http://dx.doi.org/10.1016/j.trc.2023.104361

Innovative Approaches to Addressing the Tradeoff Between Interpretability and Accuracy in Ship Fuel Consumption Prediction

Haoqing Wang
Department of Logistics and Maritime Studies, Faculty of Business, The Hong Kong Polytechnic University, Hong Kong
haoqing.wang@connect.polyu.hk

Ran Yan*
School of Civil and Environmental Engineering, Nanyang Technological University, Singapore angel-ran.yan@connect.polyu.hk

Shuaian Wang
Department of Logistics and Maritime Studies, Faculty of Business, The Hong Kong Polytechnic University, Hong Kong
hans.wang@polyu.edu.hk

Lu Zhen
School of Management, Shanghai University, Shanghai, China lzhen@shu.edu.cn

Ship fuel consumption is a major component of maritime transport costs and most of its emissions are harmful to the environment. Hence, it is essential to build an accurate ship fuel consumption prediction model, thereby providing reference to the navigation operations. However, maritime industry experts are wary of advanced black-box models since they cannot interpret the outcomes of these models. The application of advanced black-box models in the shipping industry remains limited and it is necessary to develop both accurate and interpretable ship fuel consumption prediction models. This study uses domain knowledge to develop two innovative methods for predicting ship fuel consumption—the first is a physics-informed neural network (PI-NN) model that improves the interpretability of the black-box model while maintaining accuracy and the second is a mixed-integer quadratic optimization (MIO) model that considers more forms of feature variable expressions in an additive white-box model. The proposed approaches address the tradeoff between model interpretability and model accuracy in ship fuel consumption prediction. The experiment results demonstrate that the PI-NN model improves the interpretability of the black-box model while preserving accuracy. The MIO model considers alternative variable expressions, leading to the flexibility of the white-box model. Finally, SHapley Additive exPlanations (SHAP) is used to explain how each feature value contributes to the predictions of the black-box model, thereby providing insights into how each value of feature variables affects fuel consumption. This study provides a solution to the tradeoff between model interpretability and model accuracy and can promote the application of data-driven models in ship fuel consumption prediction. Moreover, this study gives implications for the application of explainable machine learning models in practice.

Keywords: maritime transport; interpretable machine learning models; ship fuel consumption prediction; domain knowledge in shipping; mixed-integer quadratic optimization

1. Introduction

Maritime transport carries over 80% of global trade and is the foundation of the global supply chain (UNCTAD 2022). However, ship navigation brings severe pollution to the environment as ships are mainly driven by heavy oil (Wang et al. 2022a). In order to promote the sustainable development of the shipping industry, governments and scholars all pay attention to improving ship fuel consumption efficiency and thus realizing green shipping (IMO 2020, Meng et al. 2016, Fagerholt et al. 2015, Wang et al. 2018). For example, the International Maritime Organ (IMO) has promulgated a series of regulations to help achieve green shipping, such as the global sulfur content limit in fuel (IMO 2022). Ship fuel consumption is a hot, important, and ongoing research topic in the field of maritime studies (Yang et al. 2019) because fuel cost dominates the costs of a ship (Meng et al. 2016) and generates emissions, which in turn affect sustainability (Wang et al. 2022a). Academic studies on ship fuel consumption abound and developing models to predict ship fuel consumption is a key research topic (Yan et al. 2021a, Fan et al. 2022). Some literature proposes advanced models to deliver accurate ship fuel consumption prediction. However, these advanced models in literature may be hard to be implemented in practice because they are difficult to interpret and thus experts are wary of relying on these models since shipping is a traditional industry (Yan et al. 2022) and domain knowledge plays an important role in decision-making. Therefore, developing interpretable ship fuel consumption prediction models using domain knowledge is urgently needed.

Ship fuel consumption prediction models in the literature are typically categorized as either black-box or white-box models, following the classification provided in Loyola-Gonzalez (2019). Black-box models are models based on hyperplanes (e.g., support vector machines), biological neural networks such as those of animal brains (e.g., artificial neural networks), and probabilistic and combinatory logic (e.g., probabilistic logic networks) or models with local approximation functions (e.g., k-nearest neighbors) (Loyola-Gonzalez 2019). White-box models are self-explanatory and do not require an additional model to explain the results (Loyola-Gonzalez 2019). For example, linear regression is a typical white-box model. Using historical data, linear regression can yield an exact expression of the relationships between the feature variables and the estimated values. Moreover, the coefficients of the feature variables in the expressions show how these feature variables affect the outcome. White-box models are favored in practice because of their interpretability; black-box models usually have better prediction performance than white-box models (Parkes et al. 2018) but are less interpretable. Many recent studies explore the use of explainable artificial intelligence (XAI) for improving the explainability of black-box models (Lundberg et al. 2017, Ribeiro et al. 2016, Gunning et al. 2019). One way of achieving XAI is to develop an additional white-box model to interpret the results of a black-box model (Lundberg et al. 2017, Ribeiro et al. 2016, Sundararajan and Najmi 2020). That is, a black-box model is first presented and then a white-box model is used

to explain it; thus, the process involves ex-post explainability. This study is not limited to ex-post explainability. Instead, explainability or interpretability in this study indicates that the model is interpretable by itself or can be explained using a white-box model. Note that this study does not distinguish between the specific definitions of explainability and interpretability as there is no universal consensus on the definitions of either term (Doshi-Velez and Kim 2017). The scope of explainability or interpretability in this study is in line with the definition in Doshi-Velez and Kim (2017)—"to explain or to present in understandable terms".

1.1. Literature review

We review two streams of literature that are closely related to our research: i) ship fuel consumption prediction models; and ii) interpretable models in maritime studies.

1.1.1. Ship fuel consumption prediction models. Fuel consumption is a critical factor in ship routing as it generates high costs and environmental pollution. There are many studies that take into account ship fuel consumption when optimizing ship sailing speed (Fagerholt et al. 2010). In recent years, with the development of informatization and the accumulation of data, an increasing number of studies have recognized the new insights that data can provide for fuel consumption prediction. For example, Du et al. (2019) study the ship sailing speed and trim optimization problem. They first predict fuel consumption rates using a neural network model based on the noon report data and then optimize the speed of the shipping route by the dynamic programming algorithm. Utilizing historical ship voyage data to develop more accurate fuel consumption prediction models, thereby optimizing vessel operations, has become a significant topic in maritime research (Yan et al. 2021a).

The literature on ship fuel consumption prediction models is divided into black-box models and white-box models, and details are shown in Table 1. In short, the critical issue in ship fuel consumption prediction is to obtain the influencing variables and then develop an accurate ship fuel consumption prediction model based on these variables. The first stream of studies favors black-box models (Du et al. 2019, Yan et al. 2020, Le et al. 2020b), which usually outperform white-box models (Le et al. 2020b, Ma et al. 2023). However, the results of black-box models are hard to interpret. Sometimes, even experts in the maritime industry struggle to explain the outcomes of black-box models, and managers in shipping companies may consider it risky to apply models with low interpretability in practice. The second stream of studies presents statistic models which are white-box models with high interpretability. White-box models have advantages in exploring the explicit relationship between ship fuel consumption and its influencing factors (Wang and Meng 2012, Meng et al. 2016, Adland et al. 2020, Le et al. 2020a). But they usually cannot capture

Table 1 **Literature on ship fuel consumption prediction**

f2

Literature[1]	White-box model	Black-box model	Contents and findings
Wang and Meng (2012)	✓		There is a power function relationship between ship sailing speed and fuel consumption, and the power of sailing speed is between 2.7 and 3.3.
Meng et al. (2016)	✓		Analyze the relationship between fuel consumption and its influencing factors by Spearman's rank correlation coefficients.
Du et al. (2019)		✓	Develop neural network models to predict ship fuel consumption and optimize the ship sailing speed dynamically.
Yan et al. (2020)		✓	Adopt random forest to predict ship fuel consumption and optimize the ship sailing speed based on the predicted results.
Adland et al. (2020)	✓		Estimate the ship fuel consumption-speed curve and doubt the slow-steaming strategy based on empirical findings.
Le et al. (2020a)	✓		Adopt a linear regression model to predict ship fuel consumption.
Le et al. (2020b)	✓	✓	Develop a black-box multilayer perceptron artificial neural network (MLP) to predict ship fuel consumption and compare its prediction performance with two white-box multiple-regression models, showing the effectiveness of the MLP model.
Ma et al. (2023)	✓	✓	Develop both white-box model and black-box to predict ship fuel consumption and find that the white-box model has poor performance.
Uyanık et al. (2023)		✓	Develop decision tree model and neural network model to predict ship fuel consumption. The neural network model is proven to be more effective than the decision tree model.

[1] Note that studies on ship fuel consumption prediction are not limited to those listed in Table 1. As there are literature reviews on ship fuel consumption, this study does not go through it in a detailed way. Readers are referred to Yan et al. (2021a) and Fan et al. (2022) and the references therein.

complex interactions among feature variables and thus the prediction performance of white-box models is usually not as good as black-box models.

Therefore, there is a tradeoff in predicting ship fuel consumption: white-box models provide high interpretability but poor prediction performance, and black-box models provide low interpretability but good prediction performance. In practice, both model interpretability and accuracy are

important (Carvalho et al. 2019, Loyola-Gonzalez 2019). However, the literature does not address the tradeoff between interpretability and accuracy on the ship fuel consumption prediction problem. Studies on ship fuel consumption do not consider how to improve the interpretability of black-box models using constraints based on domain knowledge. Moreover, the literature mainly uses off-the-shelf white-box models and does not consider using domain knowledge available in the shipping field to develop more flexible white-box models by expanding the forms of feature variable expressions. Therefore, a theoretical solution to this tradeoff is urgently needed.

1.1.2. Interpretable models in maritime studies. A few recent studies focus on the interpretability of black-box models in the maritime domain. Kim and Lim (2022) propose machine learning models for predicting maritime accidents and develop additive white-box models based on SHAP (SHapley Additive exPlanations) and LIME (Local Interpretable Model-agnostic Explanations) to interpret the predicted values. Zhang et al. (2022) also adopt SHAP to interpret the outcomes of tree-based machine learning models in predicting maritime accidents. Veerappa et al. (2020) use SHAP to explore the internal mechanisms for classifying the types of ships. He et al. (2021) identify the factors affecting ship detention using SHAP. Yan et al. (2022) adopt SHAP to provide explanations for the port state control problem. All of these studies develop posterior explanatory models—they first build a black-box machine learning model and then develop an additive white-box model based on SHAP to interpret the results of the black-box model. In detail, SHAP assigns an important value to each feature value for all the predictions (Kim and Lim 2021). By aggregating the important values of all feature values in a sample and the mean value of the predicted target in the training dataset, SHAP develops an additive white-box model to explain the predicted values of each sample (Lundberg et al. 2017). Therefore, SHAP provides insights into how the predicted values are obtained in a black-box model and thus is suitable for ex-post explanations. In this study, after proposing two self-explanatory models, SHAP is adopted to interpret the results of an off-of-shelf black-box model. SHAP can interpret how the black-box model arrives at each predicted value of ship fuel consumption and how ship fuel consumption is affected by feature variables of different values. To the best of our knowledge, this study is the first to explore the issue of interpretability in predicting ship fuel consumption, with the aim of providing a comprehensive solution to the tradeoff between interpretability and accuracy.

Based on domain knowledge, this study proposes two approaches to using domain knowledge to address the tradeoff between interpretability and accuracy in predicting ship fuel consumption. This study differs from earlier studies on ship fuel consumption prediction in several ways. First, a black-box model is developed for predicting ship fuel consumption that uses the physics constraints identified in domain knowledge to improve the model's interpretability. Second, by considering

different forms of feature variable expressions from the perspective of domain knowledge, a mixed-integer quadratic optimization (MIO) model is solved to fit a linear regression model, which is an additive white-box model with a high level of flexibility. Last, SHAP is adopted to identify how black-box models yield predicted values and how a change in a feature value affects the predicted ship fuel consumption.

1.2. Objectives and research questions

Two innovative approaches are developed for balancing model interpretability and model accuracy. The first method applies the constraints of physics (i.e., domain knowledge) to construct a neural network model. This model is named the physics-informed neural network (PI-NN) model, which is on par with the fully-connected neural network (fully-NN) model in terms of performance but has high interpretability because the PI-NN model uses domain knowledge to make the neural network model more intuitive. The second method uses an MIO model to fit a linear regression model by selecting the best form of variable expressions that influence ship fuel consumption. This model is called the MIO model for solving the best forms (BF) of variable expressions (MIO-BF). The MIO-BF model is an additive white-box model that aims to give the best linear regression formula. The two proposed models provide solutions to the tradeoff between interpretability and flexibility (i.e., accuracy). The PI-NN model helps improve the interpretability of black-box models while preserving prediction performance, and the MIO-BF model allows the statistical models to consider more forms of variable expressions while maintaining explainability. Moreover, by solving the MIO-BF model, this study gains insights into the forms in which the variables affect fuel consumption and yield the optimal linear regression model at the same time. The performance of the MIO-BF model may be slightly poorer than that of the PI-NN model, but it is highly explainable.

By building the two models, this study answers the following three research questions.

Q1: To what extent can the PI-NN model explain the fully-NN model? That is, is there a way to build a convincing neural network model to predict ship fuel consumption using domain knowledge that improves model interpretability while maintaining model accuracy?

Q2: In what forms do feature variables affect fuel consumption? That is, what relationship does the MIO-BF model obtain between the feature variables and ship fuel consumption? Is the obtained relationship explainable in practice?

Q3: What are the differences among the MIO-BF, the PI-NN, and other artificial intelligence (AI) models (e.g., fully-NN) in terms of prediction performance?

In addition to the two proposed self-interpretable models, this study also uses SHAP for the posterior explanation of the black-box model. SHAP directly shows how the predicted values of ship fuel consumption are obtained from different feature values and offers an additive white-box

model. Unlike the MIO-BF model, SHAP has to be developed after the machine learning model is used. Thus, the MIO-BF model, the PI-NN model, and SHAP explore the issue of interpretability in ship fuel consumption prediction from different perspectives.

1.3. Innovation and contributions

The theoretical and practical contributions of this research are summarized as follows.

Theoretical contributions. This study presents significant theoretical contributions in the context of the ship fuel consumption prediction problem by introducing two innovative models: the PI-NN model and the MIO-BF model. First, applying the knowledge of physics to reconstruct a neural network model is an innovation in ship fuel consumption prediction. By dissecting the neural network into two components, one addressing air resistance and the other water resistance, interpretability is notably enhanced without compromising model accuracy. Empirical experiments confirm the effectiveness of the PI-NN model, underscoring the improvement in interpretability for black-box models. Second, the MIO-BF model yields an optimal additive model by solving the MIO programming. Unlike other additive models, MIO-BF accommodates a wider range of feature variable expressions, thereby enhancing the flexibility of white-box models. Lastly, the application of SHAP for interpreting machine learning models, while not novel in itself, marks the rare instance of SHAP being used to explain ship fuel consumption predictions. These results emphasize the paramount role of ship sailing speed in fuel consumption, particularly in operational routes where the impact of wind is less pronounced.

Practical contributions. This study also offers substantial practical contributions. First, it provides practical solutions to the trade-off between model interpretability and accuracy. Practitioners and managers can leverage their domain knowledge to enhance the interpretability of black-box models or opt for the MIO-BF model, which is more flexible and considers various forms of feature variable expressions. Second, the findings establish clear relationships between feature variables and ship fuel consumption, offering invaluable insights into the key determinants of fuel consumption. Lastly, by enhancing the interpretability of black-box models, this study encourages the adoption of such models in practice. These advanced models exhibit high predictive performance and have the potential to reduce vessel emissions, thus benefiting both industry and the environment.

In summary, this study delivers comprehensive insights into the crucial issue of interpretability versus accuracy in ship fuel consumption prediction, resulting in significant theoretical and practical contributions. The introduction of innovative models, the demonstration of how domain knowledge can strike a balance between interpretability and accuracy, and the practical implications for other industries underscore the far-reaching impact of this research. Given the limited focus on model

interpretability in previous studies, this research has the potential to promote the application of advanced, interpretable black-box models, ultimately leading to increased industry profitability and reduced environmental impact.

The remainder of this paper is organized as follows. Section 2 develops methodologies: the PI-NN model and the additive MIO-BF model. Section 3 introduces the dataset used in the experiments. Section 4 illustrates the settings for methods and shows the results. Section 4 makes a further analysis based on the results. Section 5 provides an ex-post way to understand the prediction results of machine learning models. Conclusions are presented in Section 6.

2. Methodologies

As shown in Figure 1, James et al. (2013) provide an illustration of interpretability and flexibility. Flexibility refers to the degree to which a model can capture different forms among the feature variables. For example, the linear regression model is restrictive as it can generate only a linear function between the input variables and the output. However, the linear regression model is easy to understand. In general, white-box models are highly interpretable but inflexible, whereas black-box models are flexible because they can capture complex relationships between the inputs and the output. Note that there is no clear metric in the literature for measuring model flexibility; hence, many studies use accuracy as an alternative (James et al. 2013, Gunning et al. 2019). This study argues that flexibility and accuracy are complementary—highly flexible models perform better as they can capture more complex relationships between the feature variables and the output. Therefore, this study does not strictly delineate accuracy and flexibility.

Figure 1 shows that the PI-NN model moves to a point of higher interpretability from the class of deep learning models. As the physics constraints are added to the neural network model, the flexibility of the PI-NN model decreases. The MIO-BF model shifts from least squares to a point of higher flexibility because more forms of feature variable expressions are considered. However, the interpretability of the MIO-BF model decreases slightly as it changes the original values of the feature variables. Therefore, this study provides two options for addressing the tradeoff between model interpretability and model accuracy in the ship fuel consumption prediction problem. The proposed PI-NN model makes the black-box model more interpretable while preserving accuracy. The MIO-BF model considers more forms of the feature variable expressions while developing an explainable additive model. We next introduce the two models in detail.

2.1. PI-NN model for ship fuel consumption prediction

The vector of feature variables is denoted by $\boldsymbol{x}$. When predicting ship fuel consumption, prevailing methods develop a model to solve the function:

$$y = F(\boldsymbol{x}), \tag{1}$$

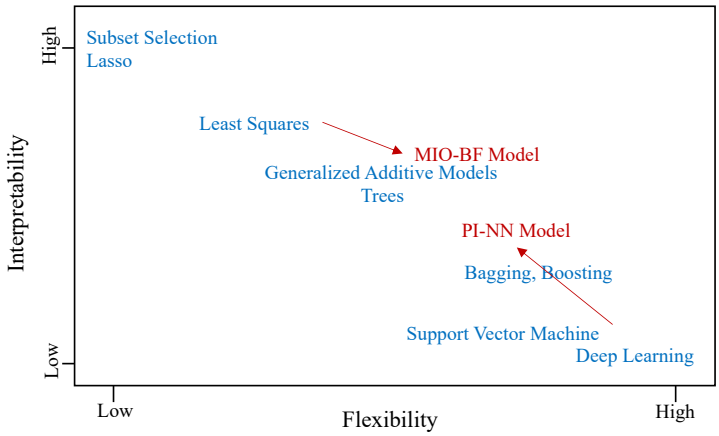

Figure 1 Tradeoff between interpretability and flexibility (excerpted from James et al. (2013) page 25 and adapted)

where y kg/s represents the fuel consumption per second (the unit of ship fuel consumption can be changed according to the recorded data). Note that it may be difficult to give explicit expressions for some complex black-box models. Here Formula (1) is just adopted to emphasize that the current research usually directly inputs feature variables into the model without additional constraints based on domain knowledge.

From a physical point of view, oil consumed by ships is used to produce energy; specifically, burning oil is the process of converting chemical energy into internal energy and then converting internal energy into mechanical energy. The mechanical energy generated mainly overcomes the resistance of water and air and is finally converted into internal energy. According to Newton's Third law, a ship gains thrust and thus velocity. $E(y)$ denotes the energy generated by y kg/s fuel. Thus, the following formula can be formulated:

$$E(y) \propto E_A(\boldsymbol{x}_A) + E_W(\boldsymbol{x}_W), \tag{2}$$

where $E_A(\boldsymbol{x}_A)$ and $E_W(\boldsymbol{x}_W)$ are the energy used to overcome air resistance and water resistance, respectively, and $\boldsymbol{x}_A$ and $\boldsymbol{x}_W$ are the vectors of variables that affect air and water resistance, respectively. Therefore, from the perspective of domain knowledge or physics, different kinds of variables could be distinguished, i.e., $\boldsymbol{x}_A$ and $\boldsymbol{x}_W$, when building the fuel consumption prediction model to improve model interpretability and persuasiveness. Raissi et al. (2020) develop a physics-informed deep-learning framework that takes the Navier-Stokes equations into account. They add a hidden layer to capture the Navier-Stokes equations. Motivated by their research, this study proposes the PI-NN model that restricts the relationship shown in Formula (2).

The structure of the proposed PI-NN model is shown in Figure 2. The input layer consists of three types of variables: variables $\boldsymbol{x}_{A'}$ that only affect E_A, variables $\boldsymbol{x}_{W'}$ that only affect E_W, and

variables $\boldsymbol{x}_{AW}$ that affect both E_A and E_W. For example, wind speed affects E_A (Meng et al. 2016), ocean current affects E_W (Chang et al. 2013), and draft affects both E_A and E_W (Rakke et al. 2012). The history data of $\boldsymbol{x}_{A'}$ and $\boldsymbol{x}_{W'}$ will input to the separated two parts of hidden layers in Figure 2. And the history data of $\boldsymbol{x}_{AW}$ will be input to all hidden layers. The neurons of the last layer in the two split hidden layer parts will be connected to two different neurons, respectively. The values of these two neurons after the activation function f are denoted by $f(\hat{y}_A)$ and $f(\hat{y}_W)$, which represent the fuel consumed by the ship to overcome air resistance and water resistance, respectively. According to the above-mentioned physics (see Formula (2)), the sum of $f(\hat{y}_A)$ and $f(\hat{y}_W)$ should be the predicted value of ship fuel consumption, denoted by $\hat{y}$. The structure indicates that the PI-NN model prunes a fully connected neural network model based on domain knowledge. Specifically, by adopting domain knowledge, this study trains two neural network models in the hidden layer level and finally combines the output using an equation provided by physics constraint. The input variables of the two separated neural network models are classified by domain knowledge, which is also the pruning of the neural network model from the input layer. Moreover, if there is only one class of variables in the input layer, the PI-NN model will become an ordinary neural network model. Note that although there are two separate parts of neurons in the hidden layer, the PI-NN is an integrated neural network model as only the final fuel consumption data can be collected in practice.

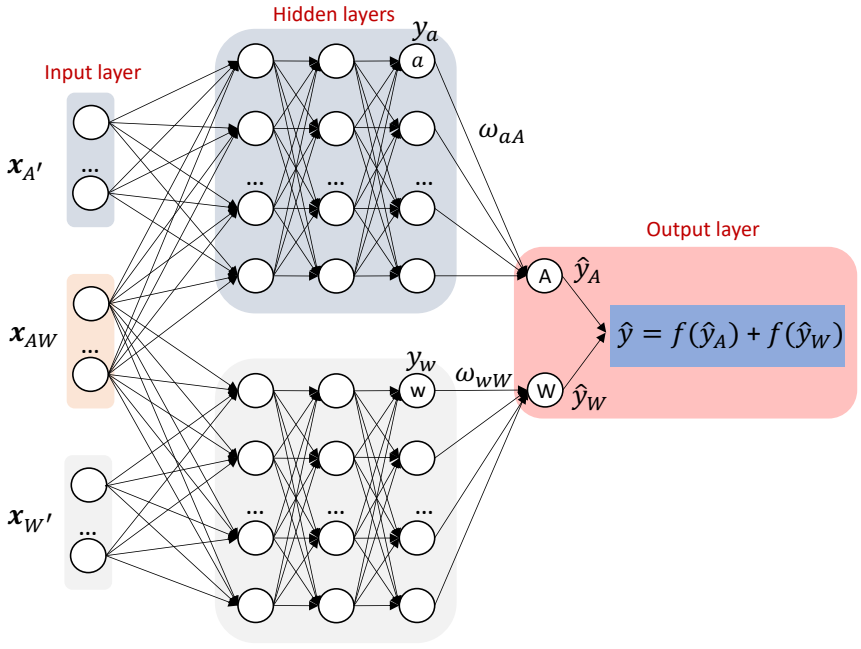

Figure 2 The structure of PI-NN model

According to the principle of the neural network model, the final predicted value $\hat{y}$ is obtained by multiplying input values and the weights of all connected neurons and then summing the values that are calculated by the activation function. The process of calculating the predicted value based on weights is called forward calculation. The key problem in the model training process is to get the optimal weights that connect consecutive neurons. The backpropagation method (McClelland et al. 1986) is used to obtain the optimal weights in the PI-NN model. Mean squared error (MSE) is used as the loss function since this study targets a regression problem:

$$L(y,\hat{y}) = \frac{1}{2}(y-\hat{y})^2 = \frac{1}{2}(y-(f(\hat{y}_A)+f(\hat{y}_W)))^2. \tag{3}$$

Notations a and w denote the neurons in the last hidden layer of the two separated hidden layers, respectively. The weights of connected neurons between the last hidden layer and neurons A and W (see Figure 2) are denoted by ω_{aA} and ω_{wW}, respectively. According to the backpropagation method, the update increments of the weights ω_{aA} and ω_{wW} should be:

$$\begin{aligned}\Delta\omega_{aA} &= -\eta\frac{\partial L(y,\hat{y})}{\partial \omega_{aA}} \\ &= -\eta\frac{\partial L(y,\hat{y})}{\partial f(\hat{y}_A)}\frac{\partial f(\hat{y}_A)}{\partial \omega_{aA}} \\ &= \eta[y-(f(\hat{y}_A)+f(\hat{y}_W))](f^{'}(\hat{y}_A)\times y_a)\end{aligned} \tag{4}$$

$$\begin{aligned}\Delta\omega_{wW} &= -\eta\frac{\partial L(y,\hat{y})}{\partial \omega_{wW}} \\ &= -\eta\frac{\partial L(y,\hat{y})}{\partial f(\hat{y}_W)}\frac{\partial f(\hat{y}_W)}{\partial \omega_{wW}} \\ &= \eta[y-(f(\hat{y}_A)+f(\hat{y}_W))](f^{'}(\hat{y}_W)\times y_w),\end{aligned} \tag{5}$$

where y_a and y_w are the values of the neurons in the last hidden layer of the two separated hidden layers, respectively. And the values of y_a and y_w can be obtained by the forward calculation of the PI-NN model. η denotes the learning rate, which determines the convergence speed of the PI-NN model. The backpropagation process of other neurons is the same as the idea of the Formula (4) and Formula (5). In summary, the PI-NN model first randomly initializes each weight and conducts forward calculation to predict the ship fuel consumption, and then optimizes the weights according to the backpropagation method. For example, the updated values of ω_{aA} and ω_{wW} are $(\omega_{aA}+\Delta\omega_{aA})$ and $(\omega_{wW}+\Delta\omega_{wW})$, respectively. Then, the forward calculation is performed according to the updated weights and the backpropagation process is conducted again to optimize the weights. When the preset number of iterations is reached, the PI-NN model outputs the final predicted value.

The used dataset will be introduced in Section 3. The detailed hyperparameter settings, e.g., the number of neurons in each hidden layer, the learning rate η, and the activation function, will be explained in Section 4. In summary, by introducing domain knowledge, the PI-NN model is proposed to give a solution to the tradeoff between accuracy and interpretability from the perspective of improving the interpretability of black-box models. The PI-NN model may not outperform the fully connected neural network model with the same network structure, but it can explain the black-box model at over 97% level as shown in results in Section 4.

2.2. MIO-BF model for ship fuel consumption prediction

Given that navigators or managers in shipping companies already have domain knowledge and have applied their knowledge and experience to make decisions for many years (Yan et al. 2021b), black-box models are not so widely used in practice in the shipping industry because even experts in the maritime industry struggle to interpret these models and thus hold the opinion that applying black-box models in practice is unreliable (Yan et al. 2022). Models with high interpretability are preferred in practice (Yan et al. 2022). In Section 2.1, the interpretability of black-box models is improved by domain knowledge. However, some white-box models, e.g., linear approximation, may be more prevalent in practice (Yan et al. 2021a) though they may not perform as well as black-box models (Uyanık et al. 2020, Parkes et al. 2018, Le et al. 2020b). Therefore, this section proposes a method to consider different forms of feature variable expressions and thus improve the flexibility of white-box models.

2.2.1. Preliminary. Linear regression is a common choice for developing highly explainable models. However, in some scenarios, there are many feature variables and informative feature variables need to be selected to build regression models that can interpret data accurately with high comprehensibility (Tan et al. 2008). The task of selecting $k, k \leq p$ out of p feature variables in a linear regression model given n observations, is the best subset selection problem (Natarajan 1995). Given predictor matrix $\boldsymbol{X} \in \mathbb{R}^{n \times p}$, response vector $\boldsymbol{Y} \in \mathbb{R}^n$, and regression coefficients $\boldsymbol{\beta} \in \mathbb{R}^p$, the best subset of feature variables can be obtained by solving the following nonconvex problem:

$$\min_{\boldsymbol{\beta}} \frac{1}{2}\|\boldsymbol{Y} - \boldsymbol{X}\boldsymbol{\beta}\|_2^2 \quad \text{subject to} \quad \|\boldsymbol{\beta}\|_0 \leq k, \tag{6}$$

where $\|\boldsymbol{\beta}\|_0 = \sum_{i=1}^{p} \mathbb{I}(\beta_i \neq 0)$ and $\mathbb{I}(\cdot)$ denotes the indicator function. The best subset selection problem is an NP-hard problem (Natarajan 1995). By solving Problem (6), the best k feature variables that interpret the target variable can be obtained.

Research on adopting optimization techniques to solve the best subset selection problem mainly lies in solving a convex approximation of Problem (6) (Bertsimas and King 2016). bo Bertsimas et al. (2016) propose an MIO approach to solve the best subset selection problem. By introducing

binary decision variables that restrict the number of selected feature variables, the solution to the MIO approach will be the solution to the best subset selection problem, i.e., Problem (6). The general MIO formulation is

$$\min_{\boldsymbol{\beta},\boldsymbol{z}} \frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 \tag{7}$$

subject to

$$\sum_{i=1}^{p} z_i \le k \tag{8}$$

$$-Mz_i \le \beta_i \le Mz_i, i=1,...,p \tag{9}$$

$$z_i \in \{0,1\}, i=1,...,p \tag{10}$$

$$\boldsymbol{\beta},\boldsymbol{z} \in \mathbb{R}^p, \tag{11}$$

where z_i is a binary variable and $\sum_{i=1}^{p} z_i$ indicates the number of nonzeros in $\boldsymbol{\beta}$. That is, Constraint (8) ensures that the number of selected feature variables cannot exceed k. M is a constant that satisfies $M \ge \|\hat{\boldsymbol{\beta}}\|_\infty$, where $\hat{\boldsymbol{\beta}}$ is the vector of estimated coefficients. Constraints (9) guarantee that $\beta_i = 0$ if $z_i = 0$. MIO is proven to handle small to moderate instances of the best subset selection problem (Hastie and Mazumder 2020).

2.2.2. Method for solving additive models exactly. Referring to the idea that adopts the MIO model to solve the best subset selection problem, an MIO model is proposed to select the best forms of feature variable expressions. Suppose that there is a set of feature variables $F = \{1,...,|F|\}$, where $f \in F$ indicates the index of feature variables and the total number of indexes is $|F|$. This study considers V different forms of feature variables and use $v = 1,...,V$ to denote each form of expression. Suppose that all the feature variables are continuous variables. Therefore, the model ends up with $|F| \times V$ feature variables. For example, there is a dataset of ship fuel consumption that contains 8 feature variables in the beginning. In addition to the 8 feature variables, the model also considers their logarithmic transformation, exponential transformation, quadratic and cubic transformation. That is, there are $|F| = 8$ feature variables in the beginning, $V = 5$ forms of expressions, and $|F| \times V = 40$ feature variables in the end. The task is to minimize the prediction error by selecting $|F|$ feature variables among $|F| \times V$ feature variables and guarantee that only one form of expression can be selected for the same index of feature variables. Although some heuristic algorithms for solving the best subset selection problem, such as forward stepwise and backward stepwise (Derksen and Keselman 1992, Hastie et al. 2020), can be revised by adding the constraints of selecting one form of expression for the same index of feature variables, they do not guarantee to provide the optimal solutions (Derksen and Keselman 1992). Therefore, MIO programming is adopted to select the best forms of feature variable expression and this study abbreviates the

model as MIO-BF. Instead of original feature variables, the MIO-BF model enables more forms of variables to be extended in a linear regression model and thus improves model flexibility. To the best of our knowledge, this is the first attempt to solve the variable expressions using MIO. The objective of the MIO-BF model is to minimize the MSE to select optimal forms of expressions of feature variables because the additive white-box model refers to linear regression in this study.

[MIO-BF]

$$\min_{\boldsymbol{\beta},\boldsymbol{z}} \frac{1}{2}\|\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{\beta}\|_2^2 \tag{12}$$

subject to

$$\sum_{v=1}^{V} z_f^v = 1, f \in F \tag{13}$$

$$-Mz_f^v \le \beta_f^v \le Mz_f^v, f \in F, v = 1, ..., V \tag{14}$$

$$z_f^v \in \{0,1\}, f \in F, v = 1, ..., V \tag{15}$$

$$\boldsymbol{\beta}, \boldsymbol{z} \in \mathbb{R}^{\mathcal{P}}, \tag{16}$$

where $\mathcal{P}$ denotes the value of $|F| \times V$, and thus $\boldsymbol{X} \in \mathbb{R}^{n\times\mathcal{P}}$, $\boldsymbol{Y} \in \mathbb{R}^n$, and $\boldsymbol{\beta} \in \mathbb{R}^{\mathcal{P}}$. Continuous decision variable β_f^v represents the coefficient value of each feature variable, and binary decision variable z_f^v represents whether a certain form of expression is selected. Constraints (13) ensure that one feature can only be formulated by one form of expression. Thus, the MIO-BF model is a typical MIO model, which can be solved by off-the-shelf optimization solvers.

The MIO-BF model is developed to take different forms of feature variable expressions into account with the aim of maintaining the interpretability of the model and improving its flexibility. The MIO-BF model gives the exact expression between ship fuel consumption and feature variables. That is, it can be known how feature variables affect ship fuel consumption from the MIO-BF model, and then infer how the predicted value of fuel consumption is obtained given certain values of feature variables. Domain knowledge is also helpful in determining the forms of variables. For example, the widely recognized relationship between ship sailing speed and fuel consumption is cubic. Wang and Meng (2012) exactly show that the power of sailing speed is between 2.7 and 3.3 using data from five ships. Therefore, when transforming the variable of shipping sailing speed, the MIO-BF model considers performing the power of 2.7, the power of 2.9, the power of 3.0, the power of 3.1, and the power of 3.3 transformations. Moreover, the obtained optimal expressions of variables also provide insights into the relationship between ship fuel consumption and feature variables in turn. The detailed information on feature variables and their transformations is discussed in Section 4.

In summary, two ways are proposed to trade-off between model accuracy and model interpretability: improve the interpretability of black-box models while maintaining accuracy or provide more

forms of feature variable expressions for white-box models while preserving interpretability. The first approach uses the PI-NN model, which combines domain knowledge and the black-box model to improve interpretability without losing too much accuracy. The second uses the MIO-BF model, which is an explainable white-box model obtained by solving an MIO model. Two solutions are provided for addressing the tradeoff by comparing the performance of the PI-NN model and the MIO-BF model. That is, the PI-NN model is more suitable for cases that require a high level of accuracy whereas the MIO-BF model is more suitable for cases that require a high level of interpretability. To the best of our knowledge, both approaches are innovative in predicting ship fuel consumption.

3. Data

A public dataset[1] of ship fuel consumption provided by Petersen (2012) is used for the experiment. As shown in Figure 3, there is a ferry sailing between Tórshavn and Suðuroy, Faroe islands. The sailing time of one voyage is about 2 hours. Taking advantage of sensors, the dataset in Petersen (2012) records the fuel consumption data and other relevant variables (e.g., port and starboard level measurements, speed through water, and wind speed) of the ferry (Petersen et al. 2012a,b). Readers are referred to Petersen (2012) for a detailed description of the data. The variables used in this study are shown in Table 2. In view of the different sampling frequencies of each sensor and thus the different statistical frequencies of each variable, 10 seconds is chosen as a unit to merge data. Next, this study introduces how to calculate the needed feature variables based on the originally recorded variables in Table 2.

Table 2 Data description

Variable	Description	Units
FD	Fuel density	kg/L
FV	Fuel volume flow rate	L/s
L_{draft}	Port level measurement	m
R_{draft}	Starboard level measurement	m
STW	Speed through water	knot
ρ	The angle of the wind relative to the heading direction of the ship	degrees
WS	Relative wind speed measured by an onboard sensor	m/s

Draft. The ferry is equipped with two level measurement devices on the port side and the starboard side (Petersen 2012). As shown in Figure 4, there is an angle θ between the device and

[1] http://cogsys.imm.dtu.dk/propulsionmodelling/data.html

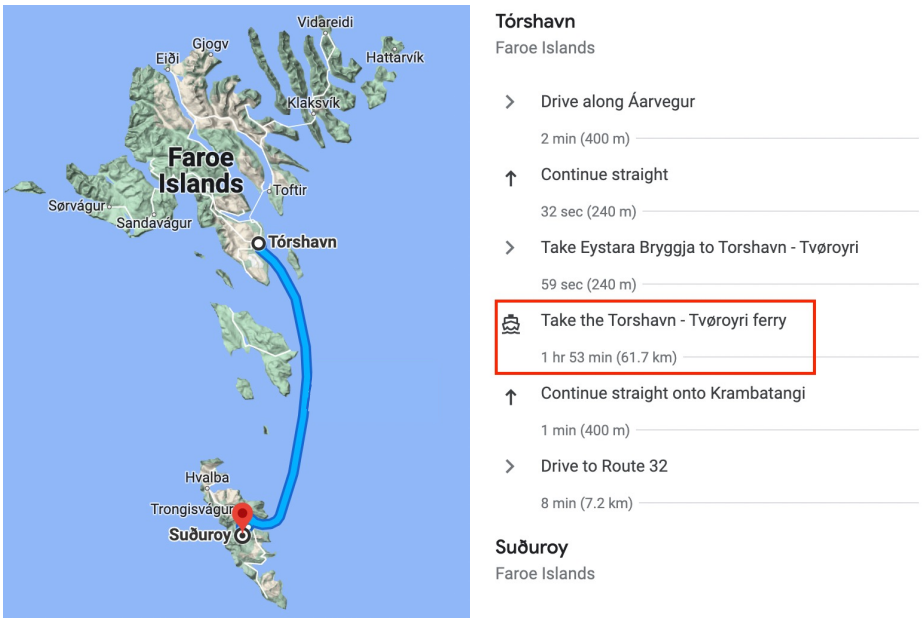

Figure 3 **The shipping line between Tórshavn and Suðuroy**

the hull, and the device detects the distance to sea level (D). Given the vertical distance between the sensor and the bottom of the hull (H), the draft of the ship (d) when the ship is sailing can be calculated by the following equation:

$$d\,(\mathrm{m}) = H - D \times \cos\theta. \tag{17}$$

The installation parameters, i.e., H and θ, of the device are known. For the device on the port side, $H = 19.3$m and $\theta = 19°$ (Petersen 2012). And for the device on the starboard side, $H = 22.1$m and $\theta = 12.6°$ (Petersen 2012). The detected distance D, i.e., L_{draft} and R_{draft}, is recorded by the device. The average of the port and starboard drafts is taken as the ship's draft. Therefore, the final obtained draft of the ship, denoted by d_{avg}, is:

$$d_{\mathrm{avg}}\,(\mathrm{m}) = \frac{(19.3 - L_{\mathrm{draft}} \times \cos 19°) + (22.1 - R_{\mathrm{draft}} \times \cos 12.6°)}{2}. \tag{18}$$

Using knowledge of physics in shipping, d_{avg} is classified as a variable affecting both E_W and E_A because the draft determines the area of the ship in contact with water and air (Rakke et al. 2012), thus affecting the friction from water and air.

Headwind and crosswind. Through relative wind speed WS and the angle of the wind relative to the heading direction of the ship (see Figure 5), the headwind S_{head} and crosswind S_{cross} can be obtained:

$$S_{\mathrm{head}}\,(\mathrm{m/s}) = WS \times \cos\rho \tag{19}$$

$$S_{\mathrm{cross}}\,(\mathrm{m/s}) = WS \times \sin\rho. \tag{20}$$

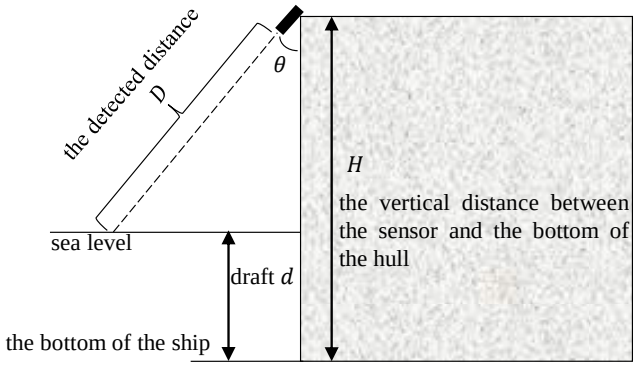

Figure 4 Draft measurement

According to Figure 5 and Formula (19), the negative value of headwind S_{head} represents tailwind and the positive value of headwind S_{head} means that the ship is sailing against the wind. According to Formula (20), the value of crosswind S_{cross} can be positive or negative, which indicates the different directions of the crosswind. Based on domain knowledge in shipping, the direction of the crosswind does not matter because the crosswind is in the vertical direction. Therefore, the absolute value of S_{cross} is used as the basic form of crosswind in the following experiments. Obviously, S_{head} and S_{cross} are variables that affect E_A.

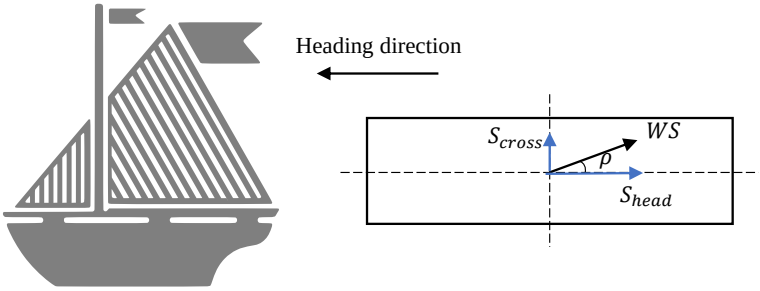

Figure 5 Wind measurement

Speed through the water. The variable STW indicates the ship's sailing speed through the water, which combines the ship's sailing speed over the ground and ocean currents (Petersen 2012). Therefore, this variable can be used directly. STW is classified as a variable affecting E_W because it measures the sailing speed over the ground and ocean currents (Petersen 2012).

Ship fuel consumption. Finally, the fuel consumption can be calculated by the following formula:

$$y\,(\text{kg/s}) = FD \times FV. \tag{21}$$

Outlier records are deleted from the experimental dataset. First, records with null values are deleted. Second, records of ship sailing speed lower than 8 knots are deleted because these records

infer that the ship is at anchor or just begins sailing. Finally, there are 150,831 records containing 245 voyages in the experimental dataset. This study randomly chooses 200 voyages (including 123,243 records) for training and 45 voyages (including 27,588 records) for testing in the experiment. The magnitude of the data used in this study is far greater than that of many other studies that use ship noon reports (one record per day) to predict ship fuel consumption (Wang and Meng 2012, Du et al. 2019, Yan et al. 2020), which makes the results more convincing.

4. Experiment

All experiments are performed on a MacBook Pro computer with an Apple M2 processor (3.5 GHz), 8 cores, and 16 GB of RAM. Gurobi 10.0.0 is used as the optimization solver.

4.1. Settings for the PI-NN model

The optimal parameter settings of the PI-NN model are shown in Table 3. The optimal hyperparameters are obtained through o3 on the training dataset consisting of the randomly selected 200 voyages. The fully-NN model for comparison is equipped with the same network structure and the neuron number between layers is 4-10-10-10-2-1. According to the literature, the feature variables are normalized when training the neural network model (Beşikçi et al. 2016).

Table 3 Parameter settings of the PI-NN model

Parameter	Searching space	Optimal setting
Number of hidden layers	[1,2,3,4]	3
Number of neurons in each hidden layer[1]	[6,8,10,12]	10
Number of neurons in the input layer	\	4
Number of neurons in the output layer	\	2
Activation function	[Relu,Sigmoid]	Sigmoid
Learning rate	[0.001,0.01,0.1]	0.1
Number of epochs	[50,80,100]	100
Number of batch size	[32,64,128]	64

[1] The number of neurons in each hidden layer consists of the number of neurons in each hidden layer of the two separate parts in the PI-NN model. Therefore, 10 neurons indicate that there are 5 neurons in each layer of the separated two parts of hidden layers.

4.2. Settings for the MIO-BF model

For variables d_{avg}, S_{head}, and S_{cross}, their square root transformation, logarithmic transformation, quadratic and cubic transformation are considered. As discussed in Section 3, the absolute value

of S_{cross} is used because the direction of the crosswind does not matter. But the direction of the headwind will affect fuel consumption. Obviously, sailing with a tailwind will save fuel consumption. Therefore, when making transformations, the sign of S_{head} needs to be preserved. Parameter μ is defined to keep the sign

$$\mu = 2 \times \mathbb{I}(S_{\text{head}} > 0) - 1, \tag{22}$$

where $\mathbb{I}(\cdot)$ is an indicator function. If $S_{\text{head}} > 0$, then $\mathbb{I}(S_{\text{head}} > 0) = 1$; if $S_{\text{head}} <= 0$, then $\mathbb{I}(S_{\text{head}} > 0) = 0$. Therefore, $\mu = 1$ if the ship is sailing against the wind and -1 otherwise. It is well-recognized that there is an approximately cubic relationship between ship sailing speed and fuel consumption. Wang and Meng (2012) show that the value of the power is between 2.7 and 3.3. Thus, the expressions of $STW^{2.7}$, $STW^{2.9}$, $STW^{3.0}$, $STW^{3.1}$ and $STW^{3.3}$ is considered. The forms of expressions of feature variables are summarized in Table 4.

Table 4 Forms of expressions of variables

Variable	Expression 1	Expression 2	Expression 3	Expression 4	Expression 5
STW	$STW^{2.7}$	$STW^{2.9}$	$STW^{3.0}$	$STW^{3.1}$	$STW^{3.3}$
d_{avg}	d_{avg}	$\ln(d_{\text{avg}})$	$\sqrt{d_{\text{avg}}}$	d_{avg}^2	d_{avg}^3
S_{head}	S_{head}	$\ln(\lvert S_{\text{head}}\rvert)$	$\mu\sqrt{\lvert S_{\text{head}}\rvert}$	μS_{head}^2	S_{head}^3
S_{cross}	$\lvert S_{\text{cross}}\rvert$	$\ln(\lvert S_{\text{cross}}\rvert)$	$\sqrt{\lvert S_{\text{cross}}\rvert}$	S_{cross}^2	$\lvert S_{\text{cross}}\rvert^3$

4.3. Results and discussion

The performance of the PI-NN model, the fully-NN model, and the MIO-BF model are shown in Table 5. To illustrate the universality of the neural network model, a tree-based machine learning model—XGBoost model (XGB)—is also applied for comparison (Chen and Guestrin 2016). And the hyperparameters of XGB are tuned by GridSearchCV (Yan et al. 2021b) the finally adopted hyperparameters are shown in Table 6. The mean absolute error (MAE) and MSE are used to measure the accuracy of the models. The variances of absolute value difference between predicted fuel consumption and real fuel consumption are calculated, providing a measure of the stability of the model performance:

$$Var = \frac{\sum_{n=1}^{N}(|y_n - \hat{y_n}| - (\frac{\sum_{n=1}^{N}|y_n - \hat{y_n}|}{N}))^2}{N}, \tag{23}$$

where N is the total number of samples in the test dataset and $n = 1, .., N$.

The results show that all four models have a small variance in their absolute errors, which indicates that the models are stable. Table 5 shows that the PI-NN model is only slightly poorer than the fully-NN model. The MAE and MSE of the fully-NN model account for 97.54% and

Table 5 Results of three models

Metrics	PI-NN	fully-NN	XGB	MIO-BF
MAE	0.0285	0.0278	0.0317	0.0353
MSE	0.0034	0.0029	0.0033	0.0035
Var	0.0026	0.0021	0.0023	0.0023

Table 6 Parameter settings of XGB model

Parameter	Searching space	Selected setting	Parameter	Searching space	Selected setting
max_depth	[4,6,8,10]	6	*colsample_bytree*	\	1
n_estimators	[50,100,150]	100	*colsample_bynode*	\	1
learning_rate	[0.05,0.1,0.2,0.3]	0.3	*min_child_weight*	[1,2,3]	1
sub_sample	[0.6,0.8,1]	1			

85.29% of the MAE and MSE of the PI-NN model, respectively. That is, the PI-NN model can replace the fully-NN model at a level of more than 97% as measured by the MSE, which means that the domain knowledge introduced into the neural network model increases model interpretability while preserving accuracy. Moreover, the number of weights in the PI-NN model is fewer than the number of weights in the fully-NN model, which indicates that the training time of the PI-NN model is less than the training time of the fully-NN model. Specifically, there are 282 weights between connected neurons in the fully-NN model, which is more than double the number of weights in the PI-NN model (135). The PI-NN model saves almost 10% of training time compared with the fully-NN model in the experiment. The performance of the XGB model is slightly poorer than both the PI-NN and fully-NN models. The neural network model is inferred to be more suitable for predicting ship fuel consumption.

The MIO-BF model does not perform as well as the other three AI models. And this result is consistent with existing literature (Uyanık et al. 2020, Parkes et al. 2018, Le et al. 2020b). However, the MIO-BF model is highly explainable. According to the values of decision variables, the corresponding formula of the MIO-BF model is:

$$y \propto STW^{2.9} + \ln(d_{\text{avg}}) + S_{\text{head}}^3 + \sqrt{|S_{\text{cross}}|}. \tag{24}$$

Formula (24) indicates that the 2.9^{th} power of ship sailing speed is proportional to fuel consumption, which is in line with previous studies (Wang and Meng 2012, Meng et al. 2016, Le et al. 2020a). The logarithmic form of the ships' draft is proportional to fuel consumption. That is, the draft has a smooth effect on ship fuel consumption. Surprisingly, headwinds seem to have a greater effect on ship fuel consumption than crosswinds because the optimal expression of the crosswind

is a root transformation but the optimal expression of the headwind is a cubic transformation. Based on common sense, the effect of the crosswind may be greater than that of the headwind when a ship sails. Such a counterintuitive result might be caused by the wind angle on the actual sailing route of the ferry. Figure 6 shows the histogram of the frequency distribution of the wind angle in the dataset. In most cases, the wind angle breaks up more wind force horizontally than vertically. This study argues that no shipping company wants to operate a sailing route that is subject to perennial crosswinds, as these create dangerous sailing conditions. Thus, as the ferry between Tórshavn and Suðuroy is already in operation, its sailing route should be appropriate for sailing, and the counterintuitive result (that the crosswind has a smaller effect than the headwind) based on the data generated by the ferry may be obtained. In summary, the result is reasonable because the experiment is based on a dataset generated by a ship in operation and the effect of the crosswind may already be taken into account by managers in the decision stage.

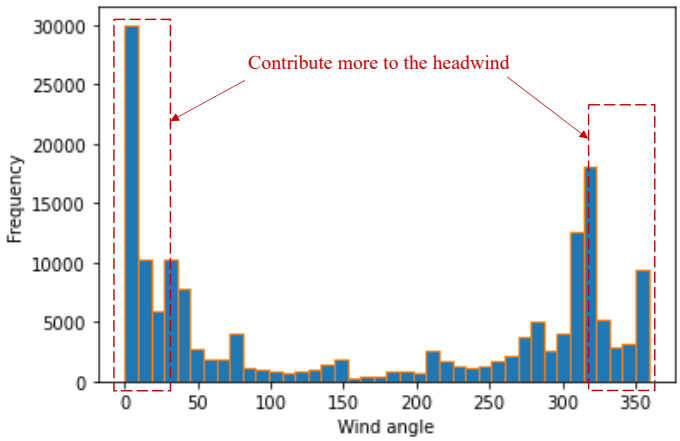

Figure 6 The distribution of wind angle

The findings answer the three questions put forward in Section 1.

R1: The proposed PI-NN model is on par with the fully-NN model to a degree of 97%. That is, the PI-NN model achieves the goal of improving model interpretability while maintaining model accuracy because it takes advantage of domain knowledge. By adding the constraints of physics, the PI-NN model becomes more acceptable to practitioners, thereby predicting ship fuel consumption more accurately.

R2: The findings indicate that the 2.9^{th} power of ship sailing speed, the logarithmic form of draft, the root transformation of the crosswinds, and the cubic transformation of the headwinds are the best formations for fitting a linear regression model to predict ship fuel consumption. The MIO-BF model is an explainable additive model and the relationship between selected forms of

variable expressions and ship fuel consumption is in line with practice. The results of the MIO-BF model indicate that sailing speed is the most important factor in ship fuel consumption. The results of the MIO-BF model also suggest that the headwind and crosswind are important variables but are not always influential because an established sailing route does not have frequent strong winds. Moreover, as wind and draft are objective, uncontrollable factors, ship captains should focus more on the effect of sailing speed on fuel consumption.

R3: The MIO-BF model is slightly poorer than the other AI models in terms of performance, which is consistent with the findings of the literature (Le et al. 2020b, Parkes et al. 2018). Moreover, the neural network model is more suitable for predicting ship fuel consumption than the state-of-the-art tree-based models. All of the four models own good stability. In summary, the PI-NN model improves the interpretability of black-box models and the MIO-BF model allows more variable expressions to be considered in a linear regression model. Managers can flexibly choose between the two models according to their needs for model accuracy and model interpretability.

5. Extension: SHAP values

SHAP is proposed by Lundberg et al. (2017). SHAP uses Shapley values from game theory to explain the prediction results and assigns a SHAP value to each feature value in each data sample (Lundberg et al. 2017, Wang et al. 2022b). SHAP provides a unified approach to interpreting model predictions and it is especially useful for explaining the prediction results of machine learning models. Different from the MIO-BF model and the PI-NN model, SHAP is developed based on a machine learning model to interpret the already predicted value of that machine learning model. Therefore, SHAP addresses the interpretability issue from the perspective of hindsight. To make the research more comprehensive, SHAP is further adopted to explore the feature importance for each predictor, i.e., d_{avg}, STW, S_{head}, and S_{cross} and quantifies the contribution of each feature value.

Referring to Wang et al. (2022b) and Yan et al. (2022), feature importance values (SHAP values) are calculated from two angles: global interpretability and local interpretability. Global interpretability means that the absolute SHAP values of each variable from the training data are averaged to measure the feature importance globally. Local interpretability shows how the contribution of an individual predictor varies across selected samples (Wang et al. 2022b). Readers are referred to Lundberg et al. (2017) for more details about SHAP. All the following figures are created by the SHAP Python module (Lundberg et al. 2017).

Figure 7 shows the contributions of each predictor from the global interpretability perspective. Note that "avgLEVEL" is d_{avg}. Variables are ranked in descending order. The top variable is STW, which indicates that the ship's sailing speed is the most important factor affecting fuel

consumption. And this result is in line with the consensus in the maritime field (Meng et al. 2016, Wang and Meng 2012). The second important feature variable is d_{avg}, followd by S_{head} and S_{cross}. The ordering of S_{head} and S_{cross} is consistent with what this study has addressed in Section 4.3.

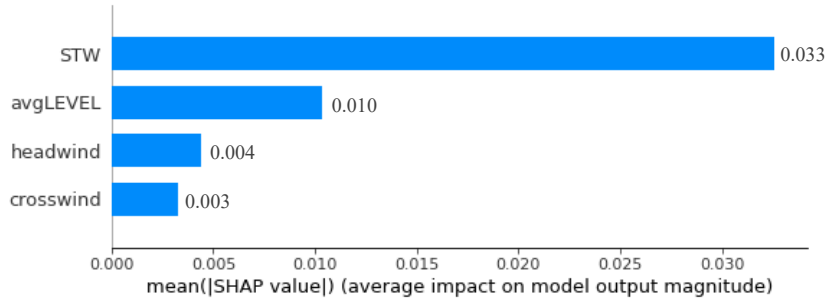

Figure 7 The SHAP variable importance on global interpretability

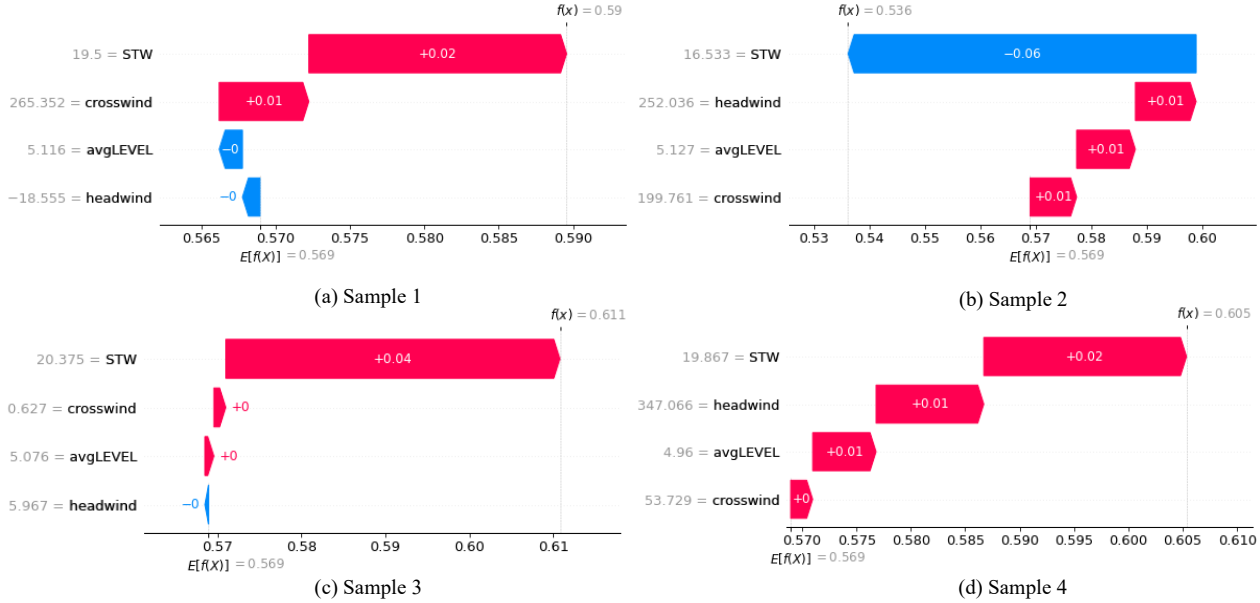

Figure 8 The SHAP value: local interpretability

Four samples are randomly selected in the dataset and analyze how the four feature variables contribute to the final predictions. The expectation value (i.e., the base value) in Figure 8 indicates that the predicted value of ship fuel consumption per second is 0.569kg/s when any values of the feature variables are not revealed. The base value is the mean value of all the ship fuel consumption records in the training dataset. Taking the values of the feature variables into account, the final prediction is the sum of the SHAP value of each feature variable and the base value. For example, the predicted value of ship fuel consumption in Figure 8(b) is 0.536kg/s, which is the sum of

0.569kg/s (the base value), -0.06kg/s (the contribution of the ship sailing speed), 0.01kg/s (the contribution of headwind), 0.01kg/s (the contribution of the draft), and 0.01kg/s (the contribution of crosswind). Note that the third digit is different as the values are displayed to the last two decimal places. Figure 8 shows that sailing speed contributes the most to ship fuel consumption in all of the samples. In Figure 8(a), the crosswind value contributes positively to ship fuel consumption because the feature value of crosswind is 265.352m/s, which is quite high and increases fuel consumption. The value of the headwind variable (-18.555m/s) decreases fuel consumption as a negative headwind value indicates a tailwind. However, the feature value is low, and thus the effect is small. In Figure 8(b), the SHAP value of STW is negative, which means when the ship is sailing at 16.533 knots, the sailing speed is less than the average ship sailing speed, creating a negative SHAP value that is subtracted from the base value in the calculation of the predicted value. It is also found that the value of ship sailing speed in Figures 8(a), Figure 8(c), and Figure 8(d) all contribute positively to ship fuel consumption because the value of STW in these three samples is large. The SHAP values for the crosswind and headwind in Figure 8(b) are all positive because there are strong crosswinds and headwinds and the ship sails against the wind. In Figure 8(c), the SHAP values of all the variables except ship sailing speed are low. Figure 8(d) indicates that the ship sails against the wind and the effect of the headwind is large. The contribution of each feature variable is not independent and the SHAP value of one feature variable is influenced by other feature variables in the sample. For example, there is no big difference between the value of draft in Figures 8(a) and 8(c) but the direction of the effect is different.

These findings show that SHAP is helpful for understanding the effect of feature values on the output. The global interpretability of SHAP makes it possible to determine the average effect of all of the feature variables on the predicted values. The local interpretability of SHAP clearly quantifies the contribution of each feature value to the final predicted value and helps in understanding the internal mechanisms in black-box models. SHAP provides a posterior alternative for explaining machine learning models. That is, a machine learning model is trained in the first stage and then SHAP is used to interpret it in the second stage. This study comprehensively explores the issue of interpretability in ship fuel consumption prediction by improving interpretability from a model-building perspective and presenting SHAP for the posterior explanation.

6. Conclusion

Ship fuel consumption is an important issue in the shipping industry. In this study, two innovative approaches are developed for predicting ship fuel consumption that addresses the tradeoff between model interpretability and model accuracy. Although some black-box models are quite advanced and can deliver accurate predictions, they lack interpretability and hence are rarely applied in

practice. The proposed PI-NN model incorporates the constraints of physics into a neural network model; the results of the experiment using real-world data demonstrate that the effectiveness of the PI-NN model is on par with that of the fully-NN model to a degree of 97%. An additive white-box model, the MIO-BF model, is also developed to consider more forms of feature variable expressions based on domain knowledge. The MIO-BF model can give an explicit expression for predicting ship fuel consumption by solving MIO programming. Practitioners can choose between the two approaches depending on their requirements: the PI-NN model is more suitable in scenarios requiring a high level of accuracy, whereas the MIO-BF model is more suitable in scenarios requiring a low level of accuracy but a high level of interpretability. SHAP, a popular interpretability method, is adopted to provide explanations for the results of the machine learning model.

This study helps to promote the application of data-driven models in maritime practice as models are developed based on domain knowledge, thereby making them more acceptable to practitioners. This study argues that using data-driven models for predicting ship fuel consumption will decrease fuel consumption, which will help to reduce operating costs, protect the environment, and achieve green shipping. AI has immense potential in the shipping industry. This study provides methods for coupling AI models with domain knowledge; this study also provides alternatives for interpreting black-box AI models. This research contributes to the application of AI in the shipping industry as the findings show that domain knowledge can complement AI models. With the help of domain knowledge, AI can lead to digital transformation, energy efficiency, and predictive analytics in the maritime industry.

References

Adland, R., Cariou, P., Wolff, F. C. (2020). Optimal ship speed and the cubic law revisited: empirical evidence from an oil tanker fleet. Transportation Research Part E: Logistics and Transportation Review, 140, 101972.

Bertsimas, D., King, A. (2016). OR forum—an algorithmic approach to linear regression. Operations Research, 64(1), 2–16.

Bertsimas, D., King, A., Mazumder, R. (2016). Best subset selection via a modern optimization lens. The Annals of Statistics, 44(2), 813–852.

Beşikçi, E. B., Arslan, O., Turan, O., Ölçer, A. I. (2016). An artificial neural network based decision support system for energy efficient ship operations. Computers & Operations Research, 66, 393–401.

Carvalho, D. V., Pereira, E. M., Cardoso, J. S. (2019). Machine learning interpretability: a survey on methods and metrics. Electronics, 8(8), 832.

Chang, Y. C., Tseng, R. S., Chen, G. Y., Chu, P. C., Shen, Y. T. (2013). Ship routing utilizing strong ocean currents. The Journal of Navigation, 66(6), 825–835.

Chen, T. and Guestrin, C. (2016). XGBoost: a scalable tree boosting system. In Proceedings of the 22nd ACM, SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

Derksen, S., Keselman, H. J. (1992). Backward, forward and stepwise automated subset selection algorithms: frequency of obtaining authentic and noise variables. British Journal of Mathematical and Statistical Psychology, 45(2), 265–282.

Doshi-Velez, F., Kim, B. (2017). Towards a rigorous science of interpretable machine learning. arXiv preprint arXiv:1702.08608.

Du, Y., Meng, Q., Wang, S., Kuang, H. (2019). Two-phase optimal solutions for ship speed and trim optimization over a voyage using voyage report data. Transportation Research Part B: Methodological, 122, 88–114.

Fagerholt, K., Gausel, N. T., Rakke, J. G., Psaraftis, H. N. (2015). Maritime routing and speed optimization with emission control areas. Transportation Research Part C: Emerging Technologies, 52, 57–73.

Fagerholt, K., Laporte, G., Norstad, I. (2010). Reducing fuel emissions by optimizing speed on shipping routes. Journal of the Operational Research Society, 61(3), 523–529.

Fan, A., Yang, J., Yang, L., Wu, D., Vladimir, N. (2022). A review of ship fuel consumption models. Ocean Engineering, 264, 112405.

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., Yang, G. Z. (2019). XAI—Explainable artificial intelligence. Science Robotics, 4(37), 7120.

Hastie, T., Tibshirani, R., Tibshirani, R. (2020). Best subset, forward stepwise or lasso? Analysis and recommendations based on extensive comparisons. Statistical Science, 35(4), 579–592.

Hazimeh, H., Mazumder, R. (2020). Fast best subset selection: coordinate descent and local combinatorial optimization algorithms. Operations Research, 68(5), 1517–1537.

He, J., Hao, Y., Wang, X. (2021). An interpretable aid decision-making model for flag state control ship detention based on SMOTE and XGBoost. Journal of Marine Science and Engineering, 9(2), 156.

IMO. (2020). Fourth greenhouse gas study 2020. https://www.imo.org/en/OurWork/Environment/Pages/Fourth-IMO-Greenhouse-Gas-Study-2020.aspx. Accessed on 22 January 2023.

IMO. (2022). MEPC.328(76). https://wwwcdn.imo.org/localresources/en/OurWork/Environment/Documents/Air%20pollution/Certified%20copy%20of%20MEPC.328(76).pdf. Accessed on 9 January 2023.

James, G., Witten, D., Hastie, T., Tibshirani, R. (2013). An introduction to statistical learning, Volume 112. New York: Springer.

Kim, G., Lim, S. (2022). Development of an interpretable maritime accident prediction system using machine learning techniques. IEEE Access, 10, 41313–41329.

Kim, D., Antariksa, G., Handayani, M. P., Lee, S., Lee, J. (2021). Explainable anomaly detection framework for maritime main engine sensor data. Sensors, 21(15), 5200.

Le, L. T., Lee, G., Kim, H., Woo, S. H. (2020a). Voyage-based statistical fuel consumption models of ocean-going container ships in Korea. Maritime Policy & Management, 47(3), 304–331.

Le, L. T., Lee, G., Park, K. S., Kim, H. (2020b). Neural network-based fuel consumption estimation for container ships in Korea. Maritime Policy & Management, 47(5), 615–632.

Loyola-Gonzalez, O. (2019). Black-box vs. white-box: understanding their advantages and weaknesses from a practical point of view. IEEE Access, 7, 154096–154113.

Lundberg, S. M., Lee, S. I. (2017). A unified approach to interpreting model predictions. Advances in Neural Information Processing Systems, 30, 1–10.

Ma, Y., Zhao, Y., Yu, J., Zhou, J., Kuang, H. (2023). An Interpretable Gray Box Model for Ship Fuel Consumption Prediction Based on the SHAP Framework. Journal of Marine Science and Engineering, 11(5), 1059.

McClelland, J. L., Rumelhart, D. E., Group, P. R. (1986). Parallel distributed processing. Explorations in the Microstructure of Cognition, 2, 216–271.

Meng, Q., Du, Y., Wang, Y. (2016). Shipping log data based container ship fuel efficiency modeling. Transportation Research Part B: Methodological, 83, 207–229.

Natarajan, B. K. (1995). Sparse approximate solutions to linear systems. SIAM Journal on Computing, 24(2), 227–234.

Parkes, A. I., Sobey, A. J., Hudson, D. A. (2018). Physics-based shaft power prediction for large merchant ships using neural networks. Ocean Engineering, 166, 92–104.

Petersen, J. P. (2012). Mining of ship operation data for energy conservation. Master thesis at. Technical University of Denmark.

Petersen, J. P., Jacobsen, D. J., Winther, O. (2012a). Statistical modelling for ship propulsion efficiency. Journal of Marine Science and Technology, 17, 30–39.

Petersen, J. P., Winther, O., Jacobsen, D. J. (2012b). A machine-learning approach to predict main energy consumption under realistic operational conditions. Ship Technology Research, 59(1), 64–72.

Raissi, M., Yazdani, A., Karniadakis, G. E. (2020). Hidden fluid mechanics: learning velocity and pressure fields from flow visualizations. Science, 367(6481), 1026–1030.

Rakke, J. G., Christiansen, M., Fagerholt, K., Laporte, G. (2012). The traveling salesman problem with draft limits. Computers & Operations Research, 39(9), 2161–2167.

Ribeiro, M. T., Singh, S., Guestrin, C. (2016). "Why should I trust you?" Explaining the predictions of any classifier. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 1135–1144.

Shao, W., Zhou, P., Thong, S. K. (2012). Development of a novel forward dynamic programming method for weather routing. Journal of Marine Science and Technology, 17, 239–251.

692 Sundararajan, M., Najmi, A. (2020). The many Shapley values for model explanation. In International
693 Conference on Machine Learning, 9269–9278.

694 Tan, F., Fu, X., Zhang, Y., Bourgeois, A. G. (2008). A genetic algorithm-based method for feature subset
695 selection. Soft Computing, 12, 111–120.

696 UNCTAD. (2022). Review of maritime transport. https://unctad.org/system/files/official-
697 document/rmt2022_en.pdf. Accessed on 20 February 2023.

698 Uyanık, T., Karatuğ, Ç., Arslanoğlu, Y. (2020). Machine learning approach to ship fuel consumption: a case
699 of container vessel. Transportation Research Part D: Transport and Environment, 84, 102389.

700 Uyanık, T., Bakar, N. N. A., Kalenderli, Ö., Arslanoğlu, Y., Guerrero, J. M., Lashab, A. (2023). A data-
701 driven approach for generator load prediction in shipboard microgrid: the chemical tanker case study.
702 Energies, 16(13), 5092.

703 Veerappa, M., Anneken, M., Burkart, N., Huber, M. F. (2022). Validation of XAI explanations for multivari-
704 ate time series classification in the maritime domain. Journal of Computational Science, 58, 101539.

705 Wang, L., Gopal, R., Shankar, R., Pancras, J. (2022b). Forecasting venue popularity on location-based
706 services using interpretable machine learning. Production and Operations Management, 31(7), 2773–
707 2788.

708 Wang, S., Meng, Q. (2012). Sailing speed optimization for container ships in a liner shipping network.
709 Transportation Research Part E: Logistics and Transportation Review, 48(3), 701–714.

710 Wang, S., Qi, J., Laporte, G. (2022a). Governmental subsidy plan modeling and optimization for liquefied
711 natural gas as fuel for maritime transportation. Transportation Research Part B: Methodological, 155,
712 304–321.

713 Wang, Y., Meng, Q., Kuang, H. (2018). Jointly optimizing ship sailing speed and bunker purchase in liner
714 shipping with distribution-free stochastic bunker prices. Transportation Research Part C: Emerging
715 Technologies, 89, 35–52.

716 Yan, R., Wang, S., Cao, J., Sun, D. (2021b). Shipping domain knowledge informed prediction and optimiza-
717 tion in port state control. Transportation Research Part B: Methodological, 149, 52–78.

718 Yan, R., Wang, S., Du, Y. (2020). Development of a two-stage ship fuel consumption prediction and reduction
719 model for a dry bulk ship. Transportation Research Part E: Logistics and Transportation Review, 138,
720 101930.

721 Yan, R., Wang, S., Psaraftis, H. N. (2021a). Data analytics for fuel consumption management in maritime
722 transportation: Status and perspectives. Transportation Research Part E: Logistics and Transportation
723 Review, 155, 102489.

724 Yan, R., Wu, S., Jin, Y., Cao, J., Wang, S. (2022). Efficient and explainable ship selection planning in port
725 state control. Transportation Research Part C: Emerging Technologies, 145, 103924.

726 Yang, L., Chen, G., Rytter, N. G. M., Zhao, J., Yang, D. (2019). A genetic algorithm-based grey-box model
727 for ship fuel consumption prediction towards sustainable shipping. Annals of Operations Research,
728 1–27.

729 Zhang, C., Zou, X., Lin, C. (2022). Fusing XGBoost and SHAP models for maritime accident prediction and
730 causality interpretability analysis. Journal of Marine Science and Engineering, 10(8), 1154.