

Ship Fuel and Carbon Emission Estimation Utilizing Artificial Neural Network and Data Fusion Techniques

Shaohan Wang^{1,2,3} , Xinbo Wang^{1,2*}, Yi Han¹, Xiangyu Wang¹, He Jiang¹, Zhexi Zhang¹

¹COSCO Shipping Technology Co., Ltd., Shanghai, China

²School of Naval Architecture, Ocean and Civil Engineering, Shanghai Jiao Tong University, Shanghai, China

³Shanghai Ship and Shipping Research Institute, Shanghai, China

Email: *wanghetongtt@hotmail.com

How to cite this paper: Wang, S.H., Wang, X.B., Han, Y., Wang, X.Y., Jiang, H. and Zhang, Z.X. (2023) Ship Fuel and Carbon Emission Estimation Utilizing Artificial Neural Network and Data Fusion Techniques. *Journal of Software Engineering and Applications*, 16, 51-72.

<https://doi.org/10.4236/jsea.2023.163004>

Received: February 18, 2023

Accepted: March 20, 2023

Published: March 23, 2023

Copyright © 2023 by author(s) and Scientific Research Publishing Inc. This work is licensed under the Creative Commons Attribution International License (CC BY 4.0).

<http://creativecommons.org/licenses/by/4.0/>



Open Access

Abstract

Ship energy consumption and emission prediction are the main concern of the shipping industry for ship energy efficiency management and pollution gas emission control. And they are attracting more global attention and research interests because of the increase in global shipping trade volume. As the core of maritime transportation, a large volume of data is collected around ships such as voyage data. Due to the rapid development of computational power and the widely equipped AIS device on ships, the use of maritime big data for improving and monitoring ship's energy efficiency is becoming possible. In this paper, a fuel consumption and carbon emission model using the artificial neural network (ANN) framework is proposed by using AIS, ship machinery, and weather data. The proposed work is a complete framework including data collection, data cleaning, data clustering and model-building methodology. To obtain the suitable parameters of the model, the number of neurons, data inputs and activate functions were tested on both AIS-based data and MRV-based data for comparison. The results show that the proposed method can provide a solid prediction of ship's fuel consumption and carbon emissions under varying weather conditions.

Keywords

Artificial Neural Network, Ship Fuel Consumption, Regression Analysis, AIS, Container Ship, IMO, Carbon Neutrality

1. Introduction

With the rapid growth of the shipping trade volume, as the main component of

international trade and economic development, carbon dioxide emission from ships is drawing more attention than before [1] [2]. In June 2021, the 76th session of the Marine Environment Protection Committee of the International Maritime Organization (IMO) formally revised the International Convention for the Prevention of Pollution from Ships (MARPOL) Attach VI. These alterations specified operational approaches to lower the carbon emission of international shipping which requires all ships to meet not only both the Energy Efficiency Existing Ship Index (EEXI) but also the annual operational Carbon Index Indicator (CII). The latter will rate all ships above 5000 tons of deadweight from A to E accordingly, and ships rated below the level of C may face a formal modification report for the next few years or even a suspension of operation [3].

Recently, for energy-saving purposes and decreasing the number of emissions from ships, many researchers focus on the ship speed and route optimizations. The speed optimization methods so far mainly are taking weather conditions as an essential factor for influencing ship's speed due to different wind speeds, wave heights and other related sea environments [4]-[10]. The route planning method, on the other hand, tries to find an economical route for minimal fuel consumption based on the sea state information in real time and related maritime affairs [11]-[16].

Thus, both methods above are under the foundation of an accurate estimation of fuel consumption and carbon emissions. Based on the ship's energy efficiency optimization method, current fuel consumption and carbon emission estimations can be categorized into two types: the white-box-model (WBM), the black-box-model (BBM) and the grey-box-model (GBM) optimization. The first one requires the details of ships in shape, system, and fuel types [17] [18] [19] [20] which involves a complicated researching framework and detailed simulated ship model along with the disadvantages of being costly and hard to be implemented into a wide range of no smart devices installed ships. WBMs relatively are precise for specific ships because it requires ship's mechanism analysis. As ships are strongly affected by the navigation environment, the use of WBMs will be a very complex process since it needs to find the governing equation for each physical law during sailing which can influence the fuel consumption. For example, some research is based on the ship's engine power and the law of resistance transfer for discovering the relationship between ship's power and energy consumption [21] [22]. Additionally, the WBMs lack the capability to adapt their parameters to the changing navigational environment, which necessitates numerous assumptions. These assumptions include the ship's resistance interaction, the cubic relationship between the ship's engine speed and fuel consumption, and the engine's constant working efficiency. Consequently, WBMs are frequently employed to analyze a ship's performance in its initial operational phases or enhance its design rather than monitor its actual performance during voyages.

In contrast to the WBMs, with the fast development of computing power and the Internet of Things (IoT), the use data analysis along with data mining tech-

niques such as machine learning (ML), deep learning (DL) and other statistical analysis are becoming possible [23]-[29]. The BBMs entirely rely on data analysis by processing multi-dimensional data and extracting hidden information from complex dataset. Then BBMs can output a reliable basis of ship's energy performance [30]. One commonly used machine learning technique for the prediction of fuel used is the regression models such as linear regression, ridge regression and lasso regression [28] [31], however, it is found that one single regression model may not be sufficient for representing the entire fuel consumptions system thus multi regressions are needed in order to evaluate the whole relationship. Besides, the regression methods are not sensitive to the non-linear problem which very often appeared in ship's energy analysis.

Subsequently, researchers turned to deep learning algorithms to capture the intricate connections between the navigation environment and a ship's fuel consumption. These models sacrifice the "explainability" feature of the model and instead focus on extracting complex data features at higher levels of abstraction. DL models typically have a layered structure, where the lower level's output computes the higher level's abstraction. This allows DL models to represent features that are difficult to explain through principles, such as the impact of wind speed, wave height, and stream speed on a ship's lost speed and fuel consumption. In these methods, data related to a ship's fuel consumption is repeatedly trained, adjusting the weight of each parameter and simulating the relationship between fuel consumption and input data. Due to their ability to accurately estimate a ship's energy performance, many studies have been conducted. For example, Du *et al.* [32] employed an artificial neural network (ANN) with voyage report data's speed to predict fuel consumption and test future voyage report accuracy. Petersen J.P. *et al.* [23] compared the performance of ANN and Gaussian Process (GP) in estimating a ship's propulsion efficiency. Farag YBA *et al.* combined ANN with polynomial regression to estimate a ship's power and fuel consumption, enabling it to operate in real-time environments and adapt to changes in the ship's environment.

Although DL models offer several advantages, their accuracy and reliability are heavily reliant on the quality of their input data. Consequently, researchers are also investigating various data sources and pre-processing techniques, such as the ship's noon report [32], sailing logs, automatic identification systems (AIS), or the MRV report. Furthermore, it has been demonstrated that even when using data from the same time period, there can be a significant difference between AIS and MRV data [28], which may be attributed to the precision of the data [33].

Despite the research so far, there still exists some research gaps based on the literature reviews:

- 1) At present, the research on ship fuel consumption models often are limited to a single type of ships, there is no existing models so far that can cover most operating ship types (*i.e.*, the bulk ships, container ships, and oil tankers).

2) A complete framework including the data collection method, and data clearing process is needed. Current studies' processing methods are varying which cannot be used to cover data from different sources.

3) Existing studies, including the GBMs, only tries to insert physical-based equations inside the DL models which can result in a very time consuming process.

4) Most studies so far are tested on a limited number of ships thus the method's robustness on all ships is questionable.

To address the research gap in ship fuel consumption and emissions by considering the weather's impact on a ship's speed and energy savings, this study proposes using an ANN with two types of weather data as inputs to predict daily fuel consumption. The study's contributions are as follows:

1) An end-to-end framework that includes a complete data processing method, leveraging maritime technical knowledge for data selection to improve data quality, from data collection to model output.

2) The study's results demonstrate the model's ability to cover the primary types of ships.

3) The proposed method was tested on 665 ships with over 90080099 AIS data points.

The remainder of this paper is organized as follows: Section 2 introduces the data collection and processing framework, including explanations of the data sources and types. Section 3 provides detailed explanations of the DL models used, namely ANN, and compares it with the two most commonly used machine learning methods for ship energy estimation: ridge regression and polynomial regression. Section 4 presents a case study of the proposed method on 665 ships, followed by fuel and carbon estimation results in Section 5. Finally, the last section concludes the study and outlines possible future research directions.

2. Data Preprocessing

Big data analysis of ship fuel consumption consists of four parts, as shown in **Figure 1**, data cleaning, data fusion, data clustering, and predictive analysis using neural networks.

The ship-related data used, such as AIS data, MRV data, and meteorological data, have a human-filled component. Therefore, the data used must contain human errors and omissions. Data cleaning eliminates outliers and deviation values from the original data. In data cleaning, if the data values collected at adjacent times differ significantly, or if the data deviate significantly from the theoretical upper and lower limits of most variables. In this study, the theoretical range of some characteristic variables of ships is calculated by using the basic information of ships, and the results of the theoretical range are used to filter and screen the original data.

Since AIS data, MRV data, and meteorological data are three different datasets, matching and fusion of the three datasets are required before building the

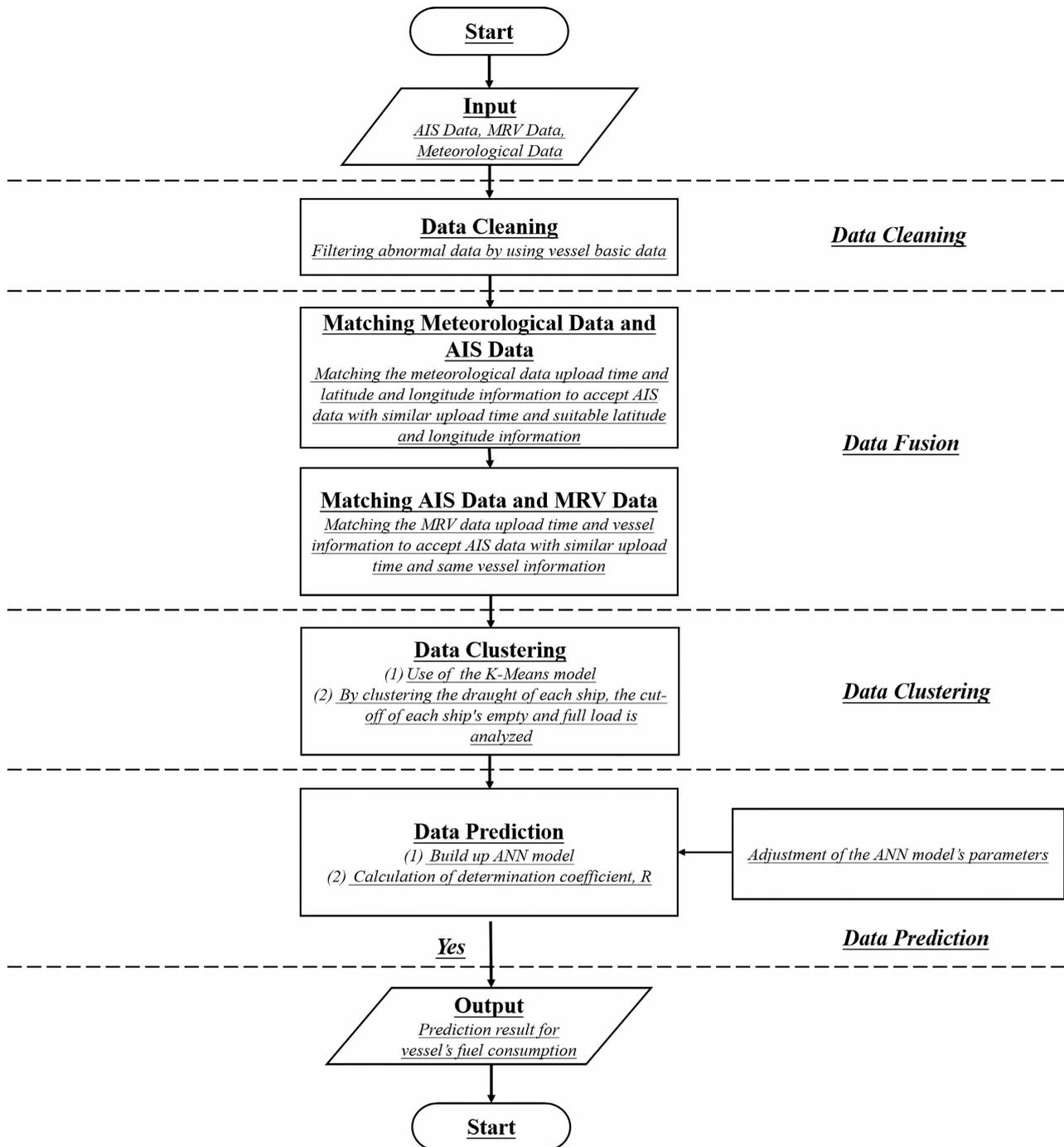


Figure 1. Methodological overview of the research.

model. In the data fusion part, the three data are matched using the time of uploading, where the geographic location information and time information are used to match the AIS data and meteorological data, and after forming the new data set, the time information and ship information are used to match with the MRV data, thus obtaining the data set used for modeling.

To model the fuel consumption of container ships, dry bulk carriers, and liquid bulk carriers, it is necessary to differentiate between empty and full load

drafts. This classification is important as container ships differ from dry bulk carriers and liquid bulk carriers. However, draught data from AIS information is subject to human error. To address this issue, the K-Means algorithm is used to cluster and classify the ship's draught into three loading states: full, empty, and between full and non-full, based on the ship's actual loading status. One-hot coding is then used to encode the three states, ensuring each state has a unique register bit, and only one is valid at any time. After verifying the fused and encoded data's compliance with the basic characteristics of the ship and its navigation, the data is processed for training and prediction using an artificial neural network.

The model's accuracy will be evaluated using commonly used model evaluation indices, with a particular focus on the total fuel consumption of the ship's voyage. Shipping companies typically measure the benefits of fuel consumption by the total fuel consumption of the ship voyage. Therefore, a comparison will be made between the predicted and real total fuel consumption of the ship, with an error value set to refine the neural network's parameters until the error between predicted and real total fuel consumption is less than the defined value.

Section 2 will provide detailed information on the dataset used, the pre-processing steps, and the prediction analysis.

2.1. AIS Data

COSCO Shipping Technology provided AIS data for a full year in 2021, containing 5 container ships, 3 dry bulk carriers, and 3 liquid bulk carriers, totaling 755,584 data. The AIS dataset used contains the following information: ship name, ship number, ship's geographical location, ship's draft, ship's heading, and ship speed.

2.2. MRV Data

China COSCO Shipping Corporation Limited provided MRV data for vessels from 2020-8-1 to the present. Measurement, reporting, and verification (MRV) record the total fuel consumption consumed by each vessel for each daily route and job. MRV data can be matched and integrated with AIS data to obtain dynamic data and fuel consumption data of each ship more easily.

2.3. Meteorological Data

Meteorological data are provided by Shanghai Meteorological Bureau. The meteorological data contains the size and direction of wind, waves, and currents. The meteorological data provided is based on the latitude and longitude grid. Therefore, the meteorological data reported in the latitude and longitude grid is used to represent all the meteorological conditions within the latitude and longitude grid.

2.4. Data Cleaning

Since there is a human-filled part of the AIS data, misreporting and omission are

inevitable. On the other hand, AIS data can be lost due to signal reception problems, thus it causes a long interval between uploads of two adjacent AIS data, which brings errors.

In order to reduce the impact of erroneous data on the prediction model, data cleaning will be performed based on the basic information of the ship object under study. Firstly, the upper and lower bounds of the theoretical sailing speed of the ship will be calculated based on the upper and lower bounds of the ship's pitch and theoretical speed as shown in Equation (1), and if the ship's sailing speed in the AIS data is not within the range of the theoretical sailing speed, the AIS data will be rejected.

$$speed = screw_pitch * engine_speed * 60 / 18520000 \quad (1)$$

where *speed* indicates the ship's sailing speed in *kn*; *screw_pitch* indicates the ship's pitch in mm; *engine_speed* indicates the ship's main engine speed in rdp/s.

2.5. Data Fusion

To enable predictive modeling, data fusion and matching based on potential associations in the three-part dataset are necessary, as three separate datasets are used. The meteorological data represents the meteorological situation within each degree of latitude and longitude grid and is matched with the time and geographic location information of the AIS data. When a ship sails into a specific latitude and longitude grid at a specific time, the meteorological data within that grid at that time is matched with the AIS data reported by the ship to obtain the AIS data with meteorological factors. The time reported by MRV and the time of AIS data upload is used to perform matching to obtain AIS data with meteorological and fuel consumption characteristics. **Figure 2** illustrates the data fusion process.

2.6. Data Clustering

K-Means is a common clustering method based on Euclidean distance. According to the characteristics of the ship's loaded cargo, it has been known in advance to classify the ship's loading into three cases: full load, empty load and between full load and empty load according to the ship's draft. And the most difficult part of K-Means is to determine the number of clusters for clustering, *i.e.* the value of *K*. However, since it has been determined to be divided into three categories, K-Means can accurately classify the ship draught into three categories according to the demand.

The basic idea of K-Means is that for a given sample set, the sample set is divided into *K* clusters according to the distance size between the samples. Let the sample points within the clusters be as closely connected as possible, *i.e.*, the clusters have high similarity; let the distance between the sample points between the clusters be as large as possible, *i.e.*, the similarity between the clusters is low.

The K-Means algorithm is based on the minimum-sum-squared error discriminant (MSE). Its cost function is shown below.

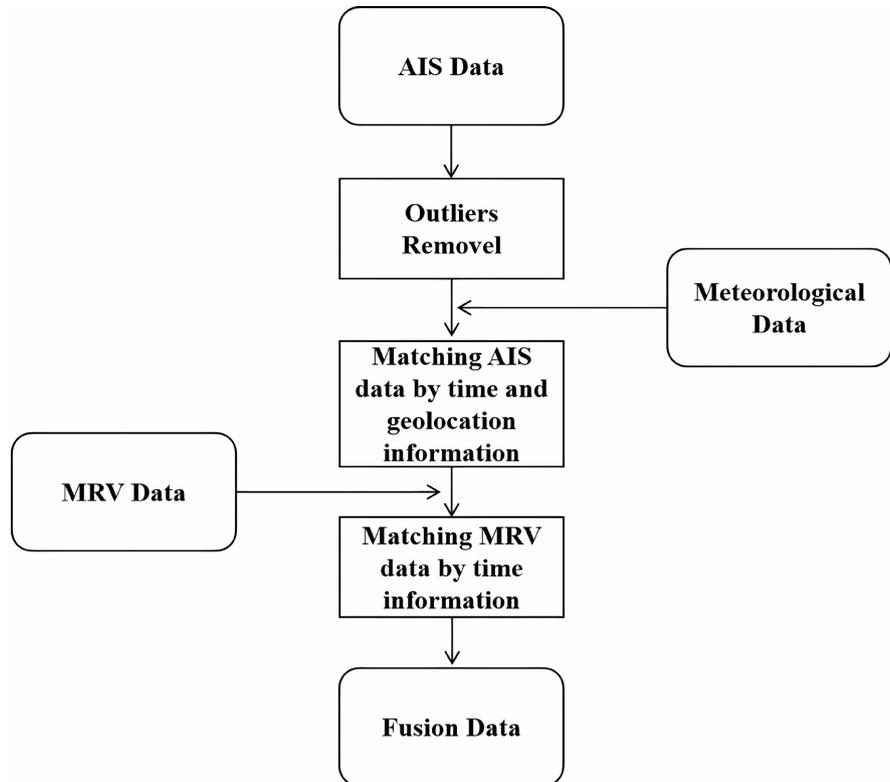


Figure 2. Fusion data processing.

$$J(c, \mu) = \sum_{i=1}^k \|x^{(i)} - \mu_{c^{(i)}}\|^2 \quad (2)$$

where $\mu_{c^{(i)}}$ denotes the closest center of cluster with the point of $x^{(i)}$, c is the center of the cluster. K -Means is the clustering of samples into K clusters, where the value of K is set artificially in advance, in this study $K = 3$. The specific algorithm is described as follows.

- 1) Randomly select K clustering prime points;
- 2) Repeat the following process until convergence.
 - For each sample, calculate the clusters to which it should belong.

$$c^{(i)} := \arg \min_j \|x^{(i)} - u_j\|^2 \quad (3)$$

- For each cluster, recalculate the center of mass of the class.

$$\mu_j := \frac{\sum_{i=1}^m I\{c^{(i)} = j\} x^{(i)}}{\sum_{i=1}^m I\{c^{(i)} = j\}} \quad (4)$$

After clustering the ship sketches using K -Means, three loading states were obtained for full, empty, and between full and empty. Since these three states are discrete and it is not easy to calculate the Euclidean distance between them. Therefore, using one-hot encoder, these three states are encoded to extend the discrete features to Euclidean space, and a certain value of the discrete features corresponds to a point in Euclidean space, so that each state has its own inde-

pendent hosting bit, so that only one state in each data entry is valid when it is input to the prediction model later. The predictive model used, the ANN, will be accepted in detail in Section 3.

2.7. Carbon Emissions Estimation

Since there is no significant carbon emission difference between heavy fuel oil (HFO) and light fuel oil (LFO), in the shipping industry, the main engine type is generally low-speed diesel engine, which corresponds to a carbon emission factor of 3.114 or 3.151, and in this study 3.114 is used as the carbon emission factor. The calculation formula is shown below [32].

$$carbon_emission = 3.114 * fuel \quad (5)$$

where *fuel* denotes the fuel consumption in tonnage.

3. DL Models for Ship Fuel Prediction

In this study, three types of prediction models are tested, namely ANN, ridge regression and polynomial regression. Each model is described in detail in Section 3.

3.1. Artificial Neural Network

Artificial neural network (ANN) is often used to solve classification and regression prediction problems. This allows us to build a nonlinear equation with input and output relations and visualize it by means of a network, which is known as an ANN. In general, ANNs can be configured to fit arbitrary nonlinear functions with a reasonable network structure, so they are also used to deal with nonlinear systems or black-box models with complex internal representations. Artificial neurons are the basic units in an artificial neural network; they receive one or more inputs and integrate them in a weighted form and produce an output. When each neuron receives the input of the variable, it will assign weights of different degrees according to the degree of the input variable, and then use the specific activation magic to carry out a weighted summation of the input variable with the assigned weights. There are different types of activation functions, such as nonlinear activation function, piecewise linear activation function, etc. Each activation function applies to different data sets and data types.

An artificial neural network consists of a group of simulated neurons. Each neuron is a node connected to other nodes by links corresponding to biological axon-synapse-dendrite connections. Each link has a weight that determines how strongly one node influences another. **Figure 3** illustrates a Fully Connected Neural Network (FNN).

FNN has the following characteristics.

1) The neurons are laid out according to layers. As shown in **Figure 3**, the leftmost is called Input layer, the middle is called Hidden layer, and the rightmost is called Output layer;

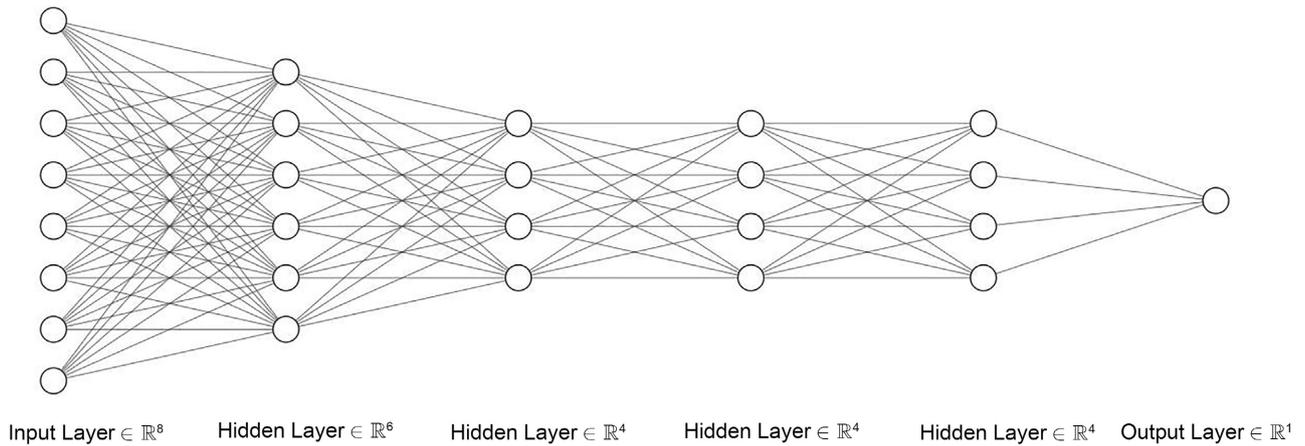


Figure 3. Neural network model for prediction model.

2) Neurons of the same layer are not connected;

3) Each neuron in layer N is connected to all neurons in layer $N - 1$ (this is what Full connected means), and the output of the neuron in layer $N - 1$ is the input of the neuron in layer N ;

4) Each neuron connection has a weight value. The input vector is denoted by $X = (x_1, \dots, x_n)$ and the output vector by $Y = (y_1, \dots, y_m)$;

5) In addition, the hidden layer can be multiple layers.

In this study, 8 input variables and 1 output variable were set. Eight input variables serve as the input layer, one output variable as the output layer, and four hidden layers are set in the middle.

Without loss of generality, assume that a training sample is $X = (x_1, \dots, x_8)$, and the corresponding output vector is $Y = y_1$, the output vector is a unique thermal encoding of categories. The input weight of the h th node in the hidden layer is v_{1h}, \dots, v_{8h} , and the corresponding offset is γ_h . The input weight of the output layer node is $\omega_1, \dots, \omega_q$, and the corresponding offset is θ . q is the number of nodes in the hidden layer. f is denoted as the activation function. Then the input of the output neuron is due to the following equation.

$$\beta = \sum_{h=1}^q \omega_h b_h \quad (6)$$

Formula (7) represents the output of the j th neuron.

$$y = f(\beta + \theta) \quad (7)$$

Formula (8) shows the input of neurons in the h th hiding layer:

$$\alpha_h = \sum_{i=1}^8 v_{ih} x_i \quad (8)$$

Formula (9) shows the output of neurons in the h th hiding layer:

$$b_h = f(\alpha_h + \gamma_h) \quad (9)$$

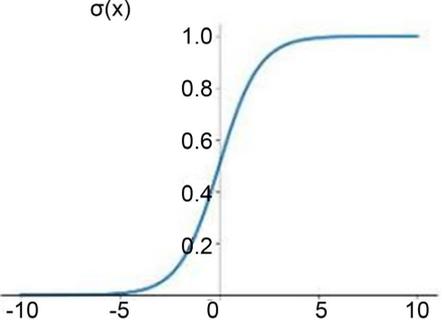
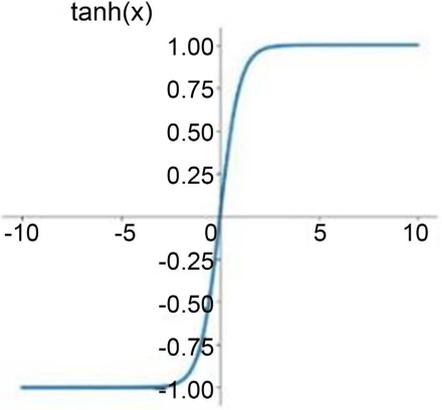
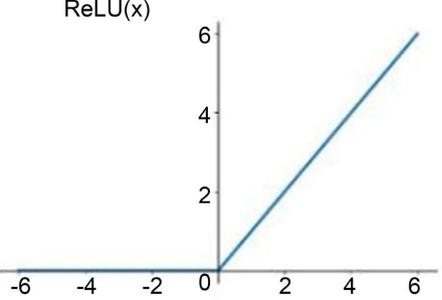
From this, the input and output of each neuron can be calculated.

Activation function is introduced to increase the nonlinearity of neural network model. If excitation function is not used, in this case, the input of nodes of

each layer is a linear function of the output of the upper layer. It is easy to verify that no matter how many layers there are in the neural network, the output is a linear combination of the input, which is equivalent to the effect without hidden layer. In this case, the most primitive Perceptron, the network's ability to approximate is very limited. The expression ability of neural network is enhanced by introducing nonlinear activation function. Common activation functions are shown in **Table 1**.

Since the sigmoid and Tanh functions have the same differential form, the sigmoid and Tanh functions can be repeatedly used. However, the sigmoid function ranges from 0 to 1, while the Tanh function ranges from -1 and 1 . As the number of hidden layers increases, the differential converges to zero rapidly, which makes the training and prediction of neural networks difficult. On the contrary, since the differential value of ReLU function is only 0 or 1, this function

Table 1. General activation function of artificial neural networks.

Activation Functions	Formula	Graph
Sigmoid	$g(x) = \frac{1}{1 + e^{-x}}$	
Tangent Sigmoid	$g(x) = \tanh(x)$	
ReLU (rectified linear unit)	$g(x) = \max(0, x)$	

solves this problem well, thus accelerating the training speed and improving the training effect of neural network [33]. Therefore, in this study, the sigmoid function, tangent sigmoid function and ReLU function will be used as activation functions respectively to build a fuel consumption prediction model. The results of fuel consumption prediction models established by different activation functions were analyzed and compared, and activation functions with higher accuracy were selected to establish the prediction model.

3.2. Other Models

The most used regression prediction model is polynomial regression (PR). Polynomial regression is a method of regression analysis that examines the polynomial relationship between a dependent variable and one or more independent variables. The regression prediction model can be expressed by the following equation.

$$Y = X\beta + \varepsilon \quad (10)$$

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_{n-1} \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & x_1 & \cdots & x_1^p \\ 1 & x_2 & \cdots & x_2^p \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & \cdots & x_n^p \end{bmatrix} \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{bmatrix} \quad (11)$$

where ε is the vector of unobserved random errors with mean zero for variable x_i , β is the vector of regression coefficients for variable x_i . β is the vector of regression coefficients for variable x_i , and p is the order of the polynomial. This vector of regression coefficients can be expressed by the least square method estimation, as shown in the following equation.

$$\beta = (X^T X)^{-1} X^T Y \quad (12)$$

Ridge regression was introduced to solve the problem of multicollinearity among the input variables. One way to eliminate multicollinearity is by regularizing the cost function by adding a penalty term. One of the regularization methods is called Tikhonov regularization, and the linear regression after this cost function regularization is called Ridge Regression.

The first term of the cost function of ridge regression is consistent with that of standard linear regression in that it is the sum of the squares of the Euclidean distances [34], except that it is followed by the square of the L_2 -parametric of a w -vector as the penalty term (L_2 -parametric means the sum of the squares of each element of the w -vector and then squared), where λ denotes the coefficient of the penalty term, which artificially controls the size of the penalty term. Since the regular term is L_2 -parametric, this regularization is sometimes called L_2 -regularization. The loss function of the ridge regression is shown below [35].

$$Cost(w) = \sum_{i=1}^n (y_i - w^T x_i)^2 + \lambda \|w\|_2^2 \quad (13)$$

As with the standard linear regression, the size of w that minimizes the cost function of the ridge regression is also found, as shown below.

$$w = \arg \min_w \left(\sum_{i=1}^n (y_i - w^T x_i)^2 + \lambda \|w\|_2^2 \right) \quad (14)$$

The analytic solution of w is obtained directly by derivation of the surrogate function, where X is an $n * m$ matrix, Y is an n -dimensional column vector, λ belongs to the set of real numbers, and I is the unit matrix of $m * m$.

$$w = (X^T X + \lambda I)^{-1} X^T Y, \lambda \in R \quad (15)$$

$$X = \begin{bmatrix} x_1^T \\ x_2^T \\ \vdots \\ x_n^T \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & \cdots & x_{1m} \\ x_{21} & x_{22} & \cdots & x_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ x_{n1} & x_{n2} & \cdots & x_{nm} \end{bmatrix} \quad (16)$$

$$Y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} \quad (17)$$

Ridge regression complements OLS by trading the loss of unbiasedness for higher numerical stability, thus improving the accuracy of the fitted predictions.

4. Case Study

4.1. Data Collection

The analysis has AIS data provided by COSCO, MRV data and meteorological data provided by Shanghai Meteorological Bureau. The three datasets were fused to obtain the dataset used for analytical modeling. The three datasets were fused to obtain the dataset used for analytical modeling. This dataset was collected from 2020/10/1 to 2022/4/1, which contains 184 container ship; 335 dry bulk carriers; 146 liquid bulk carriers. The extracted data set is divided into input variables and output variants for prediction analysis. The input variables contain wind size, wave size, sailing speed, daily main engine speed, sailing distance, full load, empty load and between full load and empty load, a total of 8 characteristic variables. Among them, full load, no load and between full load and no load are obtained by ship draft (half load), using data clustering and one-hot coding mentioned in Section 2. Eight input variables can be categorized into four types, which represent ship load status, engine operation, ship speed and meteorological conditions. The daily fuel consumption of the ship is then used as an output variable, as shown in [Table 2](#).

4.2. Data Preprocessing

After data cleaning, data fusion, data clustering and coding of the original trisection dataset, point-by-point fusion data similar to AIS data is obtained. However, since there is only one engine Speed and MRV Daily Fuel per day, while there

Table 2. Data description for ship fuel consumption prediction.

Data	No.	Parameter	Remark
Input	1	Wind_val (m/s)	Meteorological Data
	2	Wave_val (m)	
	3	Engine_speed (rpd/s)	Engine Operation
	4	Speed (kn)	
	5	Distance (km)	
	6	No_load	
	7	Full_load	Ship State
	8	Half_load	
Output	1	MRV Daily Fuel (tons/day)	Fuel Consumption

are multiple other input variables, further processing of the fused data is required to improve the accuracy and usefulness of the prediction model. The weather data and navigation speed are weighted using the adjacent time of each fused data, and the calculation formula is shown below. The sailing distance is summed to represent the total mileage of the ship sailing on this day. This results in a daily data set that represents the ship's sailing status, weather and fuel consumption.

$$Wind_val = \sum_{i=1}^n \frac{time_diff_i}{\sum_{i=1}^n time_diff_i} * Wind_val_i \quad (18)$$

$$Wave_val = \sum_{i=1}^n \frac{time_diff_i}{\sum_{i=1}^n time_diff_i} * Wave_val_i \quad (19)$$

$$Speed = \sum_{i=1}^n \frac{time_diff_i}{\sum_{i=1}^n time_diff_i} * Speed_i \quad (20)$$

where n denotes the total number of mixed data contained in a day, $Wind_val_i$, $Wave_val_i$ and $Speed$ denote the wind and wave size and sailing speed of the i th mixed data respectively, $time_diff_i$ denotes the time difference between the i th mixed data and the $i-1$ th mixed data, and $\sum_{i=1}^n time_diff_i$ denotes the total sailing time of this day.

4.3. Prediction Model of Ship Fuel Consumption Using ANN

As shown in Section 4.1, the input and output variables of **Table 2** are used as the input and output variables of the ANN model. Again, the number of neurons for input is set to 8 and the number of neurons for output is set to 1. The prediction model of ship fuel consumption is obtained using the ANN. MSE and R-square were used to check the fit of the prediction model to the given data. The MSE values and R-Square values of different activation functions are compared to determine the effect of the activation function on the prediction model.

The artificial neural network is set up with 4 hidden layers and the prediction model using the artificial neural network is compared with the other two regression models mentioned previously.

The fusion data set processed by the data processing method mentioned in Section 2 is divided into training set, test set and verification set. The training set is used to help the model to conduct predictive learning training. The test set is used to test the trained model, and the output results are used to evaluate the performance of the constructed prediction model. Validation set was used to verify the accuracy of the constructed prediction model, and the results were displayed.

5. Result

The results of the prediction model using three different activation functions are shown in **Table 3**. r-Square indicates the correlation coefficient between the predicted and true values obtained by the prediction model, and when the value of R is closer to 1, it means, the prediction model is a good fit. As mentioned before, the activation functions used here are sigmoid function, tangent sigmoid function and ReLU function. **Table 3** shows the detailed network structure information. The first number represents the dimension of the input variable in the input layer, as shown in **Table 3**. The dimension of the input layer is 8; The last number represents the dimension of the output variable of the output layer, that is, the dimension of the output layer is 1; The number in the middle indicates that there are 4 hidden layers, and the number of neurons in each layer is 6, 4, 4, 4 respectively.

MSE and R-Square are used to evaluate the accuracy and performance of fuel consumption prediction models constructed with different activation functions. The dataset is divided into four types of datasets: training set, test set, validation set and the set of all data. All data from 2021/1/1 to 2022/1/1 were utilized as the training set for the prediction model, all data from 2020/10/1 to 2021/1/1 were utilized as the test set, and all data from 2022/1/1 to 2022/4/1 were utilized as the validation set.

Table 3 shows the prediction results using three different activation functions, *i.e.*, sigmoid, tangent sigmoid, and ReLU functions, respectively. It can be seen from the table that the MSE and R-Square of the prediction model using the sigmoid function as the activation function are less effective compared to the other two. When the tangent sigmoid function is used as the activation function, it can be seen from the MSE and R-Square that the prediction results are slightly better than the prediction model using the ReLU function as the activation function.

The performance of the prediction models constructed using artificial neural networks compared with other regression models is shown in **Table 4**. In comparing the performance of each method, the same training set, test set, and validation set were used under the same conditions.

Table 3. ANN regression model using three different activation function.

No.	MSE	Iter.	Conf.	R-Square			
				Train	Test	Valid	ALL
ReLU Function	4.3269	35	8-6-4-4-4-1	0.9949	0.8868	0.9422	0.9089
Tangent Sigmoid Function	4.1069	35	8-6-4-4-4-1	0.9977	0.9418	0.92855	0.9311
Sigmoid Function	43.2897	35	8-6-4-4-4-1	0.8248	0.7256	0.7781	0.7504

Table 4. Ship fuel consumption results of different prediction models.

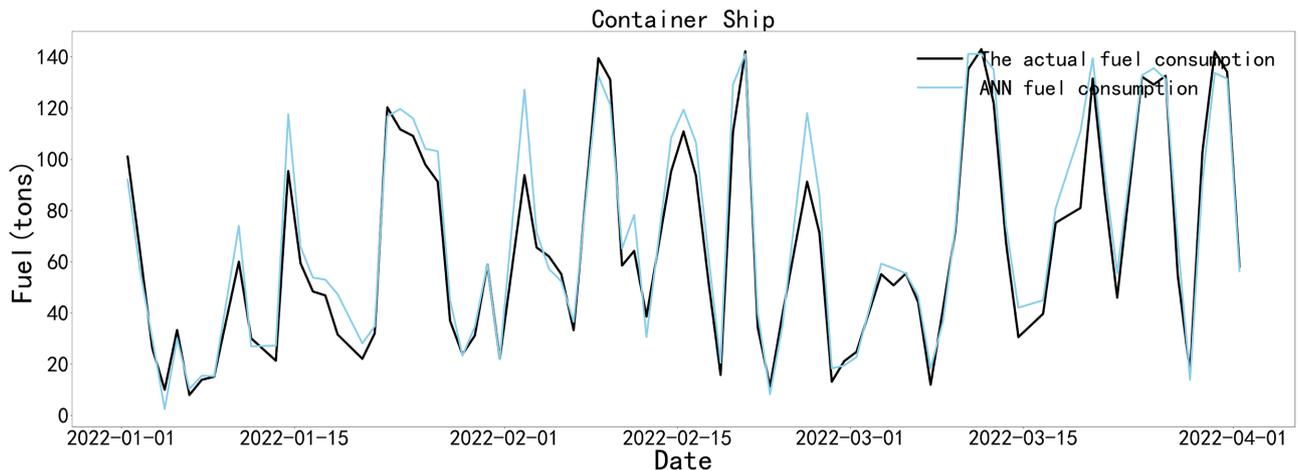
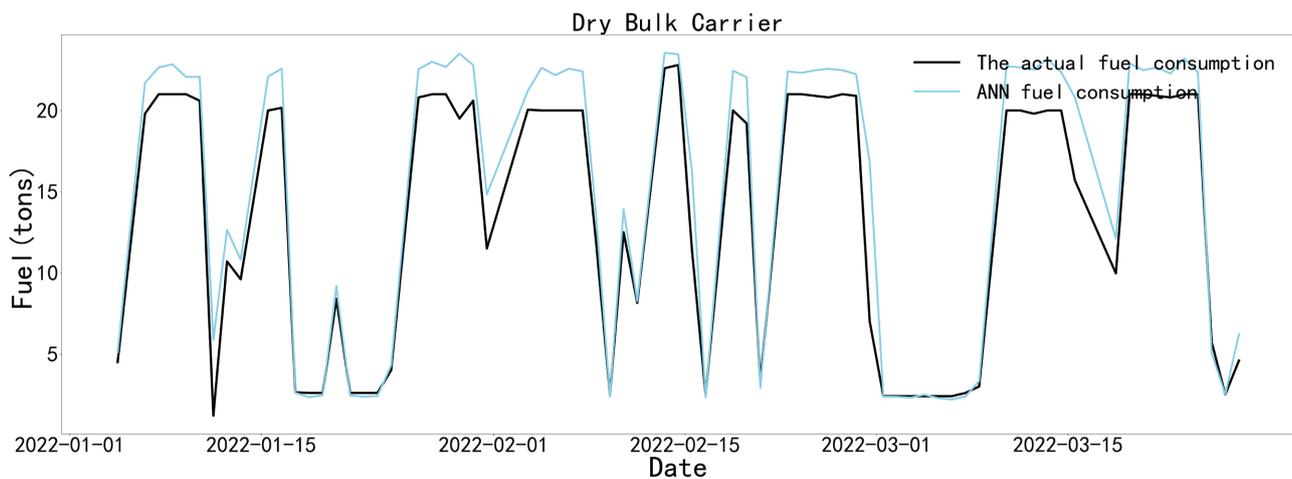
	Cumulative Error (tons)	MSE	R-Square	Time(s)
ANN	350	4.3269	0.9422	0.4
PR	877	132.1208	0.9187	4.8
Ridge Regression	1889.4761	1013.1	0.3816	0.2

All three prediction models were done in Python 3.9.7 and using normal PC specifications (i5 processor). A container ship was selected from the dataset to predict and analyze the fuel consumption of the ship from 2022/1/1 to 2022/4/1. From **Table 4**, it can be seen that ANN and Ridge regression are approximately the same in terms of computational speed, but it can be seen by the cumulative error, MSE and R-Square that the error result of Ridge regression is larger, and the effect of the model built using Ridge regression is more different from the true value. Although the performance of ANN and PR are similar in R-Square, the computational efficiency of the PR model is 12 times slower than that of ANN, and both the cumulative error and MSE are much larger compared to ANN, indicating that the prediction effect of the PR model is also deficient compared to that of the ANN model. Therefore, ANN is chosen as the preferred fuel consumption prediction model because it not only has a faster computing speed, but also ensures a higher accuracy.

Fuel consumption prediction models were constructed for each of the three different ship types using ANN. The results are shown in **Table 5**. The obtained results were obtained by selecting the data from 2022/1/1 to 2022/4/1 in the dataset as the validation set. The cumulative error response is the difference between the total predicted fuel consumption and the total actual fuel consumption from 2022/1/1 to 2022/4/1. As can be seen from **Table 5**, the cumulative error in fuel consumption for three months is about 350 tons for container ships, 68 tons for dry bulk carriers and 98 tons for liquid bulk carriers. One vessel from each of the three vessel types was selected for demonstration, and the results are shown in **Figures 4-6**. Combining **Table 5** and **Figure 4** and **Figure 5**, it can be seen that the fuel consumption prediction models established using ANN for container ships and dry bulk carriers work well, and the overall trend does not differ much from the real fuel consumption. Combining **Table 5** and **Figure 6**, it

Table 5. Fuel consumption prediction models for three different ship types.

	Avg. Cumulative Error (tons)	Avg. MSE	Avg. R-Square	Number of Ships
Container Ship	350	4.3269	0.9422	184
Dry Bulk Carrier	68	2.4857	0.9620	335
Liquid Bulk Carrier	98	11.8668	0.8105	146

**Figure 4.** Container ship fuel consumption.**Figure 5.** Dry bulk carrier fuel consumption.

can be seen that when the fuel consumption prediction model established by using ANN for the fuel consumption of liquid bulk carrier, the MSE of the model is large compared with the MSE of container model and dry bulk carrier, and the R-Square is small compared with the container model and dry bulk carrier model, but the cumulative error is only 98 tons, which is equivalent to the fuel consumption generated by the liquid bulk carrier policy sailing for 3 - 4 days, compared with the overall fuel consumption of three months, the overall error is relatively small. The reason for the error may be that there are

other potential fuel consumptions such as boiler oil in liquid bulk, and the model will be further optimized in the subsequent study to improve the model accuracy.

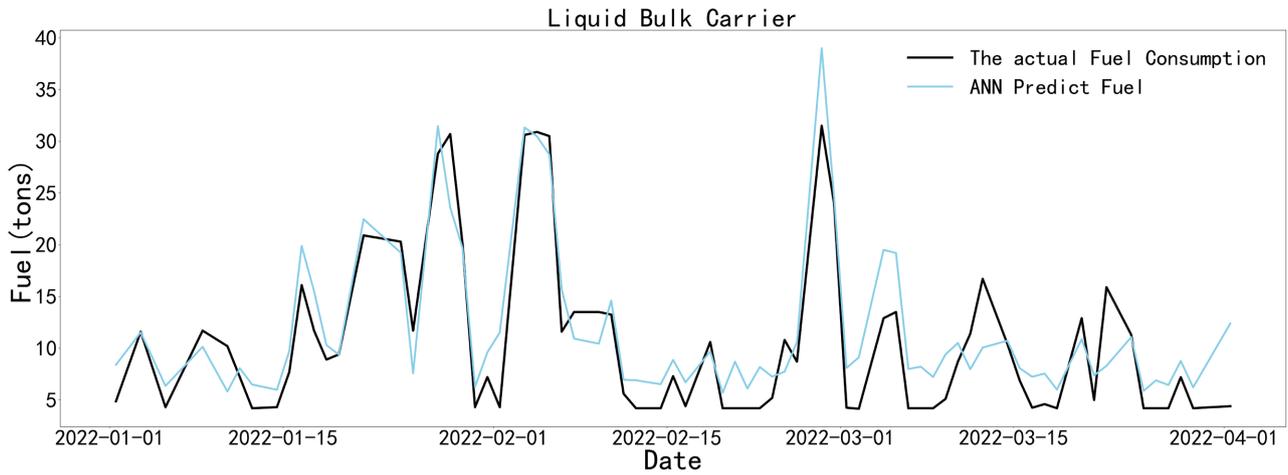


Figure 6. Liquid bulk carrier fuel consumption.

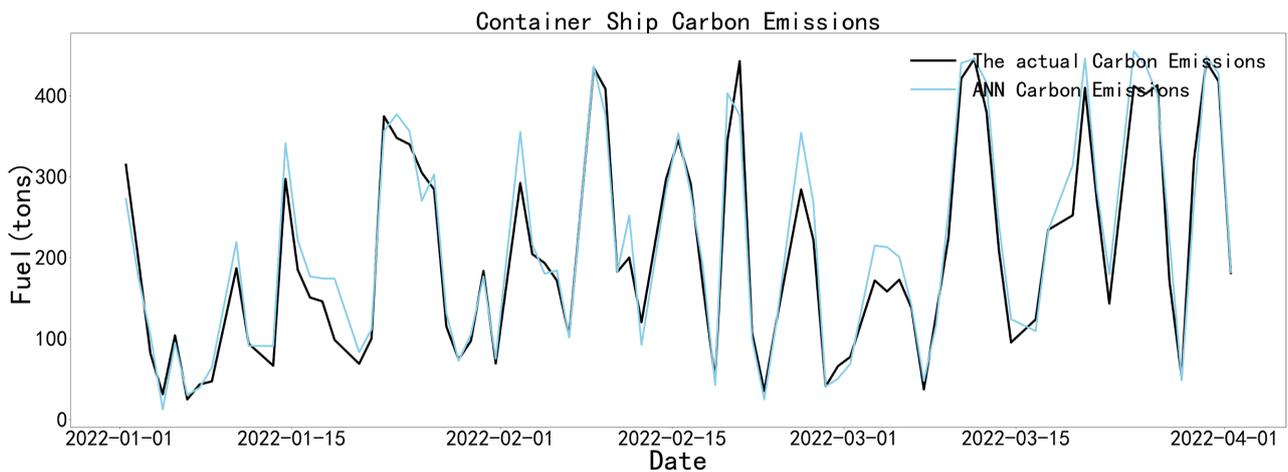


Figure 7. Container ship carbon emission estimations.

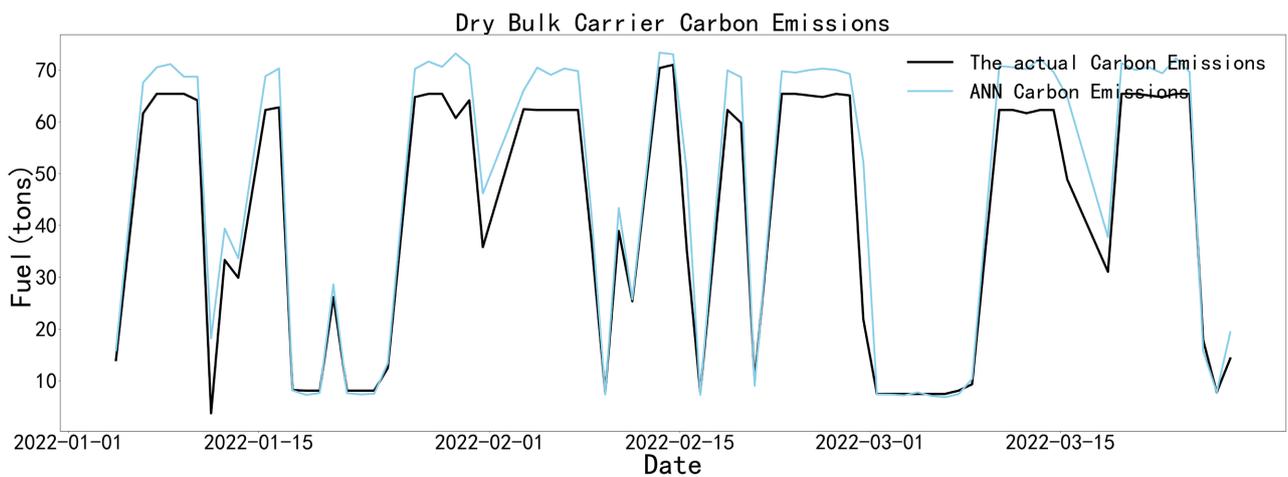


Figure 8. Dry bulk carrier carbon emission estimations.

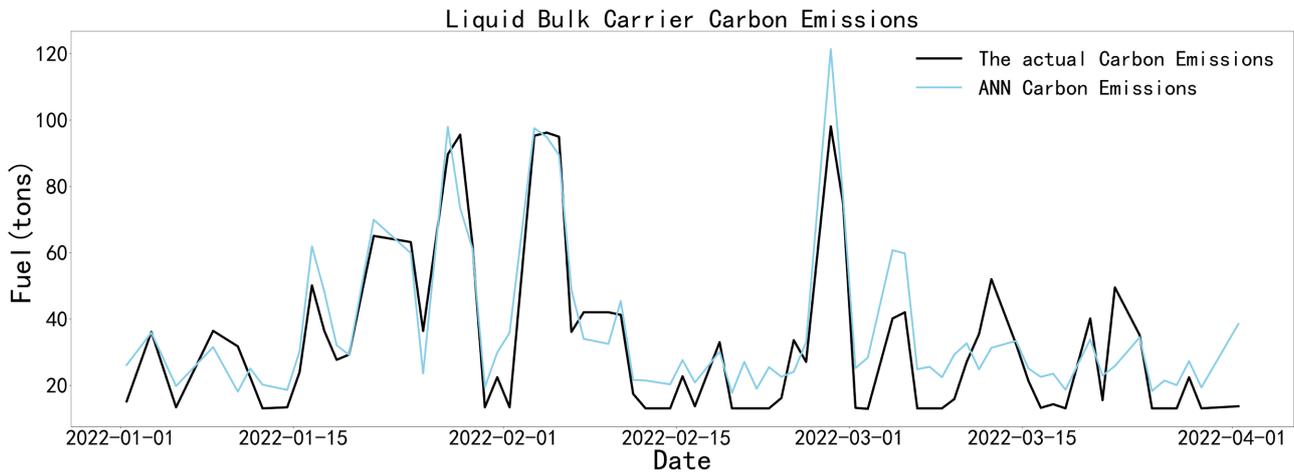


Figure 9. Liquid bulk carrier carbon emission estimations.

Using the fuel consumption prediction model established by ANN and Equation (5), the carbon emission of each ship can be estimated. It can be seen from **Figures 7-9** that the errors of carbon emissions for the selected container ships, dry bulk carriers and liquid bulk carriers are less than 10% compared with the errors of actual carbon emissions. Therefore, using this fuel consumption prediction model can effectively predict the total carbon emissions by carbon emission prediction.

6. Discussions

Several datasets are used in this work: the AIS data, MRV data and meteorological data to testify the influence of weather factors on the ship's energy performance. As the weather factors are bonded to ship's coordinate, the AIS is needed to calculate the weighted weather for each day. Besides, the construction of the ANN mode (such as the number of neurons and hidden layers) is based on the number of input and output of long with hyperparameter tuning during this work. Based on the results, the proposed ANN structures are best suited for estimation of fuels for bulk and container ships.

At the same time, it is also found that the quality of data has a strong influence on the performance of ANN predictions. As fuel consumptions from MRV are input manually by crews on ship, a large number of inaccurate data are hidden in the dataset. Besides, training dataset has lack information on serve weather conditions which can cause the ANN model's lack of accuracy when predicting ship in bad weather.

7. Conclusions and Future Work

In this study, an artificial neural network (ANN) was utilized to forecast a ship's fuel consumption and carbon emissions by incorporating various meteorological data and the main engine's daily state. The study's key findings are as follows:

- A comprehensive ship energy performance data pre-processing framework

was constructed, which included encompassed data source selection, data type selection, and parameter selection to forecast a ship's fuel consumption during navigation.

- A ship fuel consumption model was created that integrated both weather information during voyages and the ship's operational state (e.g., loading condition) into the model.
- Unlike grey box model, which incorporates physical laws into the model, this study incorporated technical knowledge of maritime transportation into the data cleaning process. Compared with the grey box and black box models, the proposed approach is more computationally efficient and explainable.
- The proposed model was evaluated on over 600 ships for three years, with 90 million AIS data points. The model's performance was valid for all three major types of ships.

Future research directions may include utilizing more precise weather data with a longitude and latitude granularity of less than 0.5, as well as studying additional ship behaviors such as berthing and drifting, in addition to the voyage period during sailing.

Conflicts of Interest

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

References

- [1] Asariotis, R. and Benamara, H. (2012) Maritime Transport and the Climate Change Challenge. Routledge, New York.
- [2] Wan, Z., Zhu, M., Chen, S. and Sperling, D. (2016) Pollution: Three Steps to a Green Shipping Industry. *Nature*, **530**, 275-277. <https://doi.org/10.1038/530275a>
- [3] Joung, T.-H., Kang, S.-G., Lee, J.-K. and Ahn, J. (2020) The IMO Initial Strategy for Reducing Greenhouse Gas (GHG) Emissions, and Its Follow-Up Actions towards 2050. *Journal of International Maritime Safety, Environmental Affairs, and Shipping*, **4**, 1-7. <https://doi.org/10.1080/25725084.2019.1707938>
- [4] Psaraftis, H.N. and Kontovas, C.A. (2013) Speed Models for Energy-Efficient Maritime Transportation: A Taxonomy and Survey. *Transportation Research Part C: Emerging Technologies*, **26**, 331-351. <https://doi.org/10.1016/j.trc.2012.09.012>
- [5] Adland, R., Cariou, P. and 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**, Article ID: 101972. <https://doi.org/10.1016/j.tre.2020.101972>
- [6] Cariou, P. and Cheaitou, A. (2012) The Effectiveness of a European Speed Limit Versus an International Bunker-Levy to Reduce CO₂ Emissions from Container Shipping. *Transportation Research Part D: Transport and Environment*, **17**, 116-123. <https://doi.org/10.1016/j.trd.2011.10.003>
- [7] Du, Y., Meng, Qi., Wang, S. and 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. <https://doi.org/10.1016/j.trb.2019.02.004>

- [8] Medina, J.R., Molines, J., González-Escrivá, J.A. and Aguilar, J. (2020) Bunker Consumption of Containerships Considering Sailing Speed and Wind Conditions. *Transportation Research Part D: Transport and Environment*, **87**, Article ID: 102494. <https://doi.org/10.1016/j.trd.2020.102494>
- [9] Fagerholt, K., Gausel, N.T., Rakke, J.G. and Psaraftis, H.N. (2015) Maritime Routing and Speed Optimization with Emission Control Areas. *Transportation Research Part C: Emerging Technologies*, **52**, 57-73. <https://doi.org/10.1016/j.trc.2014.12.010>
- [10] Wang, K., *et al.* (2018) Dynamic Optimization of Ship Energy Efficiency Considering Time-Varying Environmental Factors. *Transportation Research Part D: Transport and Environment*, **62**, 685-698. <https://doi.org/10.1016/j.trd.2018.04.005>
- [11] Prpić-Oršić, J., *et al.* (2015) Influence of Ship Routes on Fuel Consumption and CO₂ Emission. Maritime Technology and Engineering. Taylor & Francis Group, London.
- [12] Pallotta, G., Vespe, M. and Bryan, K. (2013) Vessel Pattern Knowledge Discovery from AIS Data: A Framework for Anomaly Detection and Route Prediction. *Entropy*, **15**, 2218-2245. <https://doi.org/10.3390/e15062218>
- [13] Prpić-Oršić, J., Vettor, R., Faltinsen, O.M. and Soares, C.G. (2016) The Influence of Route Choice and Operating Conditions on Fuel Consumption and CO₂ Emission of Ships. *Journal of Marine Science and Technology*, **21**, 434-457. <https://doi.org/10.1007/s00773-015-0367-5>
- [14] Lee, S.-M., Roh, M.-I., Kim, K.-S., Jung, H. and Park, J.J. (2018) Method for a Simultaneous Determination of the Path and the Speed for Ship Route Planning Problems. *Ocean Engineering*, **157**, 301-312. <https://doi.org/10.1016/j.oceaneng.2018.03.068>
- [15] Chu, P.C., Miller, S.E. and Hansen, J.A. (2015) Fuel-Saving Ship Route Using the Navy's Ensemble Meteorological and Oceanic Forecasts. *The Journal of Defense Modeling and Simulation*, **12**, 41-56. <https://doi.org/10.1177/1548512913516552>
- [16] Bui-Duy, L. and Vu-Thi-Minh, N. (2021) Utilization of a Deep Learning-Based Fuel Consumption Model in Choosing a Liner Shipping Route for Container Ships in Asia. *The Asian Journal of Shipping and Logistics*, **37**, 1-11. <https://doi.org/10.1016/j.ajsl.2020.04.003>
- [17] Fan, A., *et al.* (2022) Joint Optimisation for Improving Ship Energy Efficiency Considering Speed and Trim Control. *Transportation Research Part D: Transport and Environment*, **113**, Article ID: 103527. <https://doi.org/10.1016/j.trd.2022.103527>
- [18] Ouyang, Z.-L. and Zou, Z.-J. (2021) Nonparametric Modeling of Ship Maneuvering Motion Based on Gaussian Process Regression Optimized by Genetic Algorithm. *Ocean Engineering*, **238**, Article ID: 109699. <https://doi.org/10.1016/j.oceaneng.2021.109699>
- [19] Armstrong, V.N. (2013) Vessel Optimisation for Low Carbon Shipping. *Ocean Engineering*, **73**, 195-207. <https://doi.org/10.1016/j.oceaneng.2013.06.018>
- [20] Armstrong, V.N. and Banks, C. (2015) Integrated Approach to Vessel Energy Efficiency. *Ocean Engineering*, **110**, 39-48. <https://doi.org/10.1016/j.oceaneng.2015.10.024>
- [21] Yan, X., *et al.* (2018) Energy-Efficient Shipping: An Application of Big Data Analysis for Optimizing Engine Speed of Inland Ships Considering Multiple Environmental Factors. *Ocean Engineering*, **169**, 457-468. <https://doi.org/10.1016/j.oceaneng.2018.08.050>
- [22] Pan, P., *et al.* (2021) Research Progress on Ship Power Systems Integrated with New Energy Sources: A Review. *Renewable and Sustainable Energy Reviews*, **144**, Article ID: 111048. <https://doi.org/10.1016/j.rser.2021.111048>

- [23] Peterson, K.L., Chavdarian, P., Islam, M. and Cayanan, C. (2008) Tackling Ship Pollution from the Shore. *IEEE Industry Applications Magazine*, **15**, 56-60. <https://doi.org/10.1109/MIAS.2008.930901>
- [24] Wang, S. (2021) Vibration-Based Damage Imaging in Structures Using High-Speed Camera with Digital Image Correlation. North Carolina State University, Raleigh, NC. <https://doi.org/10.1117/12.2585246>
- [25] Wang, S., *et al.* (2019) An Efficient Augmented Reality (AR) System for Enhanced Visual Inspection. *IWHSM 2019: The 12th International Workshop on Structural Health Monitoring*, Stanford, CA, 10-12 September 2019.
- [26] Wang, S., Zargar, S.A. and Yuan, F.-G. (2021) Augmented Reality for Enhanced Visual Inspection through Knowledge-Based Deep Learning. *Structural Health Monitoring*, **20**, 426-442. <https://doi.org/10.1177/1475921720976986>
- [27] Yuan, F.-G., Zargar, S.A., Chen, Q. and Wang, S. (2020) Machine Learning for Structural Health Monitoring: Challenges and Opportunities. *Sensors and Smart Structures Technologies for Civil, Mechanical, and Aerospace Systems*, **11379**, Article ID: 1137903. <https://doi.org/10.1117/12.2561610>
- [28] Ren, F., Wang, S., Liu, Y. and Han, Y. (2022) Container Ship Carbon and Fuel Estimation in Voyages Utilizing Meteorological Data with Data Fusion and Machine Learning Techniques. *Mathematical Problems in Engineering*, **2022**, Article ID: 4773395. <https://doi.org/10.1155/2022/4773395>
- [29] Ren, F., Han, Y., Wang, S. and Jiang, H. (2022) A Novel High-Dimensional Trajectories Construction Network Based on Multi-Clustering Algorithm. *EURASIP Journal on Wireless Communications and Networking*, **2022**, Article No. 18. <https://doi.org/10.1186/s13638-022-02108-4>
- [30] Wang, K., *et al.* (2022) A Comprehensive Review on the Prediction of Ship Energy Consumption and Pollution Gas Emissions. *Ocean Engineering*, **266**, Article ID: 112826. <https://doi.org/10.1016/j.oceaneng.2022.112826>
- [31] Uyanık, T., Karatug, C. and Arslanoğlu, Y. (2020) Machine Learning Approach to Ship Fuel Consumption: A Case of Container Vessel. *Transportation Research Part D: Transport and Environment*, **84**, Article ID: 102389. <https://doi.org/10.1016/j.trd.2020.102389>
- [32] IMO (2009) Resolution MEPC. 1/Circ. 684, Guidelines for the Voluntary Use of the Ship Energy Efficiency Operational Indicator. EEOI.
- [33] Maas, A.L., Hannun, A.Y. and Ng, A.Y. (2013) Rectifier Nonlinearities Improve Neural Network Acoustic Models. *Proceedings of the 30th International Conference on Machine Learning*, **30**, 3.
- [34] Bostanabad, R., Kearney, T., Tao, S., Apley, D.W. and Chen, W. (2018) Leveraging the Nugget Parameter for Efficient Gaussian Process Modeling. *International Journal for Numerical Methods in Engineering*, **114**, 501-516. <https://doi.org/10.1002/nme.5751>
- [35] Hoerl, A.E. and Kennard, R.W. (1970) Ridge Regression: Biased Estimation for Nonorthogonal Problems. *Technometrics*, **12**, 55-67. <https://doi.org/10.1080/00401706.1970.10488634>