ARTICLE DRAFT

Modelling and optimization of ship's fuel consumption using Random Forest Regression (RFR)

Muhammad Fakhruriza Pradana[a,b], Hibatul Wafi[b], Bernd Noche[b]

[a]Department of Civil Engineering, University of Sultan Ageng Tirtayasa, Cilegon, Indonesia ; [b]Institute of Transport Systems and Logistics, University of Duisburg-Essen, Duisburg, Germany

ARTICLE HISTORY
Compiled August 16, 2023

ABSTRACT
Efforts to model energy-efficient operation of shipping operations using machine-learning methods have emerged due to volatile bunker fuel prices and stringent environmental regulations. It is widely regarded that ship speed is one of the most influential factors impacting ships' fuel oil consumption and as such, accurate modelling of ship speed is paramount to ensure the accuracy of subsequent FOC prediction.

This study proposes an intuitive data-driven modelling approach, integrating Automatic Identification System and weather data for modelling of ship states and environmental conditions' impact on FOC. Grey Box Modelling approach divides the speed and FOC prediction into stages, the first stage involves the prediction of speed over ground using Random Forest Regressor. Consequently, the FOC prediction based on predicted speed employs the empirical formula by Holtrop-Mennen, maintaining adherence with established vessel knowledge.

In the presented case study, optimised SOG prediction achieves 3.94% mean absolute percentage error (MAPE) and 93.41% R^2 score. Subsequent FOC prediction from estimated speed yields 86.57% R^2 and 12.06% MAPE. The results affirm the proposed approach's viability in predicting energy-efficient ship operations.

KEYWORDS
Energy-efficient operation; Random Forest Regression; Ship speed prediction; Fuel consumption prediction; Grey Box Model; AIS

1. Introduction

The marine industry is actively researching efficient ship operation due to rising fuel prices and stricter environmental rules. Fuel costs, known as "bunkers," comprise over 50% of voyage expenses and up to 75% of total operating costs, impacting profitability (Bialystocki and Konovessis 2016). Energy-efficient practices reduce costs and greenhouse gas emissions, crucial with shipping contributing 2.51% of global emissions (IMO 2020). This mutual motivation aligns economic benefits with environmental compliance. Stakeholders seek solutions to energy-efficient operation by considering technical and operational approaches. Technical solutions require costly structural and power system alterations (Yan, Wang, and Psaraftis 2021; Li et al. 2022), prompting interest in the cost-effective, optimisation of operational measures.

Muhammad Fakhruriza Pradana. Email: mfakhruriza@untirta.ac.id

Significant emphasis is given in this study on optimisation of ship speed due to its substantial impact on fuel consumption which is caused by a third-order non-linear correlation between fuel consumption and ship speed (Wang and Meng 2012; Du et al. 2019). However, the process of optimising the speed prediction model is intricate, appropriate features must be considered as the ship speed is influenced by factors like vessel performance and weather conditions.

Fuel consumption models based on historical data and ship parameters lack robustness and sensitivity to noise. To address this, recent research employs data-driven techniques, like machine learning (ML), for ship speed and fuel consumption prediction. ML models showcase strong generalisation capabilities and low prediction errors, although some experts are reluctant to accept the generated models by the machine learning approach due to their complexity, unintuitiveness, and potential violation of vessel physics. The success of data-driven models is also highly dependent on data quality and quantity (Yan, Wang, and Psaraftis 2021; Gkerekos, Lazakis, and Theotokatos 2019). Given volatile fuel prices, developing an accurate Fuel Oil Consumption (FOC) prediction model is valuable for maritime stakeholders. This aids in timely economic decisions without violating environmental regulations.

2. Literature Review

2.1. *Modelling Approach for Ship Operation*

Haranen et al. (2016) and Coraddu et al. (2017) categorised fuel consumption prediction models into three strategies:

White Box Models (WBM): Built on prior mechanistic knowledge and physical principles of a vessel's system, including its structure, design parameters, and propulsion configuration.

Black Box Models (BBM): Data-driven and developed using data from different sailing journeys and historical observations. The Machine Learning (ML) modelling approach focuses on the prediction of bunker consumption at different points in time.

Grey Box Models (GBM): A fusion of WBM and BBM, resulting in a single model that considers both *a priori* knowledge of the vessel and historical sailing data, This method aims to complement the performance of WBM and BBM.

Each strategy has strengths and weaknesses. WBM is transparent and comprehensible, rooted in physics and hydrodynamics, but lacks adaptability and generalisation due to its deterministic nature and dependence on prior knowledge. BBM excels in fitting and predicting data but lacks vessel-specific knowledge and can be complex. To achieve good prediction, it requires an abundance of data quantity and good data quality (Halevy, Norvig, and Pereira 2009). GBM mitigates these limitations by combining mechanistic understanding with predictive capabilities.

The modelling of FOC using GBM requires both components of WBM and BBM. For the BBM modelling part using ML approach, For black-box modelling using ML techniques, it is crucial to have sufficient high-quality data for accurate training (Halevy, Norvig, and Pereira 2009). Yan, Wang, and Psaraftis (2021) categorise the data sources for FOC modelling as follows:

Besides its intended role as a collision avoidance system, Automatic Identification System (AIS) data finds potential in ship behaviour analysis and environmental assessment. The International Maritime Organization (IMO) utilized AIS data to study Greenhouse Gas (GHG) emissions, estimating global shipping emissions (IMO 2020; Smith et al. 2015). Rakke (2016) introduced ECAIS as a methodology to compute ship emissions from AIS-derived fuel consumption data using Holtrop-Mennen and literature-based approximations. The study by Kim et al. (2020) used AIS data, ship information, and environmental data for estimating Energy Efficiency Operational Indicator (EEOI). The use of AIS data in research aims for data independence, reducing reliance on commercial databases.

2.2. *Predictive performance of tree based models*

Tree-based model is a supervised, highly interpretable BBM modelling approach using machine learning approach which is adept in classification and regression tasks. The model is inherently resistant to multicollinearity problems (Yan, Wang, and Psaraftis 2021). Several literature studies reveal its advantages and performance superiority. Soner, Akyuz, and Celik (2018) employed ferry data to predict FOC using tree-based models including bagging, random forest (RF), and bootstrap. RF achieved 43.5 L/h RMSE for fuel consumption, outperforming Artificial Neural Network (ANN) model employed by Petersen, Jacobsen, and Winther (2012).

Yan, Wang, and Du (2020) predicted FOC for a dry bulk ship's voyage using RF. The model incorporated sailing speed, cargo weight, and meteorological conditions, it is able to attain mean absolute percentage error (MAPE) of 7.91% and the RF model outperformed decision tree, ANN, LASSO, and SVR. Gkerekos, Lazakis, and Theotokatos (2019) compared ML models to predict daily FOC, RF model achieved 89% and 96% R^2 scores with noon data and ADLM system data, respectively. Li et al. (2022) fused meteorological, voyage, and AIS data to explore the effect of data on ML models for FOC prediction. Tree-based models (bagging and boosting ensembles) including ETR, RFR, AB, GB, XG, and LB were recommended for energy-efficient operation modelling, with RFR particularly displaying the best robustness among the presented ML models in the study. Abebe et al. (2020) predicted ship speed over ground (SOG) using AIS and weather data. The RF model achieved 98% R^2 score and 0.25 knots RMSE.

WBMs for predicting FOC utilize physics and hydrodynamic laws to compute the vessel's resistance, encompassing calm water resistance and additional effects like wind and waves. Then the engine power can be subsequently estimated at a specific speed, facilitating FOC calculation (Haranen et al. 2016). Holtrop-Mennen power estimation method Holtrop (1984), is applicable in a wide range (Rakke 2016; Kim et al. 2020). Rakke (2016) utilized AIS data and mechanical information to estimate engine power using Holtrop-Mennen, achieving about 5% model testing error for FOC and GHG emissions estimation. Similarly, Kim et al. (2020) estimated Energy Efficiency Operational Index (EEOI) through Holtrop-Mennen-based engine power estimation, enabled by AIS data and weather information.

3. Methodology

This chapter covers the methodology used to construct the grey box model (GBM). The grey box approach employed in this study is categorised as sequential GBM,

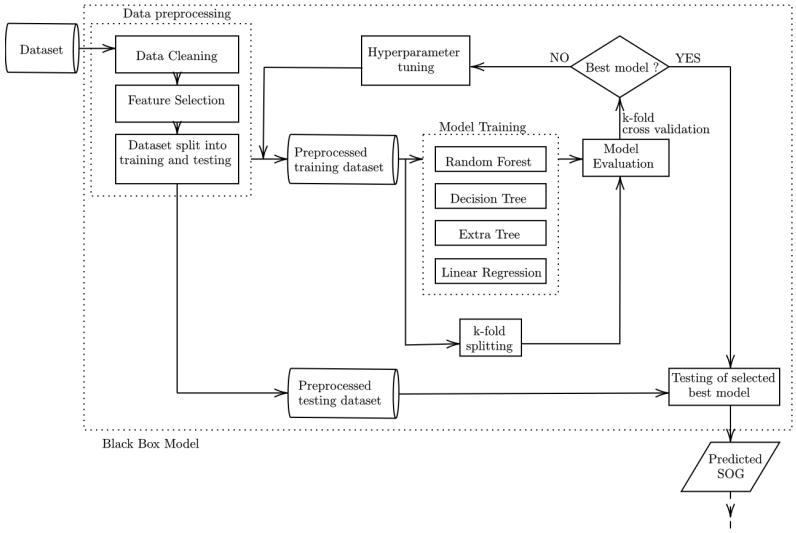

Figure 1. Scheme of proposed BBM methodology

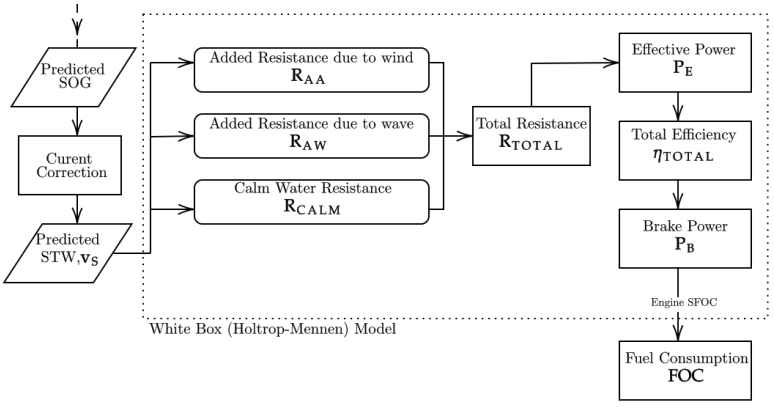

Figure 2. Scheme of proposed BBM methodology

which entails a two-stage development process. The initial stage focuses on machine learning modelling using tree-based models. The modelling is carried out using Python in conjunction with `Scikit-Learn` (Pedregosa et al. 2011).

The second stage of the modelling process revolves around the power estimation method (Holtrop 1984). This involves an initial conversion of SOG to STW for estimation of encountered resistance during the voyage, this then facilitates the estimation of the required power i.e. the energy required to propel the ship.

3.1. *Data Acquisition*

The data is collected from a ferry serving between the ports of Køge, Rønne, Ystad, and Sassnitz. The trip duration between Køge and Rønne is approximately 5 hours and 30 minutes, while the voyage between Rønne and Sassnitz takes around 3 hours and 20 minutes. The Danish Maritime Authority's (DMA) T-AIS system tracks the journey. Weather data along the ferry's route is sourced from ECMWF, providing

IMO	9812107
Type & Service	Passenger ferry
L_{OA}	158.00 m
L_{WL}	144.80 m
B (moulded)	24.5 m
T_{DESIGN}	5.70 m
T_{MAX}	5.85 m
Gross Tonnage (GT)	18,009
Deadweight (dwt)	4,830 t
Main Engines	Wärtsillä 8V31 2 x 4,880 kW
SFOC	169.4 g/kWh
Service Speed	17.7 knots
Bow Thrusters	2 x 1500 kW

Figure 3. Particular of M/S Hammershus

Figure 4. Journey of the ferry

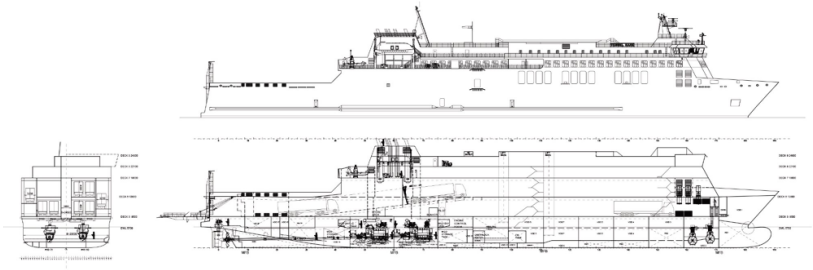

Figure 5. Schematics of M/S Hammershus

information on wind, waves, and seawater temperature. This data has a temporal resolution of 1 hour and a spatial granularity of 0.25° (longitude) x 0.25° (latitude). Current information, obtained from CMEMS, is available at a temporal resolution of 3 hours and a spatial granularity of 0.25° (longitude) x 0.25° (latitude). The resulting combined dataset maintains a temporal resolution of 1 hour. To address the temporal resolution disparity between CMEMS and ECMWF data, the weather information is synchronized. This synchronization ensures that the wind, waves, seawater temperature, and sea current data are aligned with the same weather grid and maintain consistent temporal resolutions.

3.2. *Data Pre-Processing*

This section outlines the data preprocessing steps, including data cleaning to identify anomalies and the handling of missing values. threshold application to the Speed Over Ground (SOG) and feature selection that is based on domain knowledge, ensuring alignment with vessel characteristics. The procedures are performed to ensure that the dataset is in the state required for modelling.

3.2.1. Data Cleaning

Analysis of the data points indicated an incomplete representation of the voyage between Rønne and Sassnitz due to limitations in the T-AIS system's coverage. Therefore, a latitude threshold of 55.04° N is implemented, excluding the voyage segment between Sassnitz and Rønne. To ensure that the dataset accurately captures the ship's operational conditions under steady state, a threshold is applied to the SOG. The change in

Table 1. Structure of training dataset

Training Label	
SOG [Knots]	`sog`
Training Features	
COG [°]	`cog`
Heading [°]	`heading`
Draught [m]	`draught`
Wind Speed [m/s]	`windspeed`
Air Temperature Above Oceans [K]	`oceantemperature`
Wave Period [s]	`waveperiod`
Sea Surface Temperature [K]	`surftemp`
Combined Wind Wave Swell Height [m]	`windwaveswellheight`
Current Speed [m/s]	`curspeed`
True Wind Direction [°]	`truewinddir`
True Current Direction [°]	`truecurrentdir`
True Wave Direction [°]	`truewavedir`

SOG can originate from changing sea conditions or intentional speed adjustments during port arrivals and departures. Therefore, data points with SOG below 5 knots, indicative of manoeuvring or stationary activities, are removed from the dataset (Abebe et al. 2020; Yan, Wang, and Du 2020). As a result of this filtering process, the dataset size notably reduces from 7,453 data points to 3,828 data points. This reduction highlights that approximately half of the initial data points correspond to periods when the ship was stationary or engaged in low-speed manoeuvres.

Missing and `NaN` values are imputed using the `KNNImputer` feature from `Scikit-Learn`. This step is essential as the modelling package provided by `Scikit-Learn` cannot handle instances with missing values. The choice of using a k-nearest neighbour imputation strategy is appropriate, as it aims to capture the weather conditions within the vicinity of the missing values.

3.2.2. Feature Selection

To select appropriate features for the model, feature correlation is analyzed. Feature selection aims to simplify the model and reduce computational costs during training. The High Correlation Filter, proposed by Abebe et al. (2020), is used. It treats feature pairs with correlation coefficients above 0.7 as a single entity. However, feature selection here is primarily guided by physical reasoning, prioritizing physical principles over statistics.

Features from AIS data such as *time, latitude, longitude, width,* and *length* are excluded, as they only represent ship location and constant dimensions. Features from weather data, such as *combined wind wave swell height, swell height, maximum wave height,* and *wind wave height,* are interconnected by physical relationships. The combined wind wave swell height corresponds to significant wave height $H_{1/3}$. Additionally, significant wave height can be used to identify whether the sea is dominated by swell or wind-generated waves (Bitner-Gregersen 2005). Hence, it is evident that retaining the significant wave height is essential for the model, given that various wave properties can be deduced from it.

In a statistical context, heading and COG exhibit significant correlations. However, both features are retained due to their representation of distinct ship parameters. Course Over Ground (COG) signifies the ship's course heading while heading signifies the ship's actual heading at a specific time point. A similar rationale applies to the relationship between air temperature above the ocean and sea surface temperature. Air temperature above oceans represents wind temperature, whereas sea surface temperature reflects the temperature of the water surface.

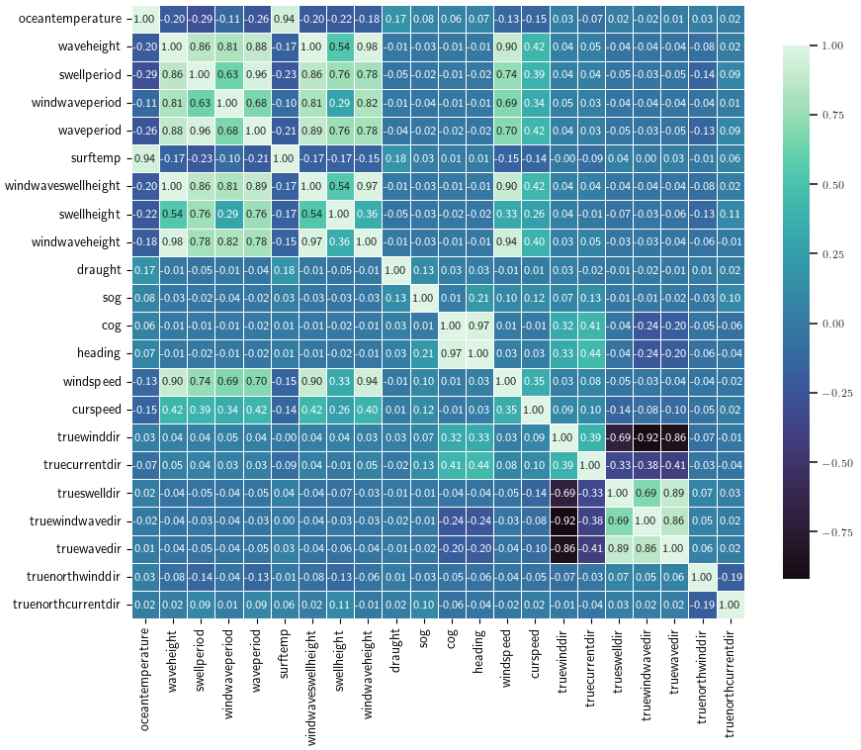

Figure 6. Correlation Heat Map

3.3. *Modelling methodologies*

3.3.1. Decision Tree (DT) Regressor

The Decision Tree operates by employing nested `if-then` statements based on predefined rules, resulting in a partitioned data space. This process can also be visualized as a binary tree, enhancing interpretability by representing diverse input responses within a single tree Kuhn and Johnson (2013); Hastie, Tibshirani, and Friedman (2009).

A Decision Tree encompasses distinct nodes: the ***Root node*** represents the top-level node; ***Leaf nodes*** (or terminal nodes) yield final prediction outcomes; and ***Internal nodes*** lie between the root and leaf nodes. The procedure of dividing a node into subsequent nodes is termed ***splitting***, where the original node is the ***parent node*** and the resultant nodes are ***child nodes***. In regression tasks, tree growth is often controlled by Mean Square Error (MSE), guided by the Classification and Regression Tree (CART) algorithm, The notation $J(\mathrm{k}, \mathrm{t}_k)$ represents the cost function that needs to be minimised.

$$\mathrm{MSE}_{\mathrm{s}_i} = \frac{1}{n_{\mathrm{s}_i}} \mathrm{SSE}_{\mathrm{s}_i} \quad \textbf{where} \quad i = (1, 2) \tag{1}$$

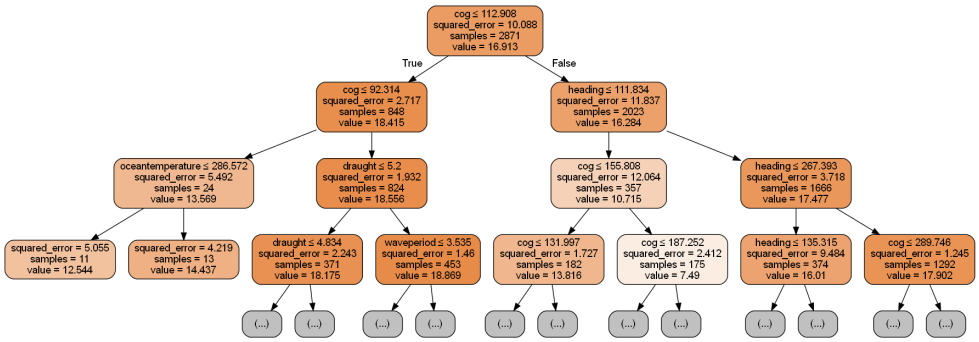

Figure 7. Structure of a Decision Tree (DT) Regressor

$$J(\mathrm{k}, \mathrm{t}_k) = \frac{1}{n_{\mathrm{s}_1}} \mathrm{SSE}_{\mathrm{s}_1} + \frac{1}{n_{\mathrm{s}_2}} \mathrm{SSE}_{\mathrm{s}_2} \begin{cases} \mathrm{SSE}_{\mathrm{s}_i} = \sum_{i \in \mathrm{s}_i} (\hat{y}_{\mathrm{s}_i} - y_{\mathrm{s}_i})^2 \\ \hat{y}_{\mathrm{s}_i} = \frac{1}{n_{s_i}} \sum_{i \in \mathrm{s}_i} y \end{cases} \tag{2}$$

The process of tree growth stops until either the number of samples for splitting reaches a predefined threshold or when no further split can be found which reduces the MSE. The decisions from optimal splits are visualised through a binary tree representation, enhancing the interpretability and ease of implementation. The inherent logic structure of decision trees enables them to handle diverse data types without extensive preprocessing, including sparse, skewed, continuous, and categorical data. Decision trees also inherently perform feature selection, which is a valuable aspect in modelling (Kuhn and Johnson 2013).

However, an unconstrained single decision tree is prone to overfitting due to its tendency to closely match the training data. This model's instability can lead to substantial changes in its structure when the data is altered, resulting in a completely different interpretation of splits (Hastie, Tibshirani, and Friedman 2009; Kuhn and Johnson 2013). To mitigate overfitting, it becomes essential to regularise the decision tree's growth during training. The following parameters control the growth of a single decision tree :

- `max_depth`: This hyperparameter is defined as the count of nodes along a path from the root node to its parent node. The default parameter allows full unpruned growth of the tree.
- `min_samples_leaf`: This hyperparameter controls the number of samples required to be at the leaf node, where the split point will be considered if the leaf contains at least `min_samples_leaf=n` training samples in each left and right branch.
- `min_samples_split`: This hyperparameter controls the minimum number of samples i.e. data points required to split a node.

3.3.2. Random Forest (RF) Regressor

Ensemble learning offers a solution to enhance the performance of Decision Tree (DT) regressors. This concept involves combining the strengths of multiple simpler base models (Hastie, Tibshirani, and Friedman 2009). One prominent ensemble method is the ***Random Forest***, introduced by Breiman (2001), which involves creating

bootstrap samples, randomly selecting splitting features, and aggregating predictions. This approach combines various learning algorithms, referred to as weak learners, with each corresponding to an individual decision tree in the Random Forest. Random forest uses the Bagging (*bootstrap aggregating*) strategy, where it trains each tree using bootstrap samples, where instances from the training set are randomly selected with replacement.

To further improve bagging, reducing the correlation between trees is applied. This involves introducing randomness during tree construction. Random split selection, as introduced by Dietterich (2000), involves selecting a feature from a random subset for each split. This, coupled with the inherent instability of a single decision tree, addresses overfitting and the lack of robustness of DTR. The Random Forest methodology addresses these issues by creating an ensemble of independent, strong learners, resulting in reduced variance and robustness against noisy data (Breiman 2001). While losing some interpretability compared to basic tree-based models, the impact of each feature in the ensemble can still be quantified (Kuhn and Johnson 2013). Random Forest performs better with larger sample sizes, and extensive parameter tuning is often unnecessary for good prediction results (Kuhn and Johnson 2013; Hastie, Tibshirani, and Friedman 2009).

In addition to the hyperparameters used to fine-tune the decision tree, the RF model provides additional hyperparameters to control the growth of the tree:

- `max_features`: This hyperparameter controls the number of features to be considered when looking for the best split. The default parameter considers all features during training.
- `n_estimators`: This hyperparameter controls the number of trees i.e. predictors in a forest.

3.3.3. Model Hyperparameter Optimisation

`Scikit-Learn` provides both the `GridSearchCV` and `RandomizedSearchCV` methods to assist in the search for optimal hyperparameters. Both approaches share a similar principle: the specified hyperparameters and their corresponding value ranges are evaluated through cross-validation to determine the best combination. The distinction between `GridSearchCV` and `RandomizedSearchCV` lies in how they search for the best hyperparameter values:

- `GridSearchCV`: This method constructs a grid comprising all possible combinations of hyperparameter values within the specified ranges. It exhaustively explores this grid to find the best combination.
- `RandomizedSearchCV`: Randomly samples hyperparameter values from the specified ranges. It offers more control over computational resources by allowing the specification of the number of iterations. This method often produces accurate results and is computationally more efficient than `GridSearchCV` (Bergstra and Bengio 2012).

Due to the computational limitation posed by exhaustive grid search, the `RandomizedSearchCV` approach will be adopted to identify optimal hyperparameters.

3.3.4. Holtrop-Mennen Method

A ship's bunker fuel consumption in actual operating conditions is affected by several factors including the operating parameter of the ship's engine, propeller efficiency, and encountered resistance by the ship. Furthermore, a ship's propulsion power is correlated to the sailing speed (SOG) and meteorological conditions (Lang 2020). Therefore, in addition to the calm water resistance R_{CALM}, the additional resistance caused by wind R_{AA} and wave R_{AW} should be considered to estimate the total resistance of the ship R_{TOTAL}. The power needed to propel a ship forward at a given ship STW v_S, to overcome R_{TOTAL} is defined as **effective power** P_e:

$$R_{TOTAL} = R_{CALM} + R_{AW} + R_{AA} \tag{3}$$

$$P_e = R_{TOTAL} \cdot v_S \tag{4}$$

The effective power P_e is transmitted through the shaft connected to the main engine of the ship which generates power to rotate the propeller of the ship, which is termed as **brake power of the engine,** P_b. The brake power can be calculated through effective power by considering the **shaft efficiency** η_s**, hull efficiency** η_h**, relative rotative efficiency** η_r **and open water efficiency** η_o:

$$P_b = \frac{P_e}{\eta_s \cdot \eta_h \cdot \eta_r \cdot \eta_o} \tag{5}$$

The bunker fuel consumption can then be calculated by multiplying the brake power P_b with the Specific Fuel Oil Consumption (SFOC) and the operation time τ_{OP}:

$$FOC = P_b \cdot SFOC \cdot \tau_{OP} \tag{6}$$

Since the speed that is represented in AIS data is SOG, v_G, given the current speed v_C and the current direction with respect to true north γ, the conversion to speed through water STW, v_S is done by the following equations (Kim et al. 2020):

$$v_{\mathrm{G}}^{x} = v_{\mathrm{G}} \cdot \sin(\alpha) \tag{7}$$

$$v_{\mathrm{G}}^{y} = v_{\mathrm{G}} \cdot \cos(\alpha) \tag{8}$$

$$v_{\mathrm{C}}^{x} = v_{\mathrm{C}} \cdot \sin(\gamma) \tag{9}$$

$$v_{\mathrm{C}}^{y} = v_{\mathrm{C}} \cdot \cos(\gamma) \tag{10}$$

$$v_{\mathrm{S}}^{x} = v_{\mathrm{G}}^{x} - v_{\mathrm{C}}^{x} \tag{11}$$

$$v_{\mathrm{S}}^{y} = v_{\mathrm{G}}^{y} - v_{\mathrm{C}}^{y} \tag{12}$$

$$v_{\mathrm{S}} = \sqrt{(v_{\mathrm{S}}^{x})^2 + (v_{\mathrm{S}}^{y})^2} \tag{13}$$

For the calculation of the total resistance, the following equations based on the study by Holtrop and Mennen (1978, 1982); Holtrop (1984) summarised in the work by Birk (2019):

$$R_{CALM} = R_F(1 + k_1) + R_{APP} + R_W + R_B + R_{TR} + R_A \tag{14}$$

$$R_F = \frac{1}{2}\rho v_S^2 S C_F \tag{15}$$

$$R_{APP} = \frac{1}{2}\rho v_S^2 (1 + k_{2_i})_{eq} C_F \sum_i S_{APP_i} + \sum R_{TH} \tag{16}$$

$$R_{TH} = \rho v_S^2 d_{TH}^2 C_{D_{TH}} \tag{17}$$

$$R_{W_a}(Fr) = c_1 c_2 c_5 \rho g V \exp\left[m_1 Fr^d + m_4 \cos(\lambda Fr^{-2})\right] \tag{18}$$

$$R_B = 0.11\rho g(\sqrt{A_{BT}})^3 \frac{Fr_i^3}{1 + Fr_i^2} e^{(-3.0P_B^{-}2)} \tag{19}$$

$$R_{TR} = \frac{1}{2}\rho v_S^2 A_T c_6 \tag{20}$$

$$R_A = \frac{1}{2}\rho v_S^2 C_A (S + \sum S_{APP}) \tag{21}$$

For the wind resistance, the following equation by Blendermann (1994)

Table 2. Coefficients to estimate wind resistance

	CD_t	CD_{lAF}	δ
Car carrier	0.95	0.55	0.8
Cargo ship, container on deck, bridge aft	0.85	0.65/0.55	0.40
Containership, loaded	0.90	0.55	0.40
Ferry	0.90	0.45	0.80
LNG Tanker	0.70	0.60	0.50
Passenger liner	0.90	0.40	0.80
Speed boat	0.90	0.55	0.60
Tanker, loaded	0.70	0.90	0.40
Tanker, in ballast	0.70	0.75	0.40

$$R_{AA} = \frac{\rho_{air}}{2} u^2 A_L C D_l \frac{\cos(\varepsilon)}{1 - \frac{\delta}{2}(1 - \frac{CD_l}{CD_t}\sin^2(2\varepsilon))} \tag{22}$$

$$u = \sqrt{u_{\mathrm{TW}}^2 + v_S^2 + 2 \cdot u_{\mathrm{TW}} \cdot v_S \cdot \cos(\beta)} \tag{23}$$

$$\frac{u_{TW}}{\sin(\varepsilon)} = \frac{u}{\sin(\beta)} \tag{24}$$

$$CD_{lAF} = CD_l \frac{A_L}{A_F} \tag{25}$$

Then the calculation of additional resistance due to wave based on STAWAVE-I is presented in the following equation:

$$R_{AWL} = \frac{1}{16} \rho g H_{1/3}^2 B \sqrt{\frac{B}{L_{BWL}}} \tag{26}$$

3.3.5. Selection and validation of optimal model

To ensure a meaningful assessment of the model's performance and its accuracy, the k-fold cross-validation technique will be employed. K-fold cross-validation involves partitioning the training set into k subsets, referred to as folds. The model will then undergo k training iterations, with each iteration using k-1 folds for training and the remaining fold for validation. During each iteration, the model's performance will be evaluated using various metrics, including the Coefficient of Determination (R^2), Explained Variance (EV), Mean Absolute Error (MAE), Root Mean Square Error (RMSE), Median Absolute Deviation (MAD), and Mean Absolute Percentage Error (MAPE). The results from each iteration will be averaged, providing an assessment of the model's accuracy, which can be further understood by considering the standard deviation. The utilization of k-fold cross-validation facilitates the evaluation of the model's robustness across different datasets.

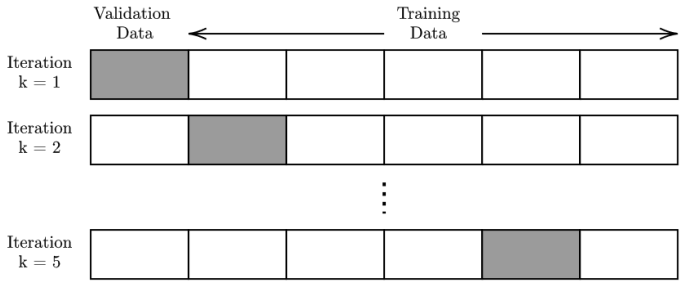

Figure 8. Visual illustration of k-folding, Grey shaded box represents the validation data while white box represents the training data

$$R^2(y, \hat{y}) = 1 - \frac{\sum_{i=1}^{n}(y_\mathrm{i} - \hat{y}_\mathrm{i})^2}{\sum_{i=1}^{n}(y_\mathrm{i} - \overline{y}_\mathrm{i})^2} \quad \textbf{where} \quad \overline{y} = \frac{1}{n}\sum_{1}^{n} y_\mathrm{i} \tag{27}$$

$$EV(y, \hat{y}) = 1 - \frac{\sigma^2_{(y-\hat{y})}}{\sigma^2_y} \tag{28}$$

$$MAE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n} |y_\mathrm{i} - \hat{y}_\mathrm{i}| \tag{29}$$

$$RMSE(y, \hat{y}) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_\mathrm{i} - \hat{y}_\mathrm{i})^2} \tag{30}$$

$$MAD(y, \hat{y}) = \mathrm{median}(|y_1 - \hat{y}_1|, \ldots, |y_\mathrm{i} - \hat{y}_\mathrm{i}|) \tag{31}$$

$$MAPE(y, \hat{y}) = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_\mathrm{i} - \hat{y}_\mathrm{i}}{y_i}\right| \cdot 100\% \tag{32}$$

4. Methodology Application

For further clarity regarding the methodology, the following steps are taken which are based on the proposed methodology. For generation of the BBM, the steps taken are:

(1) Dataset is loaded.
(2) Identify and remove any anomalies.
(3) Remove static and unneeded features.

Table 3. Optimal hyperparameter with training time of each model

Model	Training time [s]	Optimal Hyperparameter	Search Range
RFR	4.112	None	
RFR_{OPT}	3.431	`min_samples_split = 2`	[2,10]
		`min_samples_leaf = 1`	[1,10]
		`max_features = 10`	[6,12]
		`max_depth = 120`	[10,200] and [None]
		`n_estimators = 100`	[100,1000]

(4) Apply speed threshold of 5 knots.
(5) Highly correlated features are combined/removed based on physical and statistical reasoning.
(6) Impute missing values using `KNNImputer`.
(7) Split the dataset into training and testing.
(8) Train the model using the whole dataset with default hyperparameter.
(9) Evaluate model performance using k-fold cross-validation.
(10) Tune the model until the best model is obtained.
(11) For the case study, the best models will be used to predict the SOG using the test dataset.

Subsequently, for FOC calculation, the following steps are taken:

(1) The test dataset is split into seasonal data. Summer-Fall season and Winter-Spring season corresponding to data for 6 months respectively.
(2) Impute missing values using `KNNImputer`.
(3) SOG is converted to STW.
(4) Calculate calm water resistance R_{CALM}.
(5) Calculate added resistance due to wave R_{AW}.
(6) Calculate added resistance due to wind R_{AA}.
(7) Calculate total effective power P_E using total resistance R_{TOTAL}.
(8) Calculate brake power P_B from total efficiencies.
(9) Plot resulting regression line for Power-Speed curve from all models and actual case.
(10) Calculate the FOC by considering the engine SFOC and operation time.
(11) Plot resulting regression line for FOC-Speed curve from all models and actual case.
(12) Evaluate the performance of the model generated from the regression lines.

The dataset used in the case study will represent the journey of the ferry between Køge and Rønne. After data preprocessing and cleaning. There are 3828 data points. The dataset is divided into training and test datasets with a ratio of 75:25 for training and testing, respectively. This results in 2871 data points for training and 957 data points for testing.

5. Result and Discussion

5.1. *Model Optimisation*

From `GridSearchCV`, it cam be observed that the most optimal model shown in Figure 3 can reduce the training time. This is most notably caused by limiting the depth of the tree, from `max_depth = None` to `max_depth = 100`.

Figure 4 shows that the process of hyperparameter tuning for the Random Forest

Table 4. Cross Validation results of Random Forest (RF) Regressor

Model		RFR	RFR_{OPT}
R^2	[%]	89.17	89.46
expVar	[%]	89.21	89.50
MAE	[kn]	0.656	9.649
RMSE	[kn]	1.015	1.008
MAD	[kn]	0.446	0.445

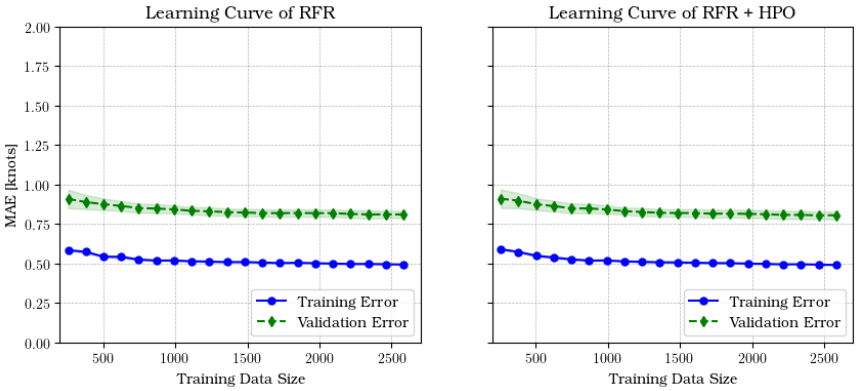

Figure 9. Correlation Heat Map

Regressor (RFR) model did not show any significant improvement in model performance. This outcome aligns with the findings of Kuhn and Johnson (2013) and Hastie, Tibshirani, and Friedman (2009).

References

Abebe, Misganaw, Yongwoo Shin, Yoojeong Noh, Sangbong Lee, and Inwon Lee. 2020. "Machine Learning Approaches for Ship Speed Prediction towards Energy Efficient Shipping." *Applied Sciences* 10 (7): 2325.

Bergstra, J, and Yoshua Bengio. 2012. "Random Search for Hyper-Parameter Optimization." *J. Mach. Learn. Res.* .

Bialystocki, Nicolas, and Dimitris Konovessis. 2016. "On the estimation of ship's fuel consumption and speed curve: A statistical approach." *Journal of Ocean Engineering and Science* 1 (2): 157–166.

Birk, Lothar. 2019. *Fundamentals of ship hydrodynamics: Fluid mechanics, ship resistance and propulsion / Lothar Birk*. 1st ed. Hoboken, New Jersey: John Wiley & Sons.

Bitner-Gregersen, Elzbieta M. 2005. "Joint Probabilistic Description for Combined Seas." In *24th International Conference on Offshore Mechanics and Arctic Engineering: Volume 2*, 169–180. ASMEDC.

Blendermann, Werner. 1994. "Parameter identification of wind loads on ships." *Journal of Wind Engineering and Industrial Aerodynamics* 51 (3): 339–351.

Breiman, Leo. 2001. "Random Forests." *Machine Learning* 45 (1): 5–32.

Coraddu, Andrea, Luca Oneto, Francesco Baldi, and Davide Anguita. 2017. "Vessels fuel consumption forecast and trim optimisation: A data analytics perspective." *Ocean Engineering* 130: 351–370.

Dietterich, Thomas G. 2000. "An Experimental Comparison of Three Methods for Constructing Ensembles of Decision Trees: Bagging, Boosting, and Randomization." *Machine Learning* 40 (2): 139–157.

Du, Yuquan, Qiang Meng, Shuaian Wang, and Haibo Kuang. 2019. "Two-phase optimal solu-

tions for ship speed and trim optimization over a voyage using voyage report data." *Transportation Research Part B: Methodological* 122: 88–114.

Gkerekos, Christos, Iraklis Lazakis, and Gerasimos Theotokatos. 2019. "Machine learning models for predicting ship main engine Fuel Oil Consumption: A comparative study." *Ocean Engineering* 188: 106282.

Halevy, Alon, Peter Norvig, and Fernando Pereira. 2009. "The Unreasonable Effectiveness of Data." *IEEE Intelligent Systems* 24 (2): 8–12.

Haranen, Michael, Pekka Pakkanen, Risto Kariranta, and Jouni Salo. 2016. "White, grey and black-box modelling in ship performance evaluation." In *1st Hull performence & insight conference (HullPIC)*, 115–127.

Hastie, Trevor, Robert Tibshirani, and J. H. Friedman. 2009. *The elements of statistical learning: Data mining, inference, and prediction / Trevor Hastie, Robert Tibshirani, Jerome Friedman.* 2nd ed., Springer series in statistics. New York: Springer.

Holtrop, J. 1984. "A statistical re-analysis of resistance and propulsion data." *Published in International Shipbuilding Progress, ISP, Volume 31, Number 363* .

Holtrop, J., and G.G.J. Mennen. 1978. "A statistical power prediction method." *Netherlands Ship Model Basin, NSMB, Wageningen, Publication No. 603, Published in: International Shipbuilding Progress, ISP, Volume 25, Number 290, October 1978* .

Holtrop, J., and G.G.J. Mennen. 1982. "An approximate power prediction method." *Netherlands Ship Model Basin, NSMB, Wageningen, Publication No. 689, Published in: International Shipbuilding Progress, ISP, Volume 29, Nr 335, 1982* .

IMO. 2020. "Fourth IMO GHG Study 2020." *International Maritime Organization London, UK* .

Kim, Seong-Hoon, Myung-Il Roh, Min-Jae Oh, Sung-Woo Park, and In-Il Kim. 2020. "Estimation of ship operational efficiency from AIS data using big data technology." *International Journal of Naval Architecture and Ocean Engineering* 12: 440–454.

Kuhn, Max, and Kjell Johnson. 2013. *Applied predictive modeling.* New York: Springer.

Lang, Xiao. 2020. "Development of Speed-power Performance Models for Ship Voyage Optimization." PhD diss.

Li, Xiaohe, Yuquan Du, Yanyu Chen, Son Nguyen, Wei Zhang, Alessandro Schönborn, and Zhuo Sun. 2022. "Data fusion and machine learning for ship fuel efficiency modeling: Part I – Voyage report data and meteorological data." *Communications in Transportation Research* 2: 100074.

Pedregosa, Fabian, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, et al. 2011. "Scikit-learn: Machine learning in Python." *the Journal of machine Learning research* 12: 2825–2830.

Petersen, Jóan Petur, Daniel J. Jacobsen, and Ole Winther. 2012. "Statistical modelling for ship propulsion efficiency." *Journal of Marine Science and Technology* 17 (1): 30–39.

Rakke, Stian Glomvik. 2016. "Ship emissions calculation from AIS." .

Smith, T.W.P., J.P. Jalkanen, B.A. Anderson, J. J. Corbett, J. Faber, S. Hanayama, E. O'Keeffe, et al. 2015. "Third IMO Greenhouse Gas Study 2014." .

Soner, Omer, Emre Akyuz, and Metin Celik. 2018. "Use of tree based methods in ship performance monitoring under operating conditions." *Ocean Engineering* 166: 302–310.

Wang, Shuaian, and Qiang Meng. 2012. "Sailing speed optimization for container ships in a liner shipping network." *Transportation Research Part E: Logistics and Transportation Review* 48 (3): 701–714.

Yan, Ran, Shuaian Wang, and Yuquan Du. 2020. "Development of a two-stage ship fuel consumption prediction and reduction model for a dry bulk ship." *Transportation Research Part E: Logistics and Transportation Review* 138: 101930.

Yan, Ran, Shuaian Wang, and Harilaos N. Psaraftis. 2021. "Data analytics for fuel consumption management in maritime transportation: Status and perspectives." *Transportation Research Part E: Logistics and Transportation Review* 155: 102489.