Regression, Support Vector Machines, Support

Vector Regression, Gaussian Process Regression, k-Nearest Neighbors, Naïve Bayes classifiers, and Decision Trees. Each method has different strengths depending on the complexity of the problem, the volume of data, and the degree of interaction among variables.

These models were chosen because they address the principal challenges in ship-design prediction. ANN is effective for capturing highly nonlinear relationships and complex feature interactions. RF offers robust generalization and resistance to overfitting through ensemble averaging. SVR performs well in high-dimensional feature spaces and provides controlled model capacity. GPR is particularly valuable because it delivers probabilistic predictions along with uncertainty estimates, which are important for engineering decision-making. Together, these four methods provide a balanced basis for comparing accuracy, robustness, and interpretability in ship design prediction.

3.3.1 Artificial Neural Network

Artificial Neural Networks (ANNs) were used in this study as nonlinear regression models for estimating ship design particulars and engine-power demand from selected operational and geometric inputs (Hopfield, 1982). In preliminary ship design, the relationships among deadweight or gross tonnage, draft, service speed, hull form, and engine power are rarely linear. ANN models are therefore suitable because they can learn complex input-output relationships directly from data without requiring an explicitly prescribed analytical equation (Pii, Schmidhuber, & Neural Networks, 2015).

The ANN model adopted in this work consists of an input layer, one or more hidden layers, and an output layer corresponding to the target design parameter (Hinton, 1992). The input layer receives the selected predictors, while the hidden layers transform these variables through weighted connections and activation functions. During training, the network adjusts its weights so that the difference between predicted and observed values is minimized. This structure enables the model to capture nonlinear interactions among vessel dimensions and operating conditions that may be difficult to represent using conventional regression alone.

For model training, RMSprop was used as the optimisation algorithm. RMSprop is an adaptive learning-rate method that updates the network parameters by scaling the gradient according to a moving average of recent squared gradients. This approach helps stabilize training when variables have different magnitudes or when the error surface changes rapidly. The RMSprop update process is expressed as follows:

$$v_t = \beta v_{t-1} + (1 - \beta) \nabla J(\theta)^2 \quad (1)$$

$$\theta_{t+1} = \theta_t - \frac{\eta}{\sqrt{v_t + \epsilon}} \nabla J(\theta) \quad (2)$$

where v_t denotes the moving average of squared gradients, $\nabla J(\theta)$ is the gradient of the loss function, η is the learning rate, and ϵ is a small constant introduced for numerical stability.

The Rectified Linear Unit (ReLU) activation function was used in the hidden layers to introduce nonlinearity into the model. ReLU is computationally efficient and helps reduce the vanishing-gradient problem commonly associated with deep or multilayer networks. It is defined as:

$$\text{ReLU}(x) = \max(0, x)$$

Through this architecture, the ANN model provides a flexible benchmark for evaluating nonlinear prediction performance across the vessel categories considered in this study.

3.3.2 Random Forest

Random Forest (RF) was selected as an ensemble-based regression method for predicting ship design parameters. The method constructs multiple decision trees from different bootstrap samples of the training data and combines their outputs to produce a final prediction (Salman et al., 2024). For regression tasks, the final RF estimate is obtained by averaging the predictions of the individual trees. This ensemble strategy reduces the instability often associated with a single decision tree and improves generalization when the dataset contains nonlinear patterns or mixed variable effects.

In this study, RF was useful because the vessel datasets include heterogeneous design characteristics across cargo vessels, oil tankers, and passenger vessels. Individual trees can capture local decision rules within the data, while the ensemble average reduces sensitivity to individual observations. This makes RF suitable for

preliminary design applications where the available data may include variation in vessel size, speed range, draft, and installed power.

For a training dataset containing input vectors X and corresponding target values Y , each tree is trained using a randomly sampled subset of the data. The predicted value for a new input x' is calculated as the average prediction from all B trees in the ensemble:

$$\hat{f} = \frac{1}{B} \sum_{b=1}^B f_b(x') \quad (3)$$

where $f_b(x')$ represents the prediction from the b^{th} regression tree and B is the total number of trees. In addition to prediction accuracy, RF also provides useful information on variable importance, which supports interpretation of the relative influence of parameters such as displacement, draft, speed, and vessel capacity on the predicted output.

3.3.3 Support Vector Regression (SVR)

Support Vector Regression (SVR) is an extension of the Support Vector Machine (SVM) framework designed for regression problems. Unlike traditional regression techniques that aim to minimize the prediction error for all data points, SVR focuses on finding a function that deviates from the actual target values by no more than a specified tolerance while maintaining maximum generalization capability (Drucker et al., 1996). This characteristic makes SVR particularly effective in handling nonlinear relationships and high-dimensional datasets commonly encountered in engineering applications.

The fundamental idea of SVR is to map the input data into a higher-dimensional feature space using a kernel function and then perform linear regression in that transformed space. Commonly used kernel functions include the linear, polynomial, and radial basis function (RBF) kernels. Among these, the RBF kernel is widely adopted due to its ability to model complex nonlinear patterns.

SVR introduces an ϵ -insensitive loss function, which ignores errors within a predefined margin ϵ . The optimization objective seeks to minimize model complexity while penalizing deviations outside this margin. The SVR optimization problem can be expressed as:

$$\begin{aligned} \min_{w, b, \xi, \xi^*} \quad & \frac{1}{2} |w|^2 + C \sum_{i=1}^n (\xi_i + \xi_i^*), \text{ Subjected to:} \\ & y_i - (w^T \phi(x_i) + b) \leq \epsilon + \xi_i, \\ & (w^T \phi(x_i) + b) - y_i \leq \epsilon + \xi_i^*, \\ & \xi_i, \xi_i^* \geq 0, \end{aligned}$$

where w is the weight vector, b is the bias term, $\phi(x)$ represents the kernel mapping, C is the regularization parameter controlling the trade-off between model complexity and prediction error, and ξ_i, ξ_i^* are slack variables.

SVR is particularly well suited for engineering regression problems due to its robustness against overfitting, effectiveness with limited datasets, and ability to model nonlinear input-output relationships. These characteristics make SVR a reliable choice for predicting ship performance parameters such as power demand and resistance, where data may be sparse and nonlinear behavior is dominant.

3.3.4 Gaussian Process Regression (GPR)

Gaussian Process Regression (GPR) is a probabilistic, non-parametric regression technique based on Bayesian inference. Unlike deterministic models, GPR treats regression as a problem of inferring a distribution over possible functions that fit the observed data (Rasmussen et al., 2006). This probabilistic nature allows GPR not only to provide predictions but also to quantify the uncertainty associated with those predictions, which is particularly valuable in engineering decision-making.

A Gaussian process is fully defined by its mean function $m(x)$ and covariance function $k(x, x')$, also known as the kernel function. The regression model assumes that the observed outputs are generated from a latent function sampled from a Gaussian process, expressed as:

$$f(x) \sim \mathcal{GP}(m(x), k(x, x'))$$

Where $m(x)$ is the mean function and $k(x, x')$ is the covariance (kernel) function.

Given training data (X, y) , the predictive distribution at a new input x' is Gaussian with:

$$\begin{aligned}\mu(x') &= k(x', X)[K(X, X) + \sigma_n^2 I]^{-1} y \\ \sigma^2(x') &= k(x', x') - k(x', X)[K(X, X) + \sigma_n^2 I]^{-1} k(X, x')\end{aligned}$$

where $K(X, X)$ is the covariance matrix of the training inputs, σ_n^2 represents the noise variance, and I is the identity matrix.

The choice of kernel function plays a crucial role in determining the smoothness and behavior of the regression model. Common kernels include the squared exponential (RBF), Matérn, and rational quadratic kernels. Hyperparameters associated with these kernels are typically optimized by maximizing the marginal likelihood.

GPR is especially advantageous for engineering applications where datasets are relatively small but prediction accuracy and uncertainty estimation are critical. In ship performance modeling, GPR enables accurate prediction of parameters such as engine power while providing confidence bounds, which can be directly incorporated into optimization and risk assessment frameworks.

3.4 Training, validation, and hyperparameter tuning

The model-selection strategy in this study was designed to balance predictive accuracy, robustness, and interpretability for a complex ship-design dataset. Four algorithms were selected for evaluation: Artificial Neural Networks (ANN), Random Forests (RF), Support Vector Regression (SVR), and Gaussian Process Regression (GPR). These methods were chosen because they provide complementary strengths for nonlinear engineering prediction.

ANNs are well suited to capturing highly nonlinear interactions among ship-design variables and are therefore effective for modeling ship power under varying operating conditions. RFs add robustness by reducing overfitting risk and by providing feature-importance information that improves interpretability. SVR is appropriate for limited or high-dimensional data because its kernel-based, margin-oriented formulation can generalize well. GPR extends the framework by producing probabilistic predictions with uncertainty estimates, which is valuable for engineering decision-making and optimisation.

Using these four models together reduces dependence on any single algorithm and allows comparison between deterministic and probabilistic approaches, improving confidence in predictive and optimisation outcomes.

Tables 3-6 summarize the hyperparameter configurations for the ML models developed for different vessel types. For ANN, a consistent architecture was adopted, with two hidden layers of 64 and 32 neurons using ReLU activation and a linear output layer for regression. Mean Squared Error was used as the loss function. The cargo vessel model employed the Adam optimizer, while the oil tanker and passenger vessel models used RMSProp with adjusted learning rates and batch sizes. Early stopping and dropout regularization were applied to prevent overfitting, and StandardScaler normalization with a fixed random seed ensured numerical stability and reproducibility.

Table 3: Hyperparameters of Neural Network

Hyperparameter(s)	Value		
	Cargo Vessel	Oil Tanker	Passenger Ship
Hidden Layers	[64, 32]	[64, 32]	[64, 32]
Activation Function	ReLU	ReLU	ReLU
Output Activation	Linear	Linear	Linear
Loss Function	Mean Squared Error	Mean Squared Error	Mean Squared Error
Optimizer	Adam	RMSProp	RMSProp
Learning Rate	0.001	0.001	0.0005
Batch Size	32	64	64
Epochs	100	100	100
Early Stopping	Patience = 10	Patience = 15	Patience = 15

Dropout Rate	0.2	0.2	0.3
Normalization	StandardScaler	StandardScaler	StandardScaler
Random Seed	42	42	42

Table 4: Hyperparameters of Random Forest

Hyperparameter(s)	Value		
	Cargo Vessel	Oil Tanker	Passenger Ship
n_estimators	100	150	170
max_depth	None	None	20
min_samples_split	2	5	3
min_samples_leaf	1	1	1
max_features	'auto'	'sqrt'	'sqrt'
bootstrap	True	True	True
oob_score	True	True	True
criterion	'squared_error'	'squared_error'	'squared_error'
random_state	42	42	42
n_jobs	-1	-1	-1

Table 5: Hyperparameters of SVR

Hyperparameter(s)	Value		
	Cargo Vessel	Oil Tanker	Passenger Ship
Kernel	'rbf'	'poly'	'poly'
C	10	10	15
ϵ (Epsilon)	0.1	0.25	0.25
Gamma	0.01	0.01	0.05
Degree	3	3	3
Shrinking	True	True	False
Tolerance (tol)	1e-3	1e-4	1e-4
Max Iterations	1000	1000	1000
Scaling	StandardScaler	MinMaxScaler	RobustScaler

Table 6: Hyperparameters of GPR (For All types of vessels)

Hyperparameter(s)	Value
Kernel Functions	Exponential, Matern 1/2, Rational Quadratic
Basis Function	Linear
Fit Method	Sparse Approximation
Prediction Method	Exact
Standardization	Enabled (Standardize = true)
Optimized Hyperparameters	Sigma, LengthScale
Optimization Algorithm	Gradient Descent
Input Normalization	Enabled using normalize()
Output Normalization	Enabled using normalize()

3.5 Comparative performance of ANN, RF, SVR, and GPR

Figure 3 reveals clear vessel-specific differences in model performance for cargo, oil tanker, and passenger ship datasets. Across all vessel classes, the four machine-learning models—ANN, RF, SVM, and GPR—achieve their best results for engine power prediction, confirming that propulsion power is the most predictable target because it has a strong and stable dependence on the principal ship design variables. Length is also predicted

with high reliability, with R^2 generally above 0.8 for cargo and passenger vessels and slightly lower but still acceptable for oil tankers, indicating that this parameter follows relatively clear patterns in the data.

Breadth prediction is more variable and shows that GPR and SVM usually outperform ANN and RF, especially for cargo and oil tanker vessels. This suggests that kernel-based and probabilistic methods are better at capturing subtler nonlinear geometric relationships. The most difficult variable to predict is block coefficient, where lower R^2 values appear consistently, particularly for ANN and RF. This reflects the sensitivity of block coefficient to multiple interacting design factors. Among all models, GPR performs most robustly for block coefficient because its probabilistic structure handles uncertainty and nonlinear variation more effectively.

The vessel-wise comparison further shows that cargo vessels benefit from GPR's balance of accuracy and stability, while RF remains robust under local fluctuations and ANN captures nonlinear patterns reasonably well. For oil tankers, all models predict length and engine power well, but block coefficient remains difficult, with GPR and other kernel-based methods producing smoother and more consistent fits than tree-based or neural approaches. Passenger vessels show the most balanced performance overall, with RF providing stable predictions and GPR remaining competitive, especially for length and engine power.

Overall, the comparative R^2 analysis shows three main findings: engine power is the easiest variable to predict, length is also predicted reliably, and block coefficient is the hardest. GPR emerges as the most consistently strong model across vessel types, followed closely by SVM, while ANN and RF show strengths that depend on the vessel class. These results support using multiple models rather than relying on a single universal predictor for inland ship design and optimisation.

3.6 Improved estimation of block coefficient through predicted displacement

A notable methodological result in the report is that direct prediction of block coefficient produced low coefficients of determination across all models. To address this, the study introduced a two-stage approach: displacement was predicted first, and block coefficient was then computed from the hydrostatic relation using the predicted displacement and principal dimensions. This approach improved the apparent estimation of block coefficient because displacement is physically more stable and more strongly linked to the available predictors than block coefficient itself. These results are presented in Figure 4.

This is an important technical contribution because it converts a difficult ratio-estimation problem into a direct regression problem followed by deterministic transformation. In effect, the model is no longer forced to learn block coefficient as an isolated target with weak statistical support. Instead, it learns displacement, which has stronger physical and statistical coupling with deadweight, speed, draft, and engineered geometric descriptors. The resulting block coefficient is therefore more physically consistent and less vulnerable to instability.

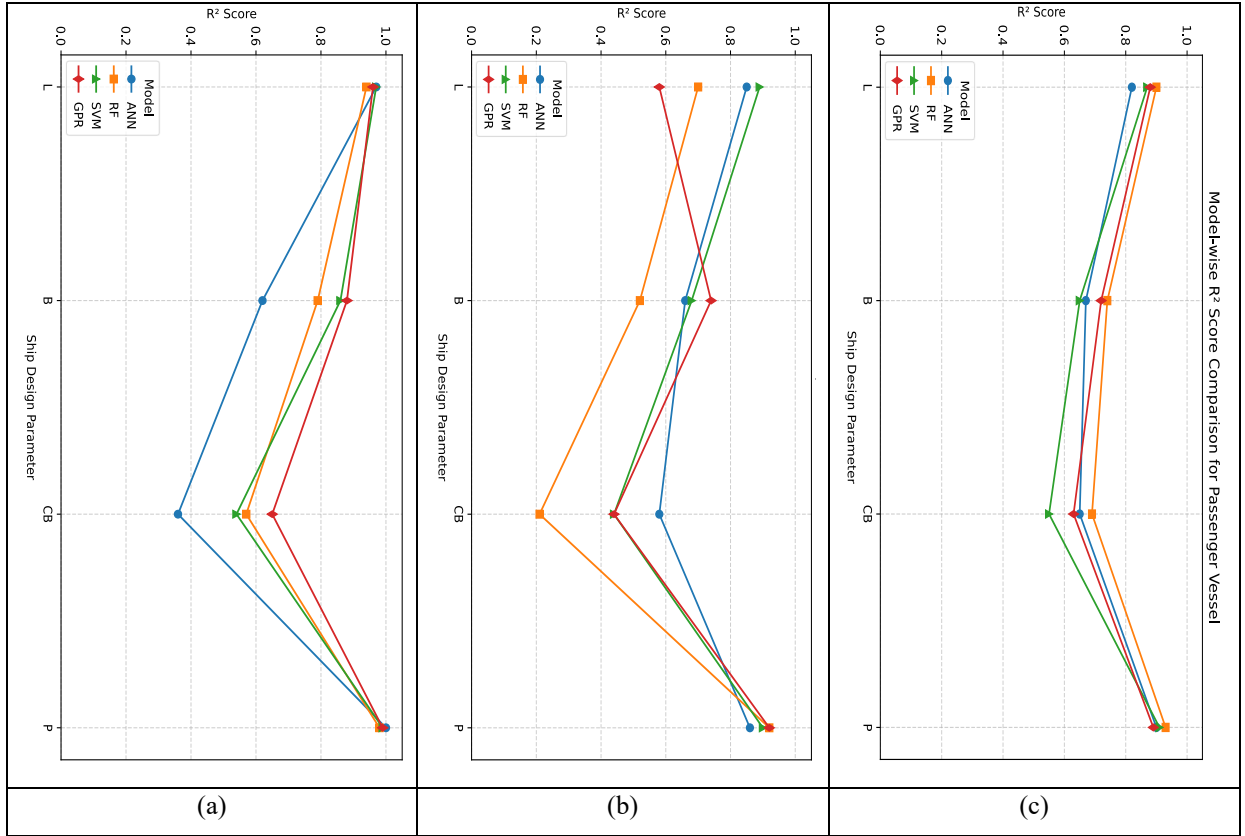


Figure 3: R^2 Comparison (a) Cargo Vessel (b) Oil Tanker (c) Passenger Vessel

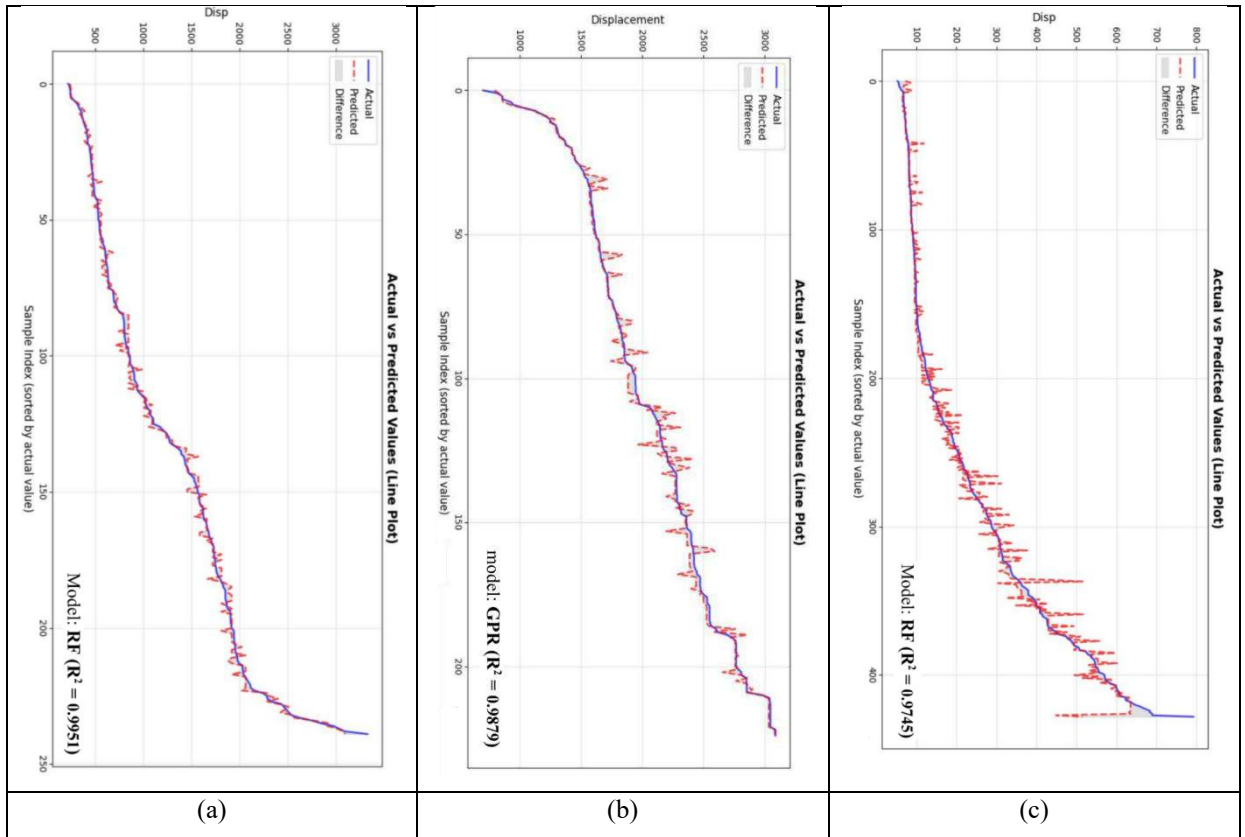


Figure 4: Best model plot of actual Vs predicted values of Displacement (a) Cargo Vessels (b) Oil Tanker (c) Passenger Ships

3.7 Constrained optimisation for design refinement

A defining feature of the research model is that prediction is not treated as an end in itself. Instead, validated machine-learning outputs were incorporated into constrained optimisation routines to support design refinement. MATLAB-based ‘**fmincon**’ and genetic algorithms were used to minimise engine power while maintaining practical design constraints on vessel particulars and operating conditions. This integration creates a direct link between prediction and engineering decision-making, allowing the model to be used for low-emission preliminary design rather than only retrospective analysis.

The optimisation framework is aligned with the vessel-specific objective functions developed in the report for cargo vessels, oil tankers, and passenger vessels. In each case, the central aim is to identify feasible design configurations that reduce energy demand while remaining within realistic bounds defined by deadweight, service speed, and hull constraints. This is a technically important step because it transforms the learning model into a design-support tool. It also ensures that the output is actionable for naval architects and operators rather than purely descriptive.

3.7.1 Mathematical Framework of the fmincon Algorithm

The ‘fmincon’ function in MATLAB is designed to minimize a given objective function while satisfying specified constraints. The process begins with the user defining the objective function to be minimized. Additionally, an initial guess for the values of the variables involved in the optimization is provided. For, this study it is predicted value, which serves as initial value. Constraints on these variables can also be defined, either as linear or nonlinear equality and inequality constraints. Furthermore, upper and lower bounds for each variable can be specified, restricting the search space for the optimization algorithm. The ‘fmincon’ function then employs various optimization algorithms, such as genetic algorithm, interior-point, active-set, trust-region-reflective or sqp methods, to iteratively minimize the objective function subject to the defined constraints and bounds. Throughout the optimization process, the algorithm continually refines the solution until it converges to a minimum within a specified tolerance level. Once convergence is achieved, the ‘fmincon’ function returns the minimum value of the objective function and the corresponding values of the variables that achieve this minimum. Mathematical algorithm of ‘fmincon’ function is as follows:

Input:
- Objective function ($f(x)$) to minimize. - Initial guess (x_0). - Constraints: - Linear inequality constraints: ($A \cdot x \leq b$). - Nonlinear inequality constraints: ($c(x) \leq 0$). - Linear equality constraints: ($A_{eq} \cdot x = b_{eq}$). - Nonlinear equality constraints: ($ceq(x) = 0$). - Bounds on variables: ($lb \leq x \leq ub$).
Output:
- Optimal solution (x^*) that minimizes ($f(x)$). - Minimum value of the objective function ($f(x^*)$).
Procedure:
1. Initialization: Set ($k = 0$) and ($x_k = x_0$).
2. Convergence Check: Check if ($\nabla f(x_k) = 0$) and all constraints are satisfied within the tolerance level. If true, terminate the algorithm and return(x_k)and ($f(x_k)$).
3. Optimization Step: Use optimization methods such as interior-point, active-set, or trust-region-reflective methods to update (x_k) to (x_{k+1}) such that:
$[x_{k+1} = \arg \min_x f(x)]$ subject to the constraints: $[A \cdot x \leq b]$ $[c(x) \leq 0]$ $[A_{eq} \cdot x = b_{eq}]$ $[ceq(x) = 0]$

and bounds:
$[lb \leq x \leq ub]$
4. Update: Set $(k = k + 1)$ and $(x_k = x_{k+1})$.
5. Repeat: Go to step 2 until convergence criteria are met.

3.7.2 Vessel-Specific Objective-Function Analysis

The vessel classes are analysed separately because the structure of the design space differs across cargo vessels, oil tankers, and passenger vessels. For each class, pairwise correlation analysis is first used to justify the regression structure, followed by residual diagnostics and actual-versus-predicted plots to verify adequacy of fit.

i. Cargo vessel

For cargo vessels, the pairwise plot shows strong positive associations among displacement, speed, draft, and engine power, with displacement acting as the dominant explanatory variable. This supports the use of a nonlinear polynomial objective function. Figure 5 presents the Pair plot and correlations of cargo vessel objective variables.

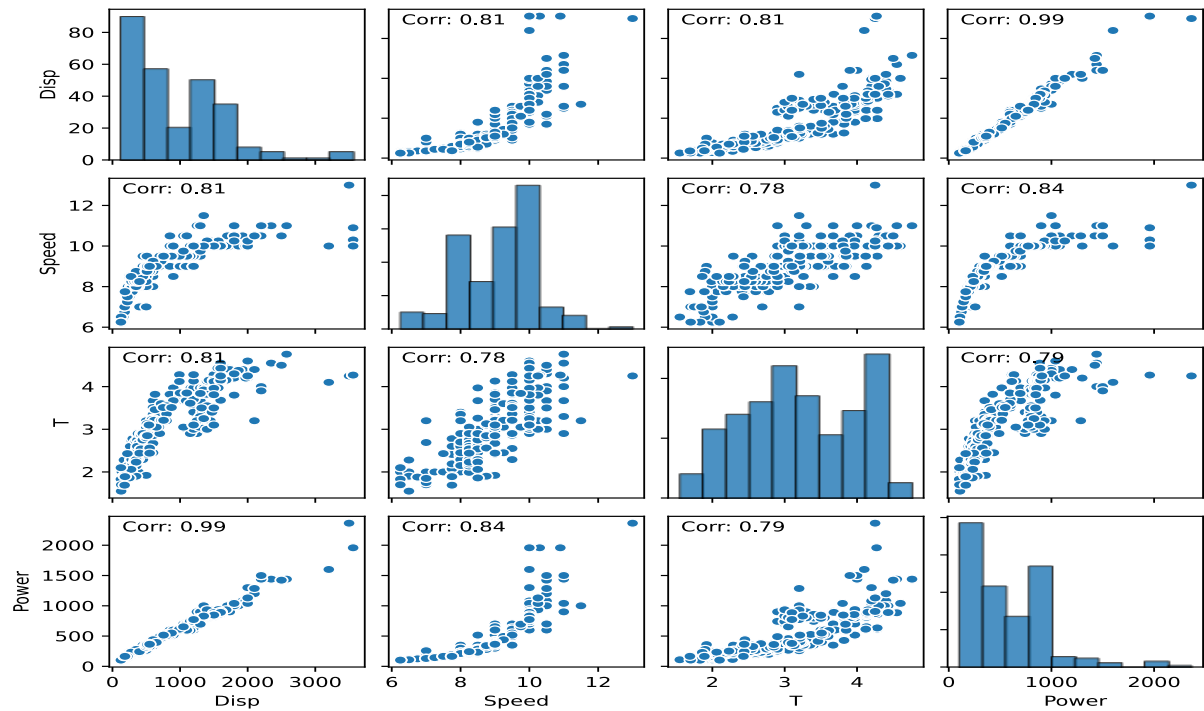


Figure 5. Pair plot and correlations of cargo vessel objective variables.

Model diagnostics indicate that the residuals (Figure 6 and 7) are centered close to zero and that the fitted relationship reproduces the observed engine power with high fidelity. The reported coefficient of determination is $R^2 = 0.989$ (Figure 14(a)), indicating an excellent fit and supporting the use of the derived function in subsequent optimization.

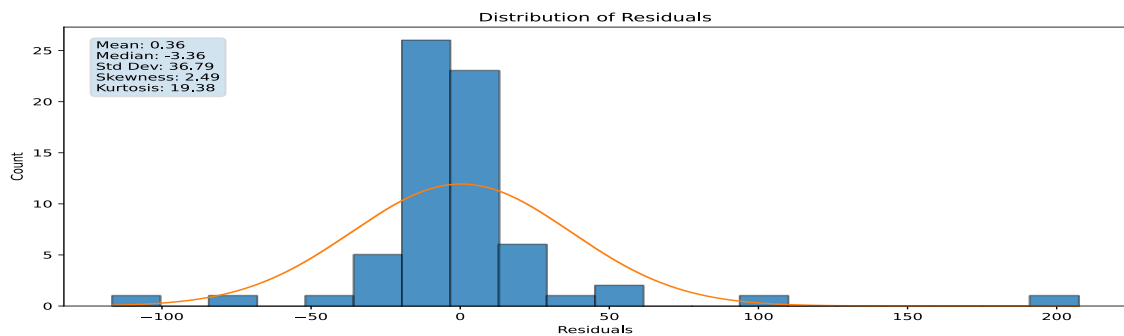


Figure 6. Distribution of residuals of cargo vessel objective function.

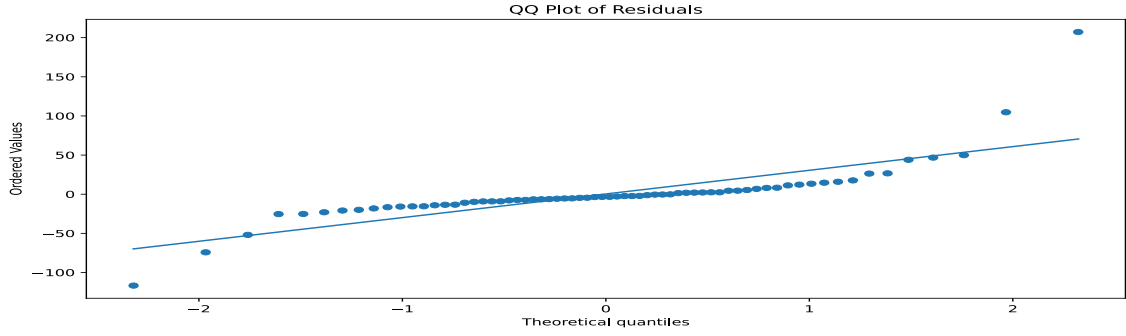


Figure 7. QQ plot of residuals of cargo vessel objective function.

ii. Oil tanker

For oil tankers, the pairwise plot again shows displacement as the most influential variable, while speed and draft also contribute substantially to power demand. The visible curvature and interaction patterns justify the same second-order formulation. Figure 8 presents the Pair plot and correlations of oil tankers objective variables.

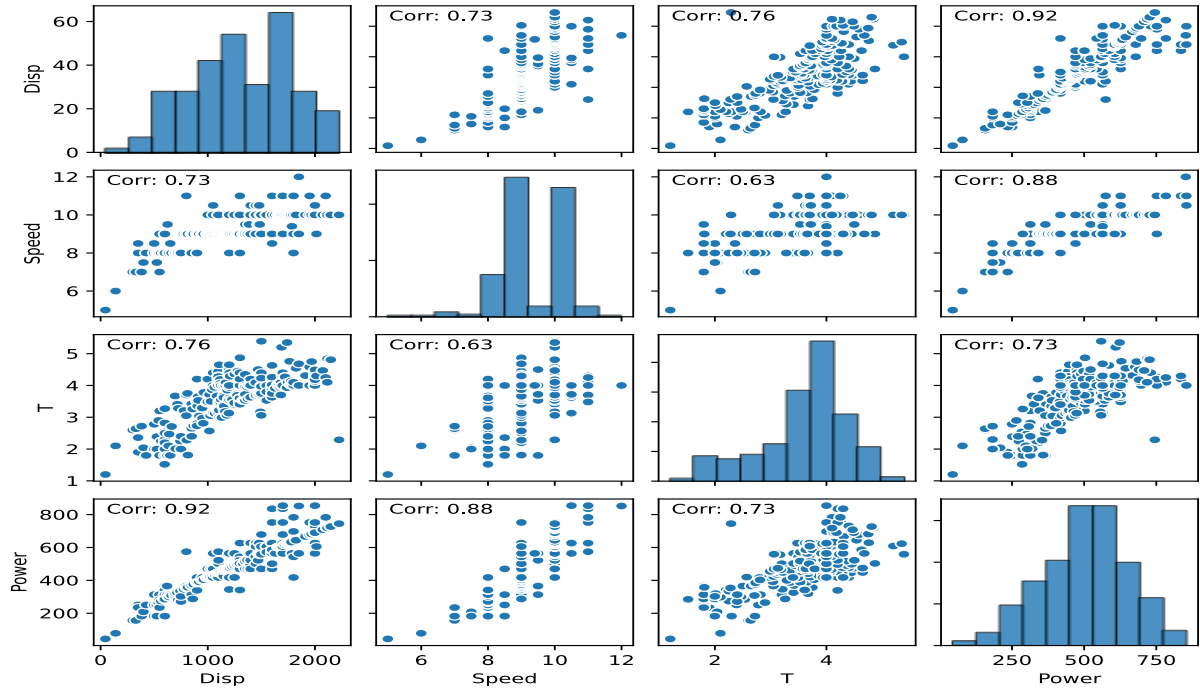


Figure 8. Pair plot and correlations of oil tanker objective variables.

The residual distribution and QQ plot indicate acceptable approximation quality, although a small number of high-power cases produce larger deviations (Figure 9 and 10). The model explains approximately 90.6% (Figure 14 (b)) of the variance in engine power ($R^2 = 0.906$), which is sufficiently strong for optimization-oriented use.

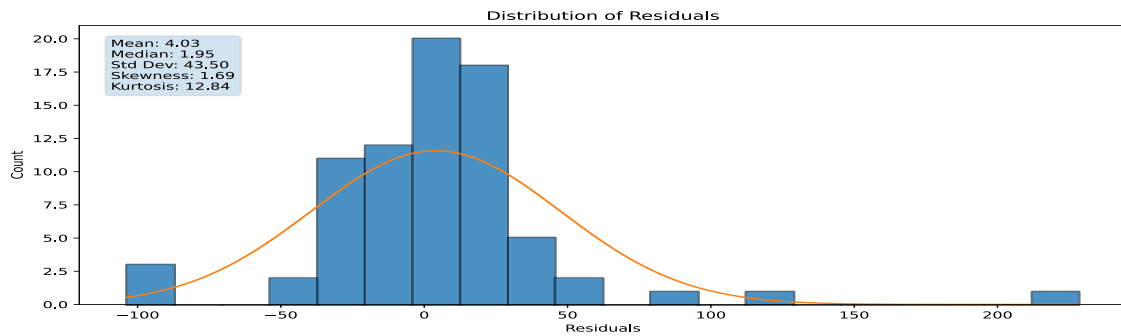


Figure 9. Distribution of residuals of oil tanker objective function.

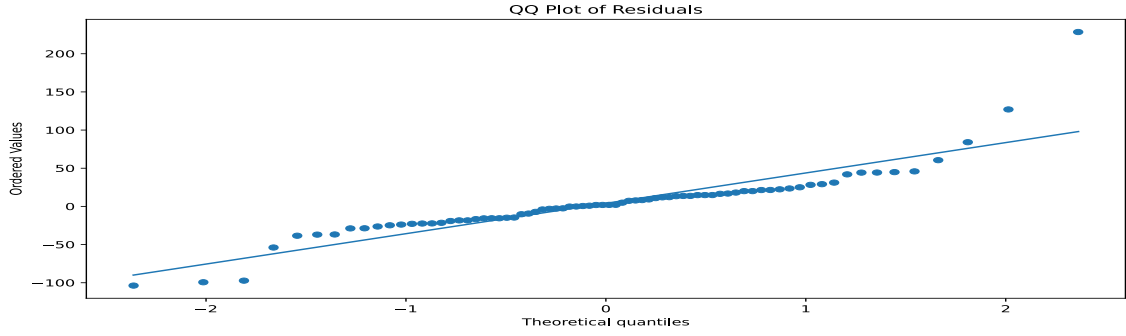


Figure 10. *QQ plot of residuals of oil tanker objective function.*

iii. Passenger vessel

For passenger vessels, displacement and speed are the most important drivers of engine power, while draft has a moderate influence. The scatter structure suggests nonlinear and interaction effects, supporting the polynomial objective function. Figure 11 presents the Pair plot and correlations of passenger vessels' objective variables.

The residual and fit plots show that the model captures the main trend in power demand, although greater scatter appears at higher power levels (Figure 12 and 13). The coefficient of determination is $R^2 = 0.901$ (Figure 14 (c)), indicating a robust but slightly less accurate fit than the passenger vessel case.

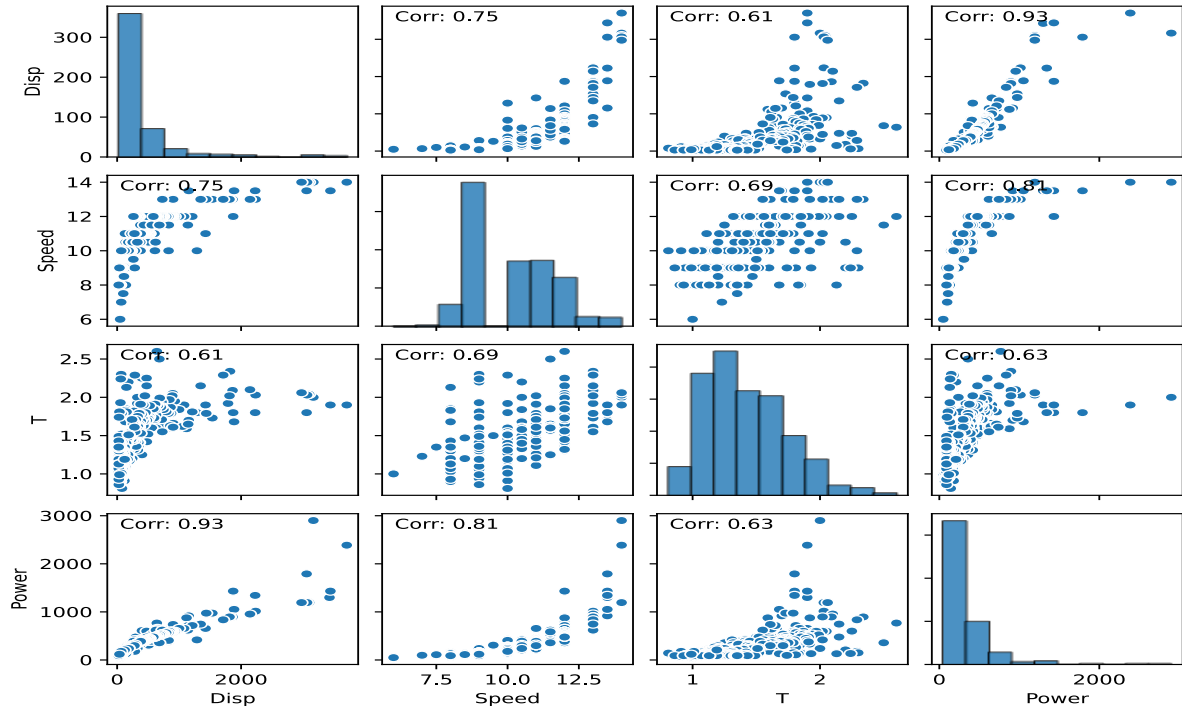


Figure 11. *Pair plot and correlations of passenger vessel objective variables.*

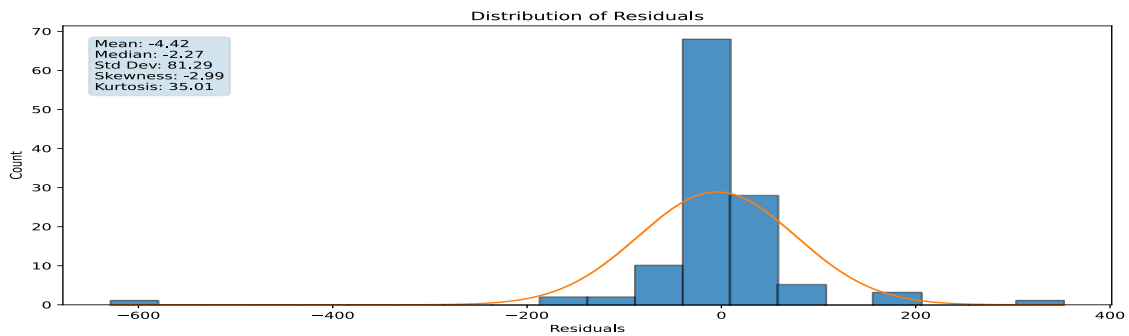


Figure 12. *Distribution of residuals of passenger vessel objective function.*

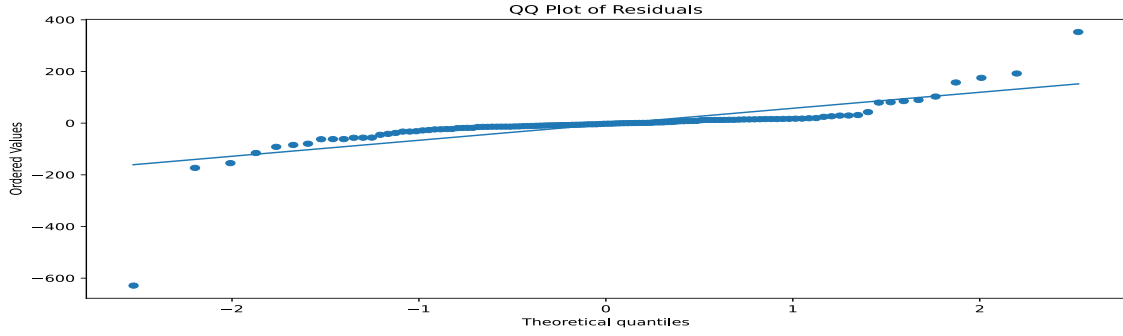


Figure 13. *QQ plot of residuals of passenger vessel objective function.*

3.7.3 Objective Function

The trained machine-learning models were incorporated into a constrained optimization framework to support efficiency-oriented preliminary ship design. In this study, optimization was used to identify combinations of principal particulars that reduce engine power while remaining within practical design limits. The models therefore served not only as predictors, but also as surrogate functions for design exploration. The optimization process was implemented in MATLAB, where the predicted relationships between the principal particulars and engine power were used to define the objective function.

A second-order polynomial formulation was adopted for each vessel class because ship resistance and propulsion demand vary nonlinearly with vessel dimensions and operating conditions. This choice allows both direct effects and interaction effects among the principal design variables to be represented in the optimization process. Separate objective functions were developed for cargo vessels, oil tankers, and passenger vessels. These are given in Equations (1), (2), and (3), respectively.

$$P_{\text{Cargo}} = 345.667 + 0.353923 \times \text{Disp} - 75.8397 \times \text{Speed} - 82.962 \times T + 3.07072 \times 10^{-5} \times \text{Disp}^2 + 0.0290548 \times \text{Disp} \times \text{Speed} - 0.0660657 \times \text{Disp} \times T + 5.734 \times \text{Speed}^2 + 5.86089 \times \text{Speed} \times T + 10.0688 \times T^2 \dots\dots\dots(1)$$

$$P_{\text{Tanker}} = -123.915 - 0.07157 \times \text{Disp} + 25.4268 \times \text{Speed} + 24.6157 \times T - 2.00845 \times 10^{-5} \times \text{Disp}^2 + 0.035213 \times \text{Disp} \times \text{Speed} - 0.00609451 \times \text{Disp} \times T + 0.937152 \times \text{Speed}^2 - 1.89017 \times \text{Speed} \times T + 0.58909 \times T^2 \dots\dots\dots(2)$$

$$P_{\text{Passenger}} = 255.975 + 1.14878 \times \text{Disp} - 20.8937 \times \text{Speed} - 402.928 \times T + 5.22504 \times 10^{-5} \times \text{Disp}^2 + 0.0125389 \times \text{Disp} \times \text{Speed} - 0.55647 \times \text{Disp} \times T - 0.567736 \times \text{Speed}^2 + 50.613 \times \text{Speed} \times T + 4.93967 \times T^2 \dots\dots\dots(3)$$

The purpose of the objective function is to identify the minimum feasible engine-power requirement for each vessel class while maintaining the design variables within realistic bounds. Displacement, service speed, and draft were used as the principal optimisation variables because they directly influence resistance, propulsion demand, and operational performance. The optimisation problem was therefore formulated as the minimisation of predicted engine power, expressed as:

$$f = \min(P) \quad (4)$$

3.7.4 Model Accuracy and Predictive Performance

Figure 14 compares the actual engine power values with those predicted by the regression-based objective function. The majority of data points cluster closely around the line of perfect agreement, indicating excellent predictive performance. The coefficient of determination, $R^2=0.989$, 0.906 and 0.901 for cargo, oil tanker and passenger ships, respectively. This demonstrates that the model confirms its high accuracy and suitability for optimization applications.

3.8 Uncertainty and Reliability Assessment

Although the present study primarily evaluates model performance using R^2 , and residual diagnostics, the results also provide insight into prediction uncertainty. The residual analyses indicate that the cargo-vessel objective function exhibits the highest reliability, followed by the oil-tanker and passenger-vessel models. While

all three objective functions achieved strong predictive performance ($R^2=0.989$, 0.906 , and 0.901 for cargo vessels, oil tankers, and passenger vessels, respectively), the passenger-vessel model showed greater residual dispersion and heavier tails, indicating increased uncertainty under atypical operating conditions. The displacement-based approach adopted for block-coefficient estimation further enhanced model robustness by replacing a weak direct prediction target with a more stable intermediate variable. Nevertheless, the study did not explicitly quantify prediction intervals or confidence bounds. Therefore, the reported results should be interpreted as demonstrating strong predictive capability within the observed data range, while future work should incorporate uncertainty quantification through confidence intervals, probabilistic prediction bounds, or external validation datasets to further assess model generalizability and reliability.

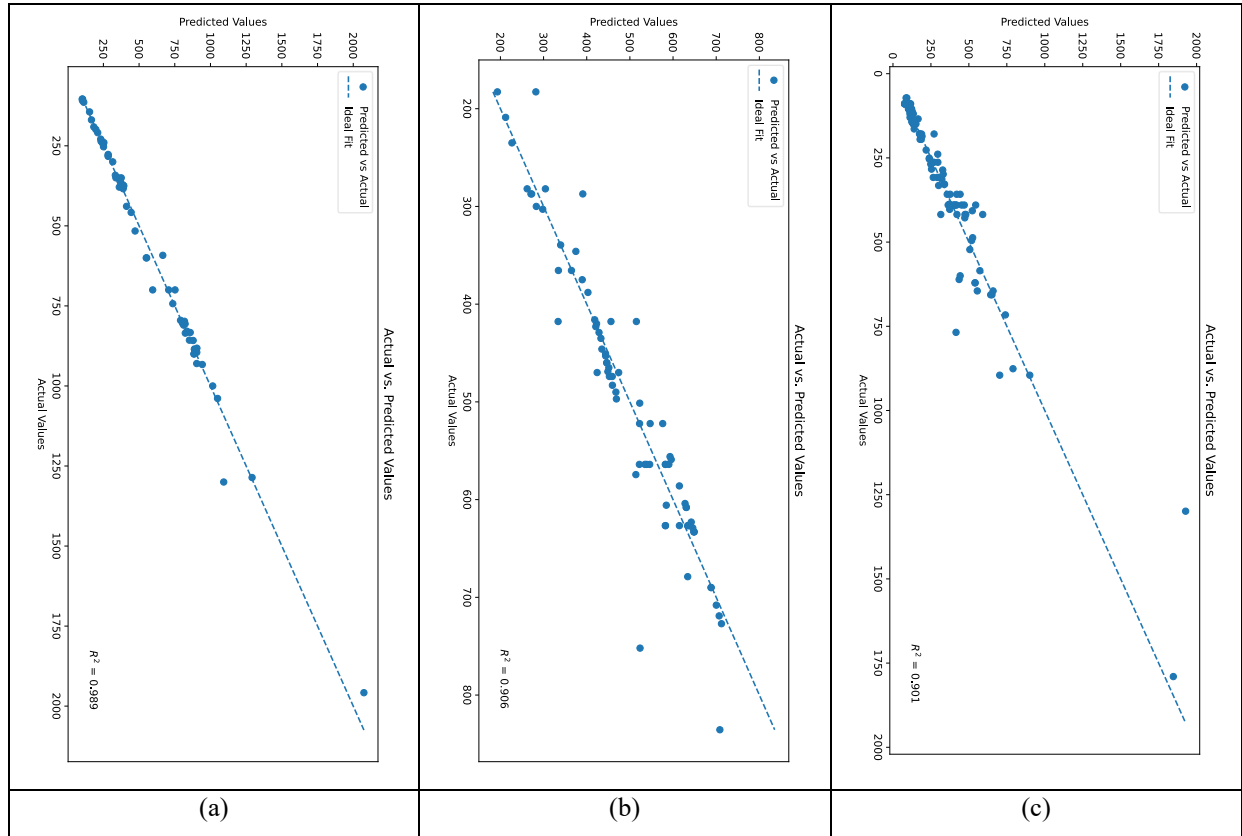


Figure 14. Actual versus predicted engine power (a) Cargo vessel (b) Oil Tanker (c) Passenger Vessel objective function

4.0 Discussion

The results indicate that machine learning can support preliminary ship-design prediction and energy-efficiency optimization, but model performance depends strongly on vessel class and target variable. Engine power was the most predictable output, whereas block coefficient remained the most difficult to estimate directly. This is consistent with the fact that power is closely linked to displacement, speed, and draft, while block coefficient is a more composite geometric parameter that is harder to infer from limited predictors. The comparative results further show that no single algorithm is universally superior: GPR was generally the most consistent model, SVR performed well for the oil-tanker subset, and RF provided stable performance for passenger vessels. Such class-dependent behaviour is expected in heterogeneous maritime datasets and supports the view that model selection should be driven by vessel structure rather than by a single universal learner.

The findings are also consistent with recent maritime ML studies, where data-driven methods have been shown to be effective for fuel-consumption prediction, performance monitoring, speed estimation, and resistance modelling. However, most of those studies focus on operational prediction rather than on principal ship-particular estimation integrated with constrained optimization. The present study extends the literature by combining prediction, displacement-based correction of a weak target, and optimization in a single workflow. This makes the framework more design-oriented and closer to practical naval-architectural decision support (Huang, 2022).

A further methodological contribution is the two-stage treatment of block coefficient. Direct prediction was weak, but estimating displacement first and then deriving block coefficient improved physical consistency and interpretability. This aligns with broader recommendations for hybrid and physics-informed maritime ML, where physically meaningful intermediate variables often produce more reliable results than direct black-box regression alone.

Limitation of this study is that model validation relied primarily on internal dataset partitioning and cross-validation, without independent external validation using an entirely separate vessel dataset. As a result, the reported performance should be interpreted as evidence of strong in-sample generalization within the observed data domain, rather than as proof of unrestricted transferability to all inland or coastal vessel classes. In addition, the available datasets remain relatively limited in size and class coverage, which may reduce robustness for rare design configurations or operating conditions not sufficiently represented in the training data. Therefore, while the proposed framework is suitable for preliminary design support within the studied range, its use for extrapolation beyond the observed vessel characteristics should be approached with caution.

5. Conclusion

This study developed an integrated data-driven framework for predicting principal ship particulars and supporting constrained optimisation for CO₂-oriented ship design. The results demonstrate that supervised learning models can capture the nonlinear relationships between ship geometry, operating conditions, and engine power with sufficient accuracy to inform preliminary design decisions.

Among the evaluated methods, GPR and SVR generally offered the strongest and most reliable performance across vessel classes, while RF provided robust predictions for heterogeneous passenger-vessel data and ANN remained effective for nonlinear trend learning. Engine power was consistently the easiest variable to model, whereas block coefficient was the most challenging. The displacement-based reformulation improved the stability of block-coefficient estimation and strengthened the physical credibility of the modelling chain.

The study confirms that machine learning, when combined with disciplined preprocessing and constrained optimisation, can function as a practical decision-support tool for low-emission ship design. The framework is particularly relevant for inland shipping, where heterogeneous fleets and limited data make rapid but reliable early-stage estimation especially valuable. Future work should expand the dataset, validate the models externally, and explore physics-informed hybrid formulations to further improve generalizability and interpretability.

Funding Declaration

This research work was funded by the University Grant Commission, Bangladesh.

Acknowledgement

The first author also thanks Rounak Saha Niloy, MSc Candidate at the University of New South Wales, Australia, for his assistance during the optimisation-related part of this research.

Data Availability

All data generated or analysed during this study are included in this published article [and its supplementary information files]

Reference

- Bentéjac, C., Csörgo, A., & Martínez-Muñoz, G. (2021). A comparative analysis of gradient boosting algorithms. *Artificial Intelligence Review*, 54, 1937–1967. <https://doi.org/10.1007/s10462-020-09896-5>
- Bertram, V. (2012). *Practical Ship Hydrodynamics* (2nd ed.). Butterworth-Heinemann, Elsevier.
- Bouman, E. A., Lindstad, E., Riialand, A., & Strømman, A. H. (2017). State-of-the-art technologies, measures, and potential for reducing GHG emissions from shipping: A review. *Transportation Research Part D: Transport and Environment*, 52, 408–421. <https://doi.org/10.1016/j.trd.2017.03.022>
- Dotse, A. (2024). Reference details to be verified for the final manuscript version.

- Drucker, H., Burges, C. J. C., Kaufman, L., Alex Smola, Vladimir Vapnik, Bell Labs, Monmouth University, & AT&T Labs. (1996). *Support vector regression machines*. https://proceedings.neurips.cc/paper_files/paper/1996/file/d38901788c533e8286cb6400b40b386d-Paper.pdf
- Durlik, I., Miller, T., Kostecka, E., & Tuński, T. (2024). Artificial intelligence in maritime transportation: A comprehensive review of safety and risk management applications. *Applied Sciences*, 14(18), 8420. <https://doi.org/10.3390/app14188420>
- Freiesleben, T., & Grote, T. (2023). Beyond generalization: A theory of robustness in machine learning. *Synthese*, 202(4), 1–28. <https://doi.org/10.1007/s11229-023-04334-9>
- Gupta, P., Rasheed, A., & Steen, S. (2022). Ship performance monitoring using machine learning. *Ocean Engineering*, 254, 111094. <https://doi.org/10.1016/j.oceaneng.2022.111094>
- Gypa, I., Jansson, M., Wolff, K., & Bensow, R. (2023). Propeller optimization by interactive genetic algorithms and machine learning. *Ship Technology Research*, 70(1), 56–71. <https://doi.org/10.1080/09377255.2021.1973264>
- Hasan, S. M. R. (2020). *Energy-efficient hydrodynamic design of ships for inland waterways of Bangladesh based on fuel consumption and emission control* (Unpublished thesis). Bangladesh University of Engineering and Technology.
- Hasan, S. R., Islam, M. S., Awal, Z. I., & Hossain, K. A. (2025). Prediction and optimization of efficient ship design particulars through advanced machine learning approaches. *Ocean Engineering*, 341, 122572. <https://doi.org/10.1016/j.oceaneng.2025.122572>
- Hassan, H. M., Elsakka, M. M., Refaat, A., Amer, A. E., & Rizk, R. Y. (2023). Optimal design of container ships geometry based on artificial intelligence techniques to reduce greenhouse gases emissions. In A. M. Negm, R. Y. Rizk, R. F. Abdel-Kader, & A. Ahmed (Eds.), *Engineering solutions toward sustainable development* (pp. 697–711). Springer. https://doi.org/10.1007/978-3-031-46491-1_44
- Hinton, G. E. (1992). How neural networks learn from experience. *Scientific American*, 267(3), 144–151. <https://doi.org/10.1038/scientificamerican0992-144>
- Hopfield, J. J. (1982). Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences*, 79(8), 2554–2558. <https://doi.org/10.1073/pnas.79.8.2554>
- Huang, L., Peña, B., Liu, Y., & Anderlini, E. (2022). Machine learning in sustainable ship design and operation: A review. *Ocean Engineering*, 266, 112907. <https://doi.org/10.1016/j.oceaneng.2022.112907>
- International Maritime Organization. (2023). *2023 IMO strategy on reduction of GHG emissions from ships*. IMO. <https://www.imo.org/en/ourwork/environment/pages/2023-imo-strategy-on-reduction-of-ghg-emissions-from-ships.aspx> (accessed 5 June 2026).
- 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. *Journal of Marine Science and Engineering*, 9(2), Article 137. <https://doi.org/10.3390/jmse9020137>
- Lundberg, S. M., & Lee, S.-I. (2017). A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems* 30, 4765–4774.
- Mekkaoui, S. E., Benabbou, L., Caron, S., & Berrado, A. (2023). Deep learning-based ship speed prediction for intelligent maritime traffic management. *Journal of Marine Science and Engineering*, 11(1), 191. <https://doi.org/10.3390/jmse11010191>
- Panda, J. P. (2023). Machine learning for naval architecture, ocean and marine engineering. *Journal of Marine Science and Technology*, 28(1), 1–26. <https://doi.org/10.1007/s00773-022-00914-5>
- Parisineni, S., & Pal, S. (2023). Reference details to be verified for the final manuscript version.
- Pii, Schmidhuber, & Neural Networks. (2015). Deep learning in neural networks: An overview. *Neural Networks*, 61, 85–117. <https://doi.org/10.1016/j.neunet.2014.09.003>
- Raissi, M., Perdikaris, P., & Karniadakis, G. E. (2019). Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational Physics*, 378, 686–707. <https://doi.org/10.1016/j.jcp.2018.10.045>
- Rasmussen, C. E., Williams, C. K. I., Baldi, P., Brunak, S., Sutton, R. S., Barto, A. G., Frey, B. J., Jordan, M. I., Spirtes, P., Glymour, C., Scheines, R., Hand, D., Mannila, H., Smyth, P., Herbrich, R., Schölkopf, B., Smola, A.

- J., & Alpaydin, E. (2006). *Gaussian processes for machine learning* (T. Dietterich, C. Bishop, D. Heckerman, M. Jordan, & M. Kearns, Eds.). MIT Press. <https://gaussianprocess.org/gpml/chapters/RW.pdf>
- Rawson, A., & Brito, M. (2023). A survey of the opportunities and challenges of supervised machine learning in maritime risk analysis. *Transport Reviews*, 43(1), 108–130. <https://doi.org/10.1080/01441647.2022.2036864>
- Salman, H. A., Kalakech, A., & Steiti, A. (2024). Random Forest algorithm overview. *Babylonian Journal of Machine Learning*, 2024, 69–79. <https://doi.org/10.58496/bjml/2024/007>
- Schneekluth, H., & Bertram, V. (1998). *Ship Design for Efficiency and Economy* (2nd ed.). Butterworth-Heinemann, Elsevier.
- Sun, Q., Zhang, M., Zhou, L., Garne, K., & Burman, M. (2022). A machine learning-based method for prediction of ship performance in ice: Part I. Ice resistance. *Marine Structures*, 83, 103181. <https://doi.org/10.1016/j.marstruc.2022.103181>
- Tadros, M., Ventura, M., & Guedes Soares, C. (2020). Optimization of the performance of marine diesel engines to minimize the formation of SOx emissions. *Journal of Marine Science and Application*, 19(3), 473–484. <https://doi.org/10.1007/s11804-020-00156-0>
- Tadros, M., Ventura, M., & Guedes Soares, C. (2023). Review of the decision support methods used in optimizing ship hulls towards improving energy efficiency. *Journal of Marine Science and Engineering*, 11(4), 835. <https://doi.org/10.3390/jmse11040835>
- Tadros, M., Vettor, R., Ventura, M., & Guedes Soares, C. (2021). Coupled engine-propeller selection procedure to minimize fuel consumption at a specified speed. *Journal of Marine Science and Engineering*, 9(1), 59. <https://doi.org/10.3390/jmse9010059>
- Tomljenovic, V., Merzifonluoglu, C., & Spigler, S. (2024). Optimizing inland container shipping through reinforcement learning. *Annals of Operations Research*, 339(1), 1–27. <https://doi.org/10.1007/s10479-024-05927-4>

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryHyperparameterTuning.docx](#)
- [SupplementaryShipDataset.docx](#)
- [SupplementaryModevaluationandcomparativeperformance.docx](#)