Enhancing transfer learning strategies for ship fuel consumption prediction under data scarcity

Downloaded from: https://research.chalmers.se, 2026-02-28 06:29 UTC

Citation for the original published paper (version of record):

Fan, A., Sun, S., Hu, Z. et al (2026). Enhancing transfer learning strategies for ship fuel consumption prediction under data scarcity. Ocean Engineering, 351.
http://dx.doi.org/10.1016/j.oceaneng.2026.124398

N.B. When citing this work, cite the original published paper.

research.chalmers.se offers the possibility of retrieving research publications produced at Chalmers University of Technology. It covers all kind of research output: articles, dissertations, conference papers, reports etc. since 2004. research.chalmers.se is administrated and maintained by Chalmers Library

(article starts on next page)

Ocean Engineering 351 (2026) 124398

Contents lists available at ScienceDirect

Ocean Engineering

journal homepage: www.elsevier.com/locate/oceaneng

Research paper

Enhancing transfer learning strategies for ship fuel consumption prediction under data scarcity

Ailong Fan [a,b,c], Siyang Sun [a,b], Zhihui Hu [d,*], Nikola Vladimir [e], Wengang Mao [f]

[a] State Key Laboratory of Maritime Technology and Safety, Wuhan University of Technology, Wuhan, China
[b] School of Transportation and Logistics Engineering, Wuhan University of Technology, Wuhan, China
[c] East Lake Laboratory, Wuhan, China
[d] Navigation College, Jimei University, Xiamen, China
[e] Faculty of Mechanical Engineering and Naval Architecture, University of Zagreb, Croatia
[f] Chalmers University of Technology, Sweden

ARTICLE INFO

Keywords:
Ship fuel consumption prediction
Data scarce
Transfer learning
Freezing strategy
Long short-term memory neural network
Random-forest algorithm

ABSTRACT

Reliable fuel consumption (FC) prediction is crucial for enhancing the energy efficiency of ships and achieving low-carbon shipping. However, the scarcity of individual ship data due to limited operation time or sensor failures remains a major obstacle to developing accurate data-driven models. This study proposes a transfer learning framework to address this challenge, which includes two model structures: bidirectional long short-term memory network (BiLSTM) and random forest (RF). By using the operation data of similar ships with sufficient historical records as the source domain, it supports FC prediction for target ships with limited data. Experimental results show that the performance of both transfer learning models is superior to that of the baseline model and the mixed data model. Compared with the baseline model, the MAE of the TL-BiLSTM and TL-RF models is reduced by 42 % and 36 %, respectively. The paper also innovatively and systematically analyzes the influence mechanism of the freezing strategy and the source-target sample ratio on the transfer performance. The proposed method provides an effective solution for FC prediction in data-scarce situations, can provide practical guidance for ship energy efficiency management.

Nomenclature

Abbreviation	Full Form
ANN	Artificial Neural Network
BiLSTM	Bidirectional Long Short-Term Memory
BPNN	Back Propagation Neural Network
CNN	Convolutional Neural Network
CORAL	Correlation Alignment
DT	Decision Tree
FC	Fuel Consumption
FFNN	Feed-Forward Neural Network
GNN	Graph Neural Network
IMO	International Maritime Organization
IQR	Interquartile Range
KNN	K-Nearest Neighbors
LM-BP	Levenberg-Marquardt Backpropagation
LSTM	Long Short-Term Memory
MAE	Mean Absolute Error
MAPE	Mean Absolute Percentage Error
MMD	Maximum Mean Discrepancy
PCA	Principal Component Analysis
PC1	Principal Component 1
PC2	Principal Component 2
PC3	Principal Component 3
R^2	Coefficient of Determination
RF	Random Forest
RFR	Random Forest Regression
RNN	Recurrent Neural Network
RMSE	Root Mean Square Error
RR	Ridge Regression
SVR	Support Vector Regression
TR-PSO-BPNN	Trust Region-based Particle Swarm Optimization Back Propagation Neural Network
TEU	Twenty-foot Equivalent Unit
TL-BiLSTM	Transfer Learning Bidirectional LSTM
TL-RF	Transfer Learning Random Forest
XGBoost	eXtreme Gradient Boosting
XGBoost-IGWO-LSTM	XGBoost-Improved Grey Wolf Optimizer-LSTM

* Corresponding author.
E-mail address: hzh@stu.shtmu.edu.cn (Z. Hu).

https://doi.org/10.1016/j.oceaneng.2026.124398
Received 8 September 2025; Received in revised form 11 January 2026; Accepted 21 January 2026
Available online 4 February 2026
0029-8018/© 2026 Published by Elsevier Ltd.

1. Introduction

1.1. Research background

With the rapid growth of global trade, the fuel consumption and emissions of ships have become a major concern, directly affecting operating costs and contributing to climate change (Fan and Li, 2023a). The International Maritime Organization (IMO) has introduced a series of mandatory measures for energy conservation and emission reduction, highlighting the need for accurate monitoring and management of ship energy consumption (Fan and Xiong, 2023).

Accurate prediction of ship fuel consumption (FC) is essential for supporting energy efficiency optimization measures such as hull line design, main engine–propeller matching, speed and trim optimization, and route selection (Fan and Yang, 2022b; Mao and Rychlik, 2016). However, FC prediction is often challenged by insufficient or low-quality operational data, especially for newly built ships that lack long-term monitoring. Sensor failures, incomplete manual reports, and limited test data further exacerbate the data scarcity problem.

In such scenarios, conventional machine learning methods may struggle to deliver reliable predictions due to the "data famine" issue. Transfer learning provides a promising solution by reusing knowledge from similar ships or historical datasets to improve prediction accuracy in data-scarce target domains. This study investigates the performance of transfer learning for FC prediction under limited data conditions and proposes two architectures based on different transfer learning modes, aiming to enhance prediction accuracy and support practical energy efficiency optimization.

1.2. Literature review

1.2.1. Data-driven prediction of ship FC

In recent years, machine learning has achieved remarkable results in various fields (Akande and Okolie, 2025; Rahman and Jamal, 2024; Sulaiman and Mustaffa, 2024). In the area of ship fuel consumption prediction, scholars typically use machine learning algorithms (such as support vector machines, random forests, neural networks, etc.) to train on historical data, in order to learn the linear or nonlinear relationships between the target variable and influencing factors, thereby accurately predicting ship fuel consumption and providing reliable basis for ship energy efficiency management and optimization (Fan and Wang, 2025; Fan and Wang, 2024; Fan and Yang, 2022a; Lang and Wu, 2022).

Early studies such as the LM-BP model proposed by (Shu and Yu, 2024) mainly relied on a small number of structured parameters such as ship speed and draught to construct the base regression relationship, at which time the data dimensions and complexity were relatively homogeneous. And with the introduction of complex algorithms, researchers began to integrate heterogeneous data from multiple sources such as ship operating attitude and environmental meteorology, and the dimension of data collection was extended to more than 10 parameter indicators. Then the emergence of integrated learning methods, with its high interpretability and adaptability to structured data quickly dominated, such models extract the nonlinear relationship between ship speed, load, meteorology and other features through the mechanism of multi-decision tree integration, which can achieve high prediction accuracy in structured tabular data. In recent years, deep learning methods have further focused on the deep mining of time-series data, capturing the long-term dependency of dynamic sequence data such as main engine operating conditions and sailing status through RNN structure, and its predictive performance improvement directly relies on large-scale and high-frequency sensor datasets. For example (Wang and Hua, 2023), used LSTM neural network to establish a ship energy consumption prediction model, and experimentally verified that the model can effectively predict ship fuel usage under different operating conditions (Han and Huang, 2021). investigated a hybrid model based on XGBoost-IGWO-LSTM approach, which significantly improves the prediction results of ship FC compared to the traditional machine learning model and single model.

With scholars' step-by-step research on various data-driven ship FC prediction models, from the early linear regression based on limited parameters such as speed and draught (e.g., LM-BP model), to the non-relational mining of multi-source heterogeneous features by integrated learning (XGBoost, RF), and then to the dynamic parsing of time-series data by deep learning (LSTM, mixture model). Although the prediction accuracy of ship FC has been improved to a great extent, the models built are also becoming increasingly complex, and the demand for large-scale, high-frequency, high-quality full-dimensional ship feature datasets for its model inputs is also more and more obvious, which leads to a significant increase in prediction costs, and how to reduce the prediction costs as much as possible while guaranteeing the prediction accuracy has become an urgent problem in this field.

1.2.2. Transfer learning research

In the field of shipbuilding and marine engineering, transfer learning has demonstrated its potential to address the issue of data scarcity (Luo and Zhang, 2025). constructed a transfer learning strategy based on ANN by leveraging the knowledge of seven other container ships to achieve FC prediction for the target ship, and explored the impact of different amounts of data used in the target domain on the transfer effect (Mavroudis and Tinga, 2025). explored using a FFNN to train a surrogate model to integrate physics-based and data-driven models, and proposed a new method to combine data fitted by a physics-based model with actual operating condition data through transfer learning to improve the accuracy of the ship shaft power prediction model (Li and Lin, 2024). utilized the transfer learning mechanism based on TR-PSO-BPNN to transfer model parameters across domains to solve the problem of segment construction time prediction in ship manufacturing under the small sample scenario. The proposed strategy reduced the MAPE by over 42.4 % compared to the traditional BPNN model (Deng and Li, 2024). first introduced transfer learning to ship shaft alignment correction, using the designed shaft simulation data to train the neural network, inputting the measured data into the pre-trained model, and achieving knowledge transfer through parameter fine-tuning, reducing the reliance on measured data (Xi and Ma, 2025), proposed a transfer learning method that combines two-stage TrAdaBoost.R2 and SVR to predict the operation time of underwater gliders (UGs) in order to address the unpredictability of the marine environment and the insufficiency of actually available data in new tasks or environments.

It is worth noting that the application of transfer learning in the field of ships is not limited to regression prediction tasks. In classification tasks, such as image recognition, fault diagnosis and status assessment, transfer learning is also widely adopted and has achieved remarkable results (Qiao and Liu, 2020). proposed a transfer learning method based on the InceptionV3 model for specific small sample scenarios in ship image classification and recognition, and proved that the prediction accuracy reached 98 % through experiments (Milicevic and Zubrinic, 2018; Yang and Yang, 2021); both utilized the CNN algorithm based on transfer learning, using the pre-trained VGG-19 model as the basis, freezing the pre-trained layer, and only fine-tuning the fully connected layer to solve the classification problem of ship spare parts and fine-grained ship types under the data scarcity scenario, and achieving an average model accuracy of 96.36 % through five-fold cross-validation of real cases (Cheng and Li, 2023); introduced semi-supervised adversarial transfer learning into the sea condition estimation field, proposed a SAFENESS framework, using the data alignment algorithm and multi-class adversarial discriminator to achieve knowledge transfer across ship types or load states (Wu and Wang, 2024); addressed the problem of ship navigation safety risk warning, through transfer

learning to share the common features of the source sea area and the target sea area, retaining the specific features of the target sea area, breaking through the geographical limitations of data, and achieving the generalization application of the model in untrained sea areas.

The successful application of transfer learning is not confined to the field of ships. Its value in addressing data scarcity and enhancing the generalization ability of models has also been fully verified in other engineering and technical fields. In the field of electric vehicles, transfer learning is employed to address the issue of poor model performance in extreme scenarios and to enhance the prediction accuracy of models for vehicle categories with insufficient data.(Tian and Liu, 2024; Wilbur and Mukhopadhyay, 2021; Xie and Jiang, 2025). In the building energy consumption field, researchers migrate similar building data through time series decomposition and seasonal trend adjustment technology, or use BIM simulation data to pre-train the model and then transfer it to the measured scenario (Bellagarda and Cesari, 2022; Ribeiro and Grolinger, 2018); In the new energy field, sample-based transfer learning is used to filter beneficial source domain data for the target domain to enhance the target domain, combined with stacked LSTM to improve the accuracy of photovoltaic prediction in data scarcity scenarios (Elissaios and Nikos, 2022), or through RNN model transfer to predict the production data of multiple wells to achieve more reasonable exploitation (Mohd Razak and Cornelio, 2022). In the commercial finance field, a dual-source transfer model is constructed to optimize financial time series prediction (He and Pang, 2019). In the retail field, deep neural networks and expert knowledge are integrated to solve the problem of insufficient new product data (Karb and Kühl, 2020); In the public health field, cross-border epidemic model migration is used to achieve dynamic prediction of COVID-19 (Gautam, 2022). In graph machine learning, transfer learning is utilized to enhance the performance of graph neural networks in classification task (Han and Liu, 2024). In information systems, transfer learning based on Transformer has been proposed for long time series data to enhance the performance of tasks such as prediction and anomaly detection (Gruetzemacher and Paradice, 2022).

Research conducted by domestic and foreign scholars on the application of transfer learning in various fields has demonstrated the remarkable effects of transfer learning from multiple perspectives, such as its ability to address data scarcity, enhance model generalization capabilities, and integrate cross-domain knowledge. This has verified the effectiveness and practicality of transfer learning as an innovative machine learning paradigm in regression prediction tasks.

1.3. Research gap and paper contributions

Recent studies have extensively employed machine learning (ML) and other data-driven approaches to predict ship fuel consumption (FC), achieving encouraging accuracy through models such as artificial neural networks (ANN), feedforward neural networks (FFNN), XGBoost, and random forests (RF). However, the effectiveness of these models largely depends on a large amount of high-quality operational data. In practical operations, newly built ships, ships with faulty sensors, or those relying on manual logs often face issues of limited or incomplete data, which significantly reduces the prediction performance.

To address the limitations of purely data-driven models, researchers have explored various solutions, including physics-informed machine learning and physics-guided machine learning, which incorporate physical knowledge to enhance model robustness in data-scarce scenarios (Lang and Wu, 2024). However, these methods are highly dependent on semi-empirical formulas or differential equations based on ship propulsion principles, making their implementation complex. Moreover, due to the unknown or inaccurate physical relationships of some new ship types, in extremely data-scarce situations (with only dozens or hundreds of data points), physics-informed models may not be able to fully calibrate physical parameters or residual models, leading to a significant decline in model performance. In such cases, transfer learning may be a better solution, which utilizes knowledge from related source domains to improve the prediction accuracy of data-scarce target domains. In the shipping industry, transfer learning has been applied to predict fuel consumption and biofouling, However, the existing research (Luo and Zhang, 2025) mainly focuses on neural network architectures such as ANN and FFNN, with the research subjects all being sister ships, resulting in poor generalization ability. There is relatively less exploration regarding the feasibility of transfer learning between non-sister ships and the key factors influencing the performance of transfer learning.

Based on this gap, this study proposes and evaluates two novel transfer learning methods for predicting ship fuel consumption:

(1) It innovatively proposes a network-based bidirectional long short-term memory (BiLSTM) and a mapping-based random forest (RF) transfer learning model. Among them, the mapping-based RF transfer learning is the greatest innovation point of this study, providing a broader perspective for transfer learning research that only focuses on neural network architectures;
(2) It systematically analyzes the impact of different freezing strategies in the bidirectional long short-term memory network structure transfer learning on the model's transfer effect, providing empirical basis for the rational selection of freezing and fine-tuning layers;
(3) Through controlled variable experiments, it reveals the dynamic impact of the sample ratio between the source domain and the target domain on the performance of transfer learning, providing practical guidance for the deployment of model sample sizes in the source domain and the target domain in data-scarce scenarios.

2. Methodology

The structure of this article is as follows: Section 2 introduces the principles, methods, and evaluation metrics of transfer learning; Section 3 explains data preprocessing and feature selection; Section 4 presents the experimental results and comparative analysis; Section 5 discusses the influencing factors of transfer learning; Section 6 summarizes the entire article and draws conclusion. The methodology and technical route of the paper are shown in Fig. 1.

2.1. Basics of transfer learning

Transfer learning refers to the process of transferring the knowledge learned from a source domain or task to a new target domain or task, in order to break through the traditional machine model's reliance on data volume and complete labeling (Pan and Yang, 2009). According to different transfer methods, transfer learning can be classified into three types: sample-based transfer learning, mapping-based transfer learning, and network-based transfer learning.

As shown in Fig. 2, sample-based transfer learning means selecting some samples from the source domain and assigning appropriate weight values to them as a supplement to the training set of the target domain.

The core idea of mapping-based transfer learning is to map the samples from the source domain and the target domain to a new data space, so that in this new space, the samples from the two domains become more similar, making them suitable for training a joint model. The general process is shown in Fig. 3.

Network-based transfer learning transfers knowledge from the source domain to the target domain by reusing the underlying network structure and parameters of the pre-trained model. As shown in Fig. 4, the neural network is similar to the processing mechanism of the human brain, being an iterative and continuous abstract process. The first layers of the network can be regarded as feature extractors, specifically responsible for extracting universal features from the input data. These features are not only valuable for the current task, but also often able to function across tasks and domains, while the higher-level networks learn

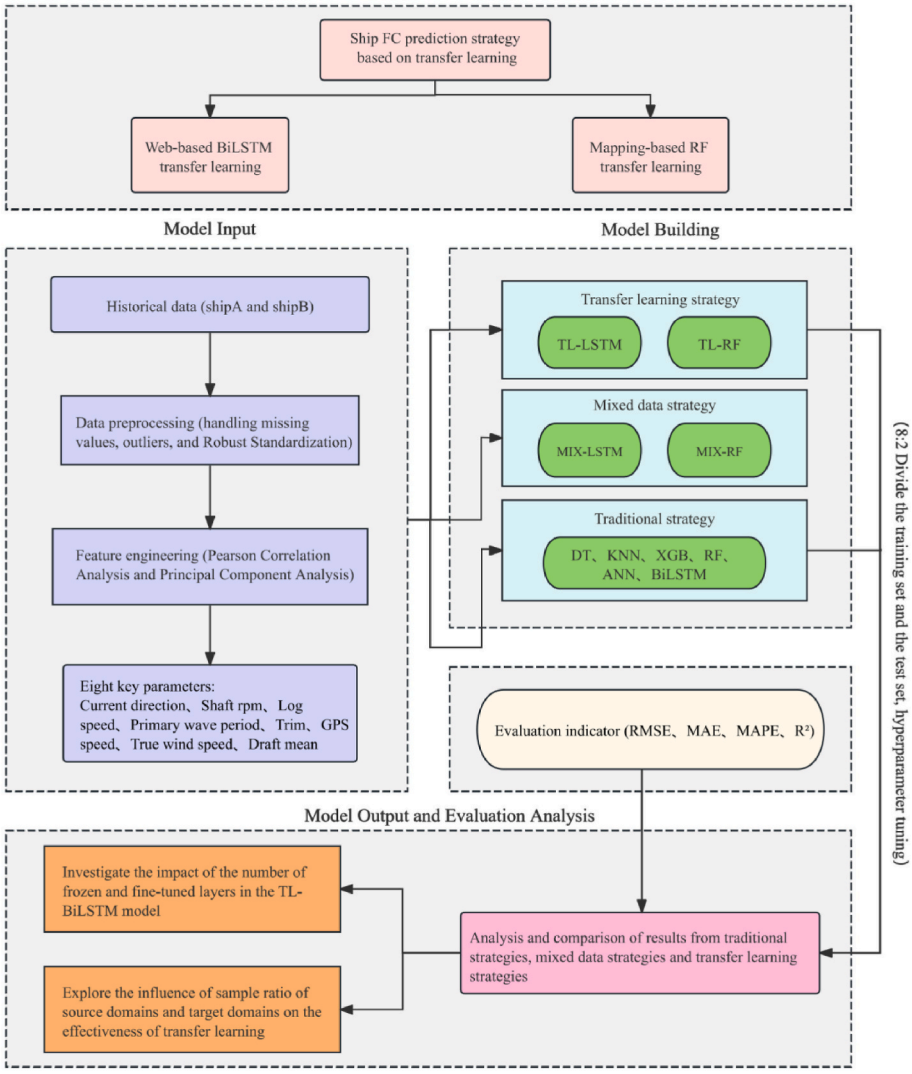

Fig. 1. Ship FC prediction model based on transfer learning.

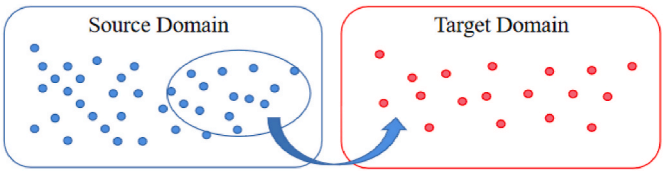

Fig. 2. Sample-based transfer learning.

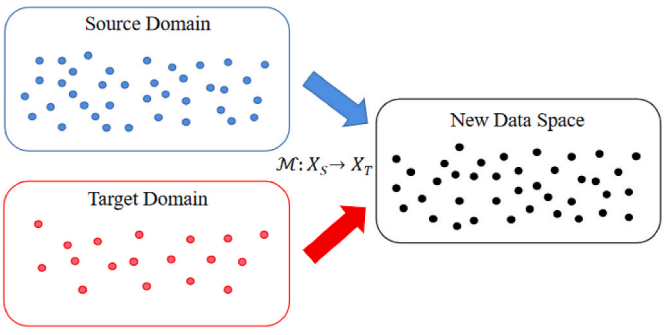

Fig. 3. Mapping-based transfer learning.

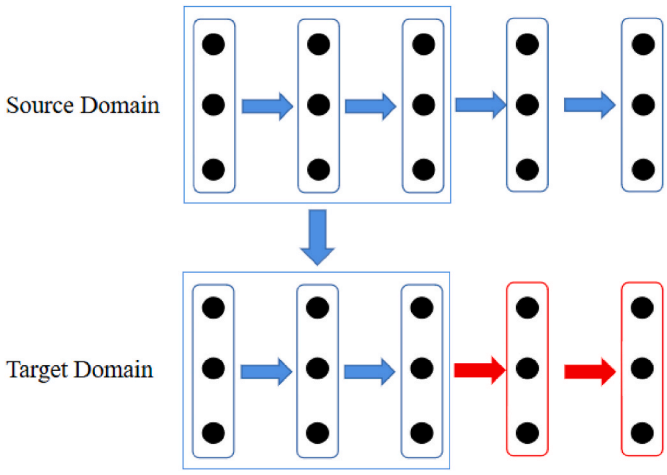

Fig. 4. Network-based transfer learning.

task-specific advanced semantics and adjust through fine-tuning or domain adaptation techniques to adapt to new tasks.

2.2. Transfer learning strategies

In the FC prediction task of ships, transfer learning can effectively transfer the fluid dynamics knowledge and the patterns related to ship operation energy consumption learned in the source domain to the target task through parameter fine-tuning or domain adaptation strategies.

Specifically, a large amount of historical labelled ship operation data is first used as the source domain $D_S = \{x_i, y_i\}_{i=1}^{n}$ $(i = 1, 2, \ldots, n)$ for pre-training to construct a teacher model, which is then migrated to the similar target domain $D_t = \{x_i, y_i\}_{i=1}^{m}$ $(i = 1, 2, \ldots, m)$ $(n \gg m$, i.e., the sample size in the source domain is much larger than that in the target domain), which is intended to assist the learning of the target domain $\mathscr{D}_t$ by using the knowledge of the source domain $\mathscr{D}_s$, so as to make the source domain serve as a knowledge supplement to the target domain that is lacking in samples.

This study mainly focuses on several transfer learning methods introduced in Section 2.1. Given that the ship operation dataset contains a large number of features and is usually time-ordered time series data, the BiLSTM and RF are taken as the benchmark learning algorithms, the mapping transfer learning model based on RF and the network transfer learning model based on BiLSTM are constructed.

2.2.1. Network-based BiLSTM transfer learning

(Hochreiter and Schmidhuber, 1997) first proposed the LSTM neural network model in the late 20th century. It is a neural network architecture specifically designed for processing and analyzing sequential data. The BiLSTM combines two LSTM layers: one processes the input sequence in a forward direction, while the other does so in a backward direction. This bidirectional processing method enables the network to simultaneously capture the forward and backward context information of each element in the sequence, thereby more comprehensively understanding the internal structure and patterns of sequential data.

This study proposes a network structural transfer learning framework based on bidirectional long short-term memory networks. This framework achieves the extraction and transfer of ship spatiotemporal features by transferring the neural network structure.

First, a BiLSTM network layer with 64 hidden units is constructed using the ship dataset D_{full} as the source domain to store and update information related to the sequence; to enhance the model's ability to focus on information from different positions in the input sequence, accelerate the training process, improve the model's stability and generalization ability, layer normalization layers and multi-head attention mechanism layers are introduced; then, an LSTM layer with 32 hidden units is used to continue focusing on the global pattern and capture the long-term dependencies in the sequence; finally, through three fully connected layers, the features from the previous layer are gradually received and linearly transformed and activated using the ReLU activation function to further extract or transform the features for output.

After building the neural network model structure of the source domain, the transfer is carried out by combining the trained neural network model structure of the source domain, using the freeze-microfine-tune strategy to freeze the structure and parameters of the bottom N-2 layers of the model, and then transferring it to the model structure of the target domain ship dataset D_{scarity}. This is done to retain the general spatiotemporal feature extraction ability of ships, including capturing factors such as propeller power characteristics, speed, and fuel consumption relationships. Then, in the remaining top layers of the target domain (the last 2 layers), fine-tuning is carried out through domain adaptation technology to achieve feature space transformation, enabling the pre-trained model to adapt to the unique data distribution and features of the target domain. The specific steps are shown in Fig. 5.

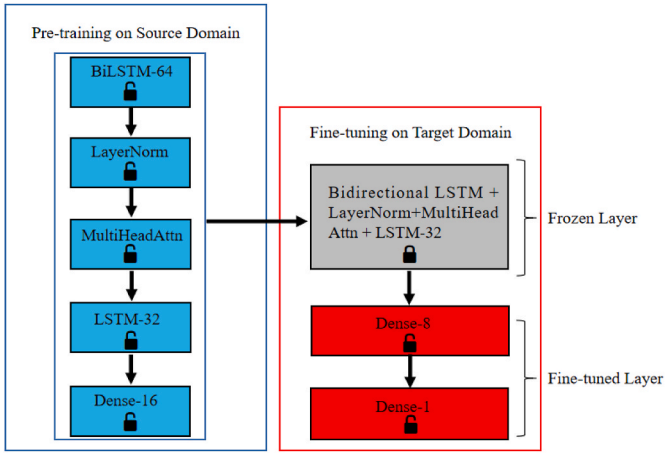

Fig. 5. The process of Network-based BiLSTM transfer learning.

2.2.2. Mapping-based RF transfer learning

RF is a tree-based model algorithm based on ensemble learning proposed by (Breiman, 2001). It improves model performance by constructing multiple decision trees and integrating their prediction results. The core idea is to use the bootstrap method to randomly select multiple sub-samples from the original data to train decision trees, and at each node split of each tree, only a portion of the features are randomly selected for evaluation. This algorithm can efficiently handle high-dimensional feature datasets, is robust to missing values and outliers.

Correlation Alignment (CORAL) is an effective domain adaptation method (Sun and Saenko, 2016). Its core idea is to reduce the distribution discrepancy between the source domain and the target domain by aligning their covariance matrices and mapping them to a common feature space through linear transformation, where the distributions of the two domains become more similar, thereby effectively solving the domain adaptation problem.

Under the theoretical framework of transfer learning, this study proposes a mapping-based random forest transfer learning method to address the issue of scarce data in the target domain. The method first calculates the covariance matrices of the source domain ship dataset D_{full} and the target domain ship dataset D_{scarity}. By using the CORAL method, a linear transformation matrix is derived to align the feature distributions between the two domains. The formula is as follows:

$$A = C_S^{-\frac{1}{2}} C_T^{\frac{1}{2}} \tag{1}$$

where: A is the linear transformation matrix; C_S is the covariance matrices of the source domain; C_T is the covariance matrices of the target domain.

$$X_S' = X_S A \tag{2}$$

where: X_S is the is the original data matrix of the source domain.; X_S' is the mapped data matrix.

This transformation effectively maps the source domain data to the feature distribution space of the target domain and minimizes the difference in second-order statistics between the two domains through the CORAL loss function. The formula is as follows:

$$D_{CORAL} = \frac{1}{4d^2} \| C_S - C_T \|_F^2 \tag{3}$$

where: D_{CORAL} is the CORAL loss function; d is the characteristic dimension; $\|\cdot\|_F$ is the Frobenius norm.

The alignment process is guided by an optimization objective to reduce the distribution difference metric, ensuring that the transformed source domain features are statistically similar to the target domain

features. In the newly created joint feature space, the rich knowledge contained in the source domain data can effectively compensate for the insufficiency of target domain samples, thereby achieving the purpose of knowledge transfer. The algorithm process can be summarized as follows:

Step1: Calculate the covariance matrices of the source domain and target domain features, and use the CORAL method to learn a linear transformation matrix to map the source domain data to a space with a distribution similar to that of the target domain.
Step2: Merge the aligned source domain data with the original target domain data to form a dataset with consistent distribution and enhanced samples.
Step3: Train a random forest model on this joint dataset. The model acquires better generalization performance by learning this target domain data that integrates the calibrated source domain knowledge.

The technical implementation framework is shown in Fig. 6.

2.3. Evaluation of model performance index

In this paper, the following indicators are used to evaluate the effectiveness of the model predictions:

(1) Root Mean Square Error (RMSE)

RMSE is a way to measure the difference between the predicted value and the true value, and in practice, the smaller the value of RMSE, the more accurate the prediction of the model is, and its unit is the same as that of the original data. The formula is as follows:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{4}$$

where: RMSE denotes the root mean square error; n is the total number of observations, i.e., the sample size; y_i is the actual value of the i observation; $\hat{y}_i$ is the predicted value of the i observation.

(2) Mean absolute error (MAE)

MAE is the average of the absolute values of the difference between the predicted and actual values of the model. Compared to MSE and RMSE, MAE penalizes the prediction error more linearly and therefore it is less sensitive to outliers (i.e. extreme errors). The formula is calculated as follows:

$$\mathrm{MAE} = \frac{1}{n}\sum_{i=1}^{n}|y_i - \hat{y}_i| \tag{5}$$

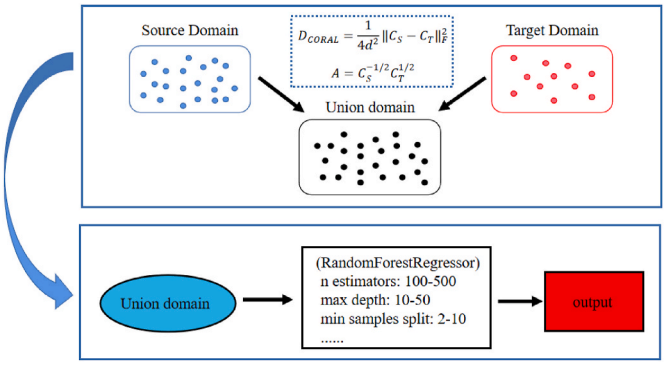

Fig. 6. The process of Mapping-based RF Transfer Learning.

where: MAE denotes the mean absolute error; n is the total number of observations, i.e., the sample size; y_i is the actual value of the i observation; $\hat{y}_i$ is the predicted value of the i observation.

(3) Mean absolute percentage error (MAPE)

MAPE is a statistical indicator that measures the degree of difference between predicted and actual values. It is often used to assess the accuracy of forecasting models, especially in time series forecasting and regression analysis. The formula is as follows:

$$\mathrm{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left(\left|\frac{y_i - \hat{y}_i}{y_i}\right| \times 100\%\right) \tag{6}$$

where: MAPE denotes the mean absolute percentage error; n is the total number of observations, i.e., the sample size; y_i is the actual value of the i observation; $\hat{y}_i$ is the predicted value of the i observation.

(4) Coefficient of determination (R^2)

R^2 is a statistic of the goodness of fit of the regression model, which indicates the degree of correlation between the predicted and actual values of the model. The value of R^2 lies between 0 and 1, and the closer the value is to 1, the better the fit of the model. The formula is as follows:

$$R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{7}$$

where: R^2 indicates the goodness of fit of the model; n is the total number of observations, i.e., the sample size; y_i is the actual value of the i observation; $\hat{y}_i$ is the predicted value of the i observation.

3. Case study

3.1. Case ship

Two container ships, ship A and ship B, are used in this study. Both of them belong to the large ocean-going container ships of over 9000 TEUs. Although there are certain differences in size and specifications between these two ships (Table 1), The underlying physical laws related to energy consumption such as the navigation resistance, the relationship between speed and power, and the loading conditions have similarities. Although the distribution of external environmental variables is different, the mechanisms that affect energy consumption are similar (for example, an increase in wind speed leads to an increase in fuel consumption). This provides the necessary conditions and theoretical basis for achieving cross-domain feature alignment through transfer learning.

3.2. Data pre-processing

Under the condition of meeting the migration prerequisites, this paper selects the working condition data sets of ship A from January 3, 2017 to May 30, 2017, and ship B from October 31, 2017 to November 7, 2017. The parameter characteristics include time, Shaft rpm, GPS speed,

Table 1
Specification parameters of ship A and ship B.

Ship type	Ship A	Ship B
Host model	MAN-B&W 8S90ME-C10.5	WinGD 7X92
Total length (m)	299.9	366.0
Width (m)	48.2	51.2
Depth (m)	24.8	30.2
Draught (m)	9.5	14.1
Capacity (TEU)	9400	14566

Log speed, True wind speed, Trim, Relative wind speed, Relative wind direction, Current direction, Primary wave direction, Wind wave direction, Primary wave period, Draft mean, Fuel consumption rate.

First, the original data of the two ships are cleaned and filled to ensure the quality and accuracy of the data. The missing values are filled using interpolation and forward filling methods. Under the condition of maintaining the trend and change of the data distribution, reasonable values are filled in the time dimension to conform to the distribution pattern of the original data; abnormal values are mainly screened out after filtering out outliers and data with obvious problems collected by sensors to ensure that the prediction model is not affected by extreme sample points. The processed source domain (ship A) sample size is 3005, and the target domain (ship B) sample size is 489. The data distribution of both is shown in Table 2.

The Maximum Mean Discrepancy (MMD) is a statistical metric based on kernel methods, used to measure the difference between two probability distributions and widely applied in domain adaptation tasks in transfer learning. When MMD is less than 0.05, it is generally considered that the two task distributions are very close. A value between 0.05 and 0.2 indicates a certain difference, but it can be eliminated through domain adaptation methods. A value greater than 0.3 indicates a significant distribution difference, which may lead to negative transfer. To verify whether the domain difference is significant and determine the feasibility of transfer between the two domains, we used the MMD distance to measure the variable differences of the two ships, and the results are shown in Fig. 7. Except for "Current direction" and "relative wind direction", the MMD distances of most variables are within the moderate difference range of 0.1–0.3. The average MMD distance metric between the two domains is approximately 0.2, indicating that transfer is possible.

Thus, Robust Scaler is used to further standardize the data. It centralizes the data using the median and the IQR, scaling data with different magnitudes into the same interval, thus eliminating the negative effects of size differences and different tonnages when migrating the same type of ship, and at the same time avoiding distortions of the scaling results from outliers, making the scaling results significantly less sensitive to the sensitivity to extreme values is significantly reduced. The formula is as follows:

$$X_{\text{scaled}} = \frac{X_i - \text{median}(X)}{\text{IQR}(X)} \tag{8}$$

where: IQR is the 75th percentile (Q3) minus the 25th percentile (Q1), X_i denotes the sample of data for each ship, and X_{scaled} denotes the data after normalization.

After the missing value filling, outlier correction and Robust normalization, it is necessary to screen the features that affect the fuel consumption of the ship to extract the features that have a greater impact on the fuel consumption of the main engine as inputs to the model, in order to improve the prediction accuracy and interpretability of the model.

Table 2
The distribution of the processed data results.

	Ship A (count:3005)			Ship B (count:489)		
	Min	Mean	Max	Min	Mean	Max
Shaft rpm (RPM)	12.06	61.27	74.09	51.47	59.81	63.37
GPS speed (kn)	1.65	17.65	20.48	10.66	15.43	20.56
Log speed (kn)	0.68	17.40	19.70	11.50	15.18	19.45
Relative wind speed (kn)	1.80	7.33	12.70	3.51	10.34	13.76
Current direction (°)	17.80	106.41	358.60	14.69	42.25	69.80
Primary wave direction (°)	32.25	129.19	147.80	3.10	108.92	126.77
Relative wind direction (°)	10.10	85.07	121.19	4.28	133.79	210.94
Wind wave direction (°)	18.64	95.27	143.20	4.58	103.81	165.80
Primary wave period (s)	2.70	4.94	5.90	2.70	3.94	4.86
Ture wind speed (kn)	1.51	11.47	18.96	0.83	9.24	17.89
Draft mean (m)	12.79	13.03	13.51	10.29	11.32	12.68
Trim (m)	0.48	0.81	1.02	0.09	0.94	1.60

3.3. Feature engineering

Pearson Correlation Coefficient r is a core indicator in statistics to measure the degree of linear correlation between two continuous variables, which takes the range of [−1, 1]: when r = 1, it indicates that the two variables are completely positively and linearly correlated; when r = −1, it is completely negatively and linearly correlated; and when r = 0, it indicates no linear association.

Principal Component Analysis (PCA) is a classic unsupervised dimensionality reduction method that maps high-dimensional data to a low-dimensional space through linear transformation while retaining the main information in the data as much as possible. Its core idea is to find the directions with the largest variance in the data and use these directions to reconstruct the data to achieve feature compression and denoising.

In this study, multi-dimensional feature screening of ship operating conditions dataset was carried out based on Pearson Correlation Coefficient and PCA. As shown in Figs. 8 and 9, the correlation between the Shaft rpm, Log speed and Main Engine Fuel Consumption is significant (r >0.7), and the correlation between GPS speed, Current direction, Primary wave period, and True wind speed and the target variables is also certain. Furthermore, through PCA dimensionality reduction to eliminate multicollinearity interference, the first three principal components (PC1, PC2, and PC3) were selected for analysis. To identify key features exhibiting high correlation with the principal components, this study computed the Pearson Correlation Coefficients between the original variables and the principal component scores. Core variables highly correlated ($|r| > 0.6$) with the principal components were selected for subsequent modeling. The correlation between Shaft rpm and PC1 is 0.71, between Log speed and PC1 is 0.80, between GPS speed and PC2 is 0.84, between Trim and PC3 is 0.69, between Current direction and PC1 is 0.65, and between Primary wave period and PC3 is 0.65. Selecting these variables allows for the integration of highly correlated information from PC1, PC2, and PC3, ensuring that the model can effectively capture the main directions of variation in the data.

Finally, eight key parameters are identified as input variables of the model: Primary wave direction, Shaft rpm, Log speed, Primary wave period, Trim, GPS speed, True wind speed, Draft mean. These feature parameters together constitute the core input variables of the ship fuel efficiency prediction model, and their cumulative variance contribution rate reaches 80.20 %, which fully verifies the statistical significance of the feature screening system.

4. Results analysis

4.1. Prediction results

After data preprocessing and feature engineering, the dataset of ship A is used as the source domain and the dataset of ship B is used as the target domain for prediction. In order to verify the necessity of transfer learning and ensure the reasonableness of the experimental conclusions, this study compares the two transfer learning strategies mentioned in Section 2.2 with two groups of control experiments, the control group ① uses the traditional DT, KNN, XGB, RF, ANN, BiLSTM as the benchmark model to compare with the transfer learning model. At the same time, a control group ② was also set up where the data of ship A and ship B were directly mixed to form the target domain, and then RF and BiLSTM were used for prediction to determine whether the same effect of transfer learning could be achieved. The basic flow of the experiment is shown in Fig. 10. To ensure the accuracy of the experimental results, five-fold

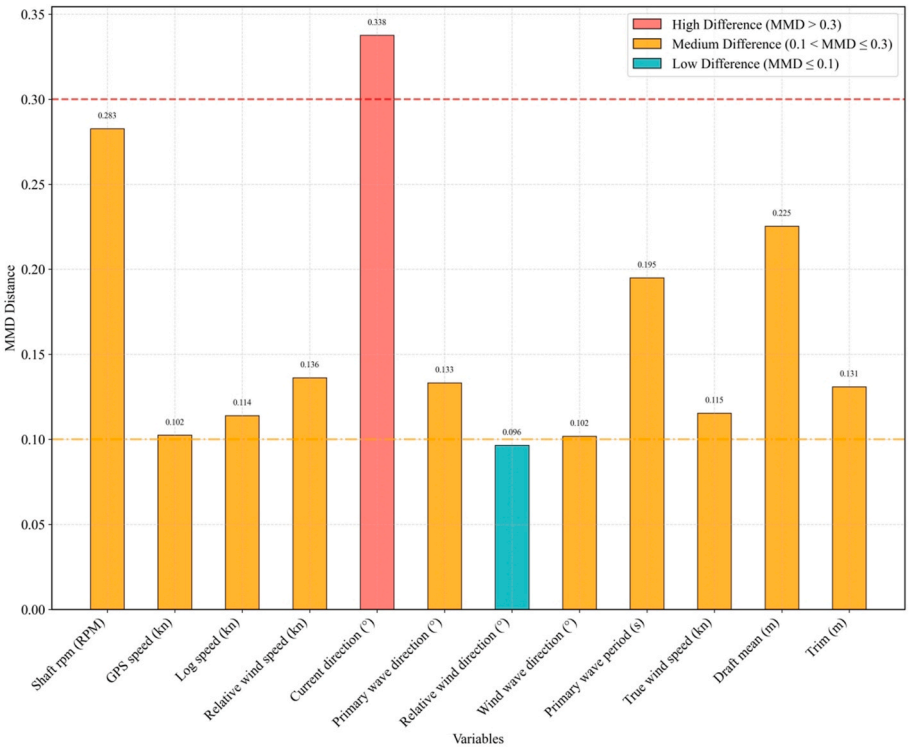

Fig. 7. The MMD distance metric between Ship A and Ship B.

time series cross-validation was conducted for each group of experiments, the expanding window method was adopted, and throughout the training process, the training set and the test set were divided in an 8:2 ratio, while identical optimal hyperparameter configurations were maintained for both the experimental and control model groups, as shown in Appendix A. All processes were implemented using libraries including pandas, numpy, and scikit-learn in Python 3.11 (64-bit), and the experiments were performed on a workstation with an Intel Core i7-10750H CPU, 16 GB RAM. The final prediction results are presented in Table 3.

4.2. Analysis of baseline model and mixed data model

By systematically comparing the visualized prediction results of control group ① (as shown in Fig. 11). In terms of the error index system, the XGBoost and RF models under the ensemble learning paradigm demonstrate significant predictive advantages. Their RMSE and MAE values are consistently below 0.50, confirming that ensemble methods can effectively capture complex data patterns through the collaborative integration of multiple base learners. In contrast, the error performance of ANN, BiLSTM, and KNN is slightly inferior to that of the ensemble models, indicating that in data-scarce scenarios, the predictive capabilities of deep learning models and distance-based metric approaches are constrained, and their accuracy is generally lower than that of ensemble tree models. Among all models, the DT model exhibits significantly higher error values across various metrics. This is attributable to the tendency of a single decision tree to overfit and capture noise in small-sample settings, which increases prediction errors and undermines generalization performance.

In terms of model fitting performance, the comparison of the coefficient of determination R^2 further supports the above conclusion. The R^2 values for XGB and RF are both 0.98, indicating that their independent variables can explain over 95 % of the variance in the dependent variable, thereby confirming the strong explanatory power of ensemble tree models in high-dimensional feature spaces. For ANN and LSTM, the R^2 values are 0.97 and 0.96, respectively, highlighting the significant dependence of deep learning models on sample size and revealing the "double-edged sword" nature of such models: while they possess powerful representational capabilities, they require substantial data to mitigate overfitting risks. When training samples are insufficient, overfitting of model parameters leads to a notable increase in generalization error. The R^2 values for DT and KNN are 0.92 and 0.97, respectively, which underscores the limitations of partitioning and neighborhood search methods based on simple rules in data-scarce scenarios.

Regarding the prediction results of control group ②, directly combining the data from ship A and ship B as the target domain yields the poorest performance. Specifically, across all three prediction accuracy metrics, this approach performs significantly worse than both baseline and transfer learning models, with its R^2 value ranking at the bottom. These results clearly indicate that simply merging source and target domain samples does not effectively address the issue of limited target domain sample size. This may be attributed to the small sample size of ship B; when data are naively combined, the proportion of ship B's samples in the merged dataset becomes even smaller and more diluted. Consequently, during training, the model is inherently biased toward learning the data distribution of ship A. As a result, when applied to predict outcomes for ship B, the model fails to capture its unique data patterns, leading to suboptimal prediction performance.

4.3. Analysis of transfer learning model

To assess the statistical significance of the performance difference between the transfer learning model and the baseline model in regression tasks, we compared the results of each model's five runs. Given that the experiments used the same hyperparameters and data splits, we employed a one-tailed paired Wilcoxon signed-rank test ($\alpha = 0.05$). The test results are shown in Tables 4 and 5.

Compared with the prediction results of the two control groups (i.e., the baseline model and the mix data model), the TL-BiLSTM and TL-RF models proposed in this study have significantly improved the prediction accuracy of the baseline model at the confidence level ($\alpha = 0.05$), and have performed excellently in terms of error and R^2 indicators.

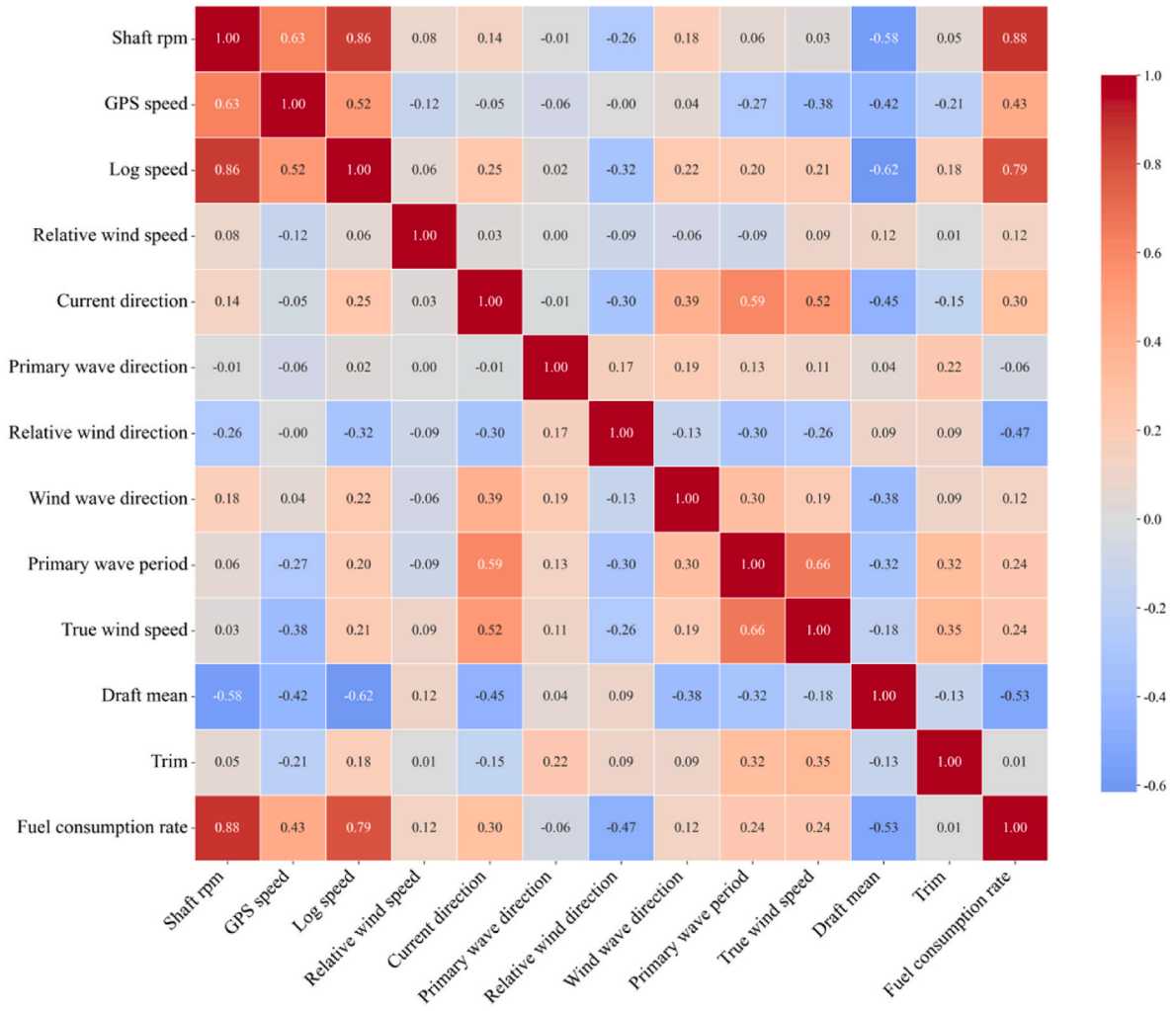

Fig. 8. Pearson correlation coefficient matrix.

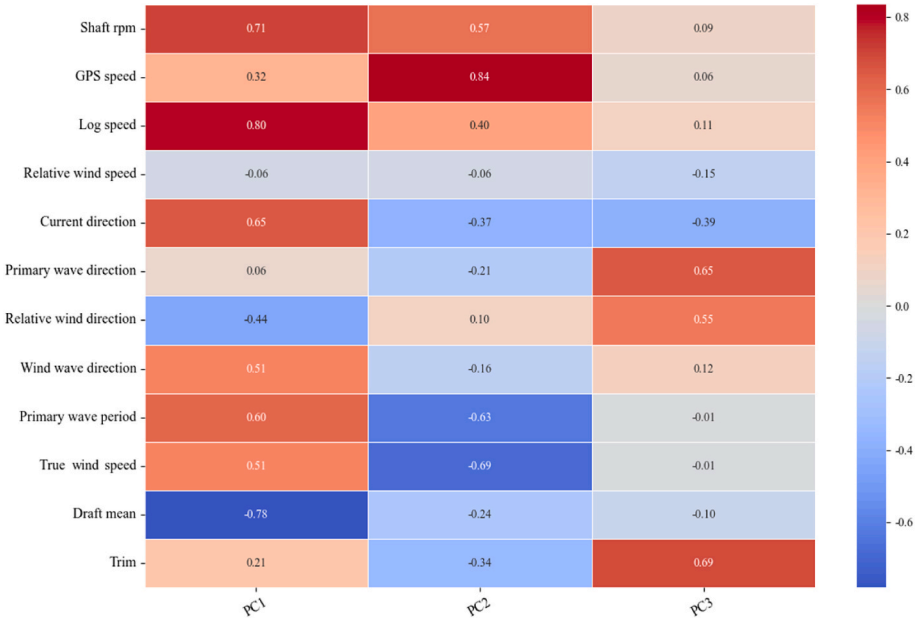

Fig. 9. PCA thermal load matrix.

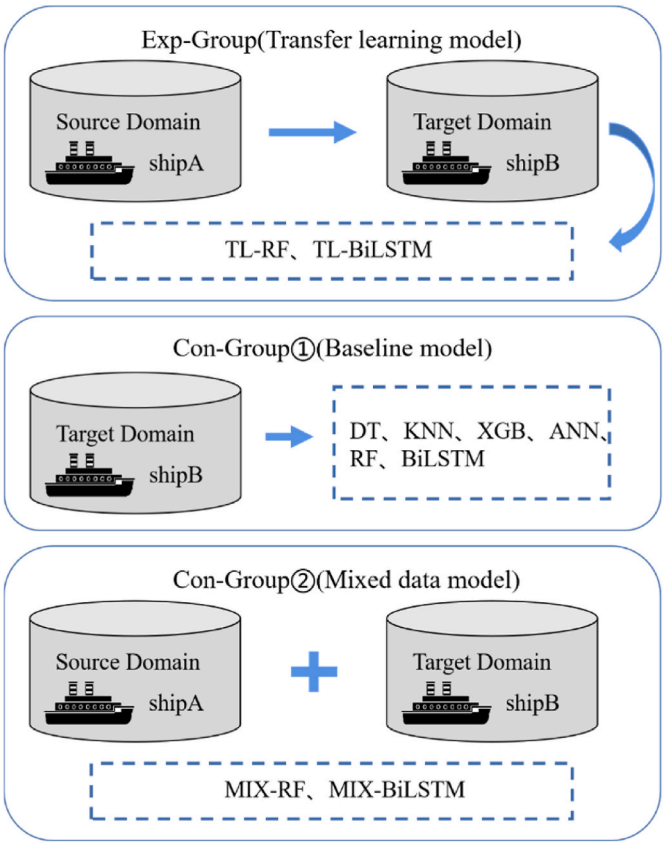

Fig. 10. Basic flow of the experiment.

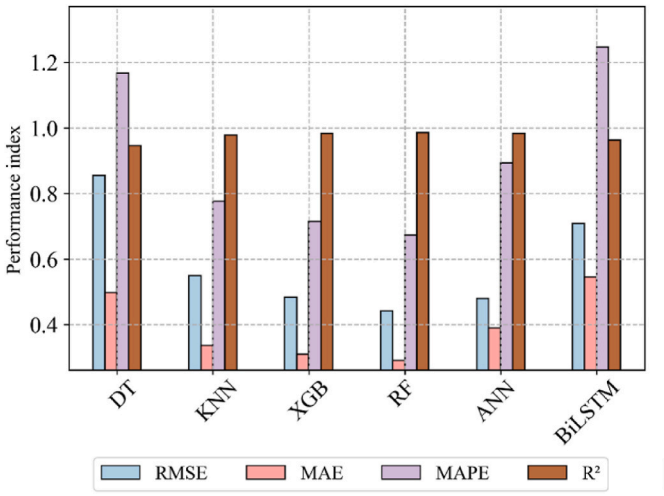

Fig. 11. Comparison of baseline model performance.

Table 3
Comparison of experimental prediction results.

	RMSE (t/day)	MAE (t/day)	MAPE (%)	R^2
DT	0.85	0.51	1.18	0.92
KNN	0.55	0.34	0.78	0.97
XGB	0.48	0.31	0.72	0.98
RF	0.46	0.28	0.66	0.98
ANN	0.45	0.37	0.84	0.97
BiLSTM	0.72	0.55	1.22	0.96
TL- BiLSTM	0.48	0.32	0.74	0.98
TL-RF	0.33	0.18	0.43	0.99
MIX-BiLSTM	3.07	1.74	9.23	0.81
MIX-RF	3.25	1.24	3.87	0.85

Table 4
The statistically significant differences between TL-BiLSTM models and baseline models.

	BiLSTM (median)	TL-BiLSTM (median)	Wilcoxon W	P value	Significant or not ($\alpha = 0.05$)
RMSE (t/day)	0.73	0.35	0.00	0.03	✓
MAE (t/day)	0.59	0.21	0.00	0.03	✓
MAPE (%)	1.34	0.48	0.00	0.03	✓
R^2	0.96	0.99	15.00	0.03	✓

Table 5
The statistically significant differences between TL-RF models and baseline models.

	RF (median)	TL-RF (median)	Wilcoxon W	P value	Significant or not ($\alpha = 0.05$)
RMSE (t/day)	0.46	0.33	0.00	0.03	✓
MAE (t/day)	0.29	0.19	0.00	0.03	✓
MAPE (%)	0.66	0.43	0.00	0.03	✓
R^2	0.98	0.99	15.00	0.03	✓

As shown in Fig. 12, in terms of error, the RMSE of TL-BiLSTM is 0.48, the MAE is 0.32, and the MAPE is 0.74 %, and the RMSE of TL- RF is 0.33, the MAE is 0.18, and the MAPE is 0.43 %, and the MAPE of TL-BiLSTM is reduced from the baseline model's 1.22 %–0.74 %, with the reduction up to 39 %, MAE from 0.55 to 0.32, a decrease of 42 %, and RMSE from 0.72 to 0.48, a decrease of 33 %; the MAPE of TL-RF decreased from 0.66 % to 0.43 % in the baseline model, a decrease of 35 %, MAE from 0.28 to 0.18, a decrease of 36 %, and RMSE from 0.46 to 0.33, a reduction of 28 %.In terms of R^2, as shown in the right axis scale in Fig. 12, the R^2 values of TL-BiLSTM and TL- RF reach 0.98 and 0.99, which are improved by 2 % and 1 % compared to 0.96 and 0.98 of the baseline models BiLSTM and RF, respectively. As shown in Fig. 13, the curves of predicted and actual values are very close to each other, indicating that the models are well fitted, not only maintaining a high degree of consistency with the real data in terms of the overall trend, but also demonstrating a keen ability to capture short-term fluctuations and sudden change points in the data. In particular, the TL- RF model is even able to predict the subtle trend fluctuations of two adjacent data samples more accurately, showing the same distribution ripples as the original data, with strong generalization and robustness.

This indicates that transfer learning, through the mechanism of learning general features by using the data of Ship A during pre-training and focusing on Ship B during fine-tuning, effectively enhances the generalization ability and cross-domain adaptation efficiency of the ship FC prediction model, thereby improving the prediction performance of the target model.

5. Discussions

5.1. Impacts of the number of frozen fine-tuned layers

In the BiLSTM-based transfer learning framework, the freezing strategy of the number of network layers directly determines the dynamic balance between the knowledge retention in the source domain and the adaptation in the target domain, so different freezing fine-tuning strategies may have different impacts on the transfer effect. The previous

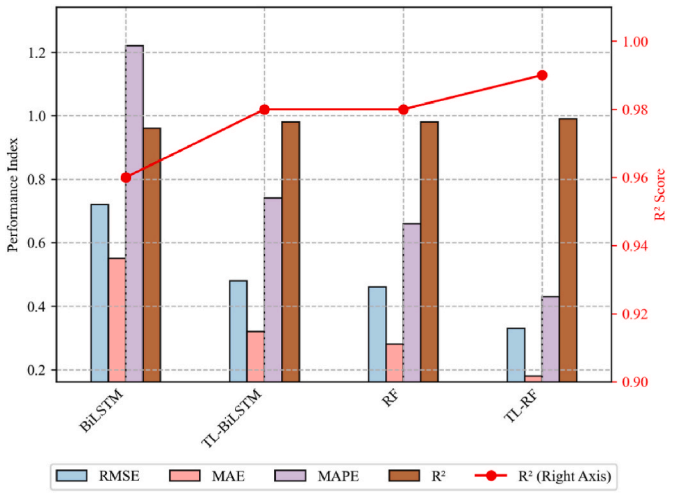

Fig. 12. Comparison of the transfer learning model with the baseline model.

strategy of fixing the freezing of the underlying N-2 layers (Fig. 5) was directly adopted in Section 2.2.1 without considering the impact of the number of frozen fine-tuned layers on the transfer learning effect, which may lead to insufficient feature capture in the target domain as well as a restricted proportion of effective transfer parameters. Therefore, further systematic experiments are needed to verify the quantitative relationship between the number of frozen layers and transfer performance.

Under the condition that the sample size ratio between the control source domain (n = 3005) and the target domain (n = 489) is (6:1) and other hyperparameters remain unchanged, the following four groups of freeze strategies experiments are set up: (a) Full Freeze Group: Only fine-tune the weights of the output layer. (b) Deep Freeze Group: Freeze the first N-3 layers and fine-tune the weights of the last 3 layers. (c) Moderate Freeze Group: Freeze the first N-4 layers and fine-tune the weights of the last 4 layers. (d) Shallow Freeze Group: Freeze the first N-5 layers and fine-tune the weights of the last 5 layers. Each group of experiments is run 5 times and the average result is taken to eliminate the influence of accidental errors on the results.

Combining the results from Table 6 shows that the deep freeze strategy (fine-tuning the last 3 layers) and the control group (fine-tuning the last 2 layers) exhibit optimal performance on the target task, with the lowest level of RMSE, MAE, and MAPE metrics, and the highest R^2value. As the number of model freeze layers decreases, i.e., more deep parameters are fine-tuned, the performance of the medium freeze (fine-tuning the last 4 layers) and shallow freeze strategies (fine-tuning the last 5 layers) shows a gradient decline, which is reflected in the gradual rise of the three prediction error metrics, and the decline of the explanatory power metric, R^2. Further, the performance of the total freeze strategy (fine-tuning only the output layer) falls off a cliff, with an average increase of more than 50 % in its error metrics and a decrease of 6 percentage points in the R^2 value, verifying the negative effect of over-retention of source domain parameters on cross-domain knowledge transfer.

Analyzing the reasons behind the above experimental results, freezing the bottom layers preserves the common feature representations from the source domain and mitigates overfitting caused by limited target domain samples, while fine-tuning the top layer enables adaptation to target-specific patterns. The deep freeze strategy achieves an optimal balance between feature retention and task adaptation. In contrast, full freezing hinders task-specific adaptation, leading to feature mismatch, whereas shallow freezing disrupts cross-domain shared features due to excessive parameter updates. Although increasing the number of frozen layers reduces trainable parameters and eases optimization, exceeding a critical threshold restricts model adaptability: domain differences must then be compensated solely through output layer adjustments, whose limited capacity fails to correct mismatches in deeper feature representations, resulting in accumulated prediction bias. By moderately freezing lower layers, the model protects pre-trained source knowledge from being overwritten by target-domain noise, while top-layer fine-tuning allows for effective feature recalibration via localized updates. Excessive fine-tuning, however, risks introducing target-specific noise and undermining the structural priors of the pre-trained model.

Table 6
Experimental results of the freeze and control groups.

	RMSE (t/day)	MAE (t/day)	MAPE (%)	R^2
Total Freeze Group	1.03	0.74	1.67	0.92
Deep Freeze Group	0.33	0.19	0.44	0.99
Medium Freeze Group	0.61	0.39	0.87	0.97
Shallow Freeze Group	0.64	0.41	0.92	0.97
Control Group	0.48	0.32	0.74	0.98

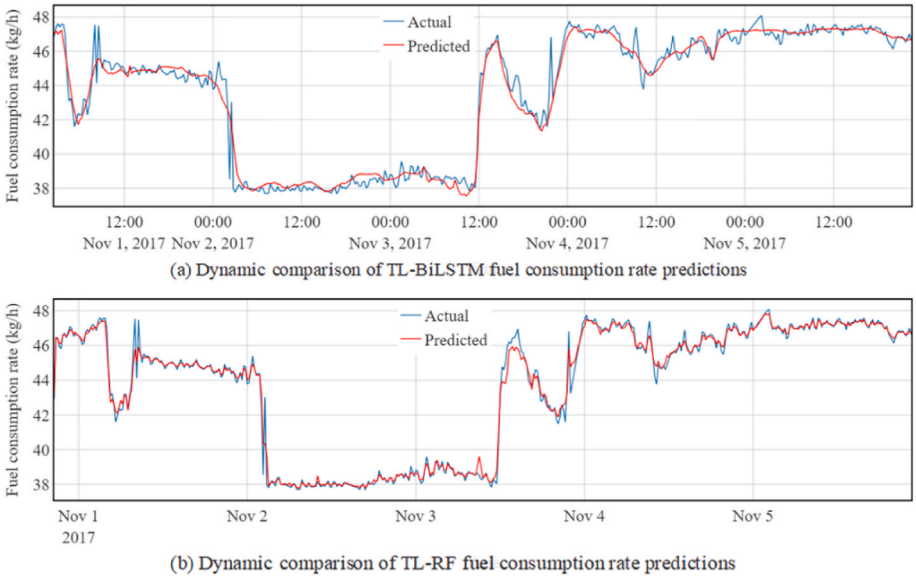

Fig. 13. Comparison of transfer learning model fuel consumption rate predictions.

This section investigates the dynamic equilibrium between feature retention and task adaptation in transfer learning by systematically varying the number of fine-tuned layers in the TL-BiLSTM model. The deep freeze strategy preserving low-level generic features while fine-tuning the final layer enables optimal cross-domain performance. Deviating from this configuration, either by freezing too many or too few layers, disrupts this balance and degrades model effectiveness. Therefore, careful selection of the number of layers subject to fine-tuning is crucial for architecture-aware transfer learning.

5.2. Impacts of source and target domain sample size

In order to deepen the research on the data interaction law between source and target domains in the transfer learning mechanism, and to systematically explore the quantitative impact of the sample sizes of the source and target domains on the generalization performance of the model, this study constructs a multi-group comparative experiment using the control variable method. Specifically, taking the base experiment containing 3005 samples in the source domain and 489 samples in the target domain in Section 4.3 as the baseline control group, and keeping the experimental conditions such as hyper-parameter configurations, network architecture, etc., strictly consistent, we further design four groups of experimental groups of graded sample size experiments (labelled as A, B, C, and D). By adjusting the sample size ratios of the source and target domains, as shown in Table 7, the experimental groups focus on the change law of the transferability of knowledge in the source domain under different sample sizes and the impact of the scarcity of samples in the target domain on the efficacy of transfer learning under the premise of keeping the characteristics of data distribution unchanged.

Due to the chance of the results of a single experiment, we will experiment A, B, C, D are independently repeated five times to take the average to eliminate the effect of random error, to ensure the scientific and interpretable results, the results of each run of the experiment as shown in Table 8, Table 9.

(1) Experiments with decreasing gradient of the ratio of the sample size in the source domain to the sample size in the target domain

The results of five independent experiments based on the TL-BiLSTM model and the TL-RF model show that, as shown in Figs. 14 and 15, under the premise of keeping the sample size of the target domain constant, when the sample size of the source domain decreases sequentially from the control group (n = 3005) to the experimental group A (n = 2000) and the experimental group B (n = 1000), the predictive performance of the model exhibits regular changes, and the RMSE, MAE, and MAPE all three error indicators have a gradual trend of small increase, while the R^2 coefficient of determination remains relatively stable with a fluctuation of <0.5 %. This phenomenon reveals that the prediction accuracy of the transfer learning model exhibits a gradient decrease characteristic as the ratio of the sample size of the source domain to the target domain decreases from 6:1 to 4:1 and 2:1. This may be due to the fact that the reduction of the source domain sample size directly weakens the learning base of migratable features and the scope of migratable learning knowledge, resulting in the model's difficulty in fully capturing domain-invariant features with universal applicability.

Table 7
Sample size of source and target domains for experimental and control groups.

	Source domain sample size	Sample size of the target domain	Proportions
Control Group	3005	489	6:1
Experiment A	2000	489	4:1
Experiment B	1000	489	2:1
Experiment C	3005	250	12:1
Experiment D	3005	100	30:1

Table 8
Results of five experiments of TL-BiLSTM strategy taking the mean value.

	RMSE (t/day)	MAE (t/day)	MAPE (%)	R^2
Control Group	0.48	0.32	0.74	0.98
Experiment A	0.56	0.37	0.79	0.97
Experiment B	0.56	0.38	0.82	0.97
Experiment C	0.55	0.34	0.84	0.97
Experiment D	0.48	0.30	0.68	0.98

Table 9
Results of five experiments of TL-RF strategy taking the mean value.

	RMSE (t/day)	MAE (t/day)	MAPE (%)	R^2
Control Group	0.33	0.18	0.43	0.99
Experiment A	0.37	0.21	0.48	0.99
Experiment B	0.40	0.24	0.55	0.98
Experiment C	0.33	0.15	0.38	0.99
Experiment D	0.28	0.12	0.35	0.99

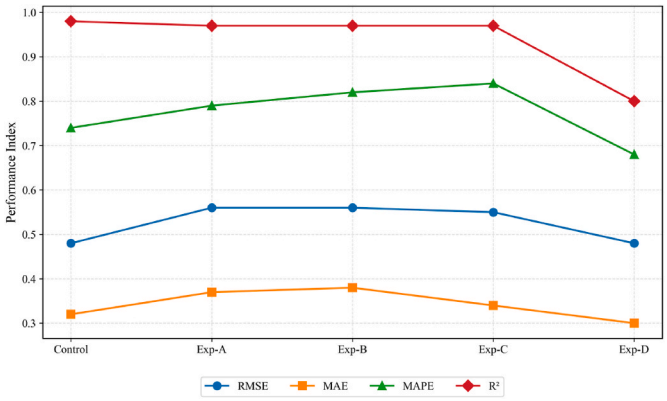

Fig. 14. Comparison of performance index of TL-BiLSTM.

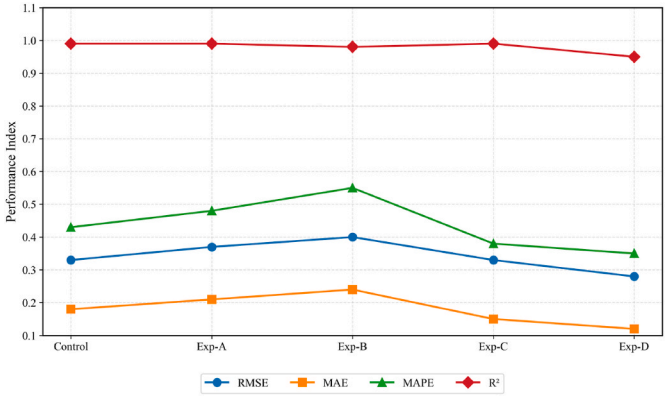

Fig. 15. Comparison of performance index of TL-RF.

(2) Experiments with gradient increase in the ratio of the sample size in the source domain to the sample size in the target domain

When the sample size of the target domain is sequentially reduced from the control group (n = 489) to experimental group C (n = 250) and experimental group D (n = 100), the prediction performance of the transfer learning model likewise exhibits a significant pattern of change. Except for the TL-BiLSTM architecture in experimental group C, which

exhibits an abnormal fluctuation contrary to the overall trend, the three error metrics of the rest of the experimental groups all show a systematic and small decreasing trend. This experimental phenomenon suggests that as the sample size of the target domain decreases, i.e., when the ratio of the sample size of the source domain to that of the target domain gradually increases from 6:1 to about 12:1 and 30:1, the prediction accuracy of the transfer learning model instead gains a steady improvement. This may be when the data in the target domain is scarce, the model effectively compensates for the deficiency caused by the lack of information in the target domain by strengthening the knowledge transfer from the source domain. Moreover, the sparser the data in the target domain is, the stronger the effect of the transfer compensation will be.

In this section, a quantifiable analysis of transfer learning data dependency is performed by controlling the sample proportion relationship between the source and target domains in ship FC prediction transfer learning, which provides an empirical basis for understanding the evolutionary pattern of model performance in data-limited transfer scenarios.

6. Conclusions

This study proposes a transfer learning framework for ship fuel consumption prediction in data-scarce scenarios, using operational data from a well-documented source ship to enhance prediction accuracy for a target ship with limited data. Two model architectures were developed: TL-BiLSTM and TL-RF. The models were evaluated against baseline and mixed-data models, and the effects of freezing strategies and source–target sample ratios were systematically analyzed. Key conclusions are as follows:

(1) Prediction performance: Both TL-BiLSTM and TL-RF achieved more than approximately 30 % reductions in RMSE, MAE, and MAPE compared to baselines, with R^2 values improved to 0.98 and 0.99, confirming the effectiveness of transfer learning in ship fuel consumption prediction.
(2) Freezing strategy in TL-BiLSTM, the deep-freeze strategy (fine-tuning the last three layers) achieved the best results, with RMSE, MAE, and MAPE of 0.33 t/day, 0.19 t/day, and 0.44 % respectively, and R^2 of 0.99, outperforming other freezing settings.
(3) Sample ratio effect: Increasing the source–target sample ratio enhanced prediction accuracy. In TL-RF, raising the ratio from 6:1 to 30:1 reduced RMSE from 0.33 t/day to 0.28 t/day and MAPE from 0.43 % to 0.35 %, showing the benefit of richer source-domain data in data-scarce scenarios.

These findings offer practical guidance to maritime operators and ship designers, enabling them to accurately predict the energy consumption of new ships or ships with limited sensors, thereby achieving cost-effective energy efficiency optimization, reducing the need for experimental testing, and supporting the decarbonization goals of the International Maritime Organization. Future research will incorporate dynamic environmental variables to eliminate temporal deviations caused by differences in weather and operational conditions, enhancing the model's adaptability to seasons and various time offsets. Meanwhile, the improvement in computational resources and costs of transfer learning compared to baseline models poses a significant challenge for its implementation in the real world. Further exploration of online learning mechanisms and lightweight deployment solutions is needed to meet the demands of real-time sensor data processing and promote the intelligent development of ship energy efficiency management.

CRediT authorship contribution statement

Ailong Fan: Writing – review & editing, Validation, Supervision, Resources, Project administration, Data curation. **Siyang Sun:** Writing – original draft, Visualization, Validation, Software, Methodology, Investigation, Formal analysis, Conceptualization. **Zhihui Hu:** Writing – review & editing, Validation, Supervision, Resources, Project administration, Data curation. **Nikola Vladimir:** Writing – review & editing, Validation. **Wengang Mao:** Writing – review & editing, Validation.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

This work is supported by the National Natural Science Foundation of China (52571365, 52502419), the National Key R&D Program of China (SQ2025YFE0102169), Hubei International Science and Technology Cooperation Project (2024EHA038), and Fund of State Key Laboratory of Maritime Technology and Safety (No. 14-9-1).

Appendix A. Optimal hyperparameter in different machine learning models

Table A.1
Optimal hyperparameter in the TL-BiLSTM and BiLSTM

Parameter	Searching space	Value
Seq length	[10, 50]	24
Epochs	[100, 300]	200
Batch Size	[16, 64]	32
Optimizer	[adam, sgd]	adam
Verbose	[0, 0.1]	0
Activation	[relu, tanh]	relu

Table A.2
Optimal hyperparameter in the TL-RF and RF

Parameter	Searching space	Value
N estimators	[100, 300]	158
Max depth	[10, 50]	50
Min samples leaf	[1, 4]	1
Max features	[auto, sqrt]	sqrt
Min samples split	[2, 10]	4

Table A.3
Optimal hyperparameter in the DT

Parameter	Searching space	Value
Min samples split	[2, 8]	8
Min samples leaf	[1, 4]	1
Max features	[auto, sqrt]	sqrt
Max depth	[10, 70]	50

Table A.4
Optimal hyperparameter in the KNN

Parameter	Searching space	Value
N neighbors	[3, 15]	4
P	[1, 2]	2
Weights	[uniform, distance]	distance

Table A.5
Optimal hyperparameter in the XGB

Parameter	Searching space	Value
N estimators	[100, 300]	100
Max depth	[3, 10]	6
Learning rate	[0.01, 0.3]	0.3
Subsample	[0.6, 1]	1

Table A.6
Optimal hyperparameter in the ANN

Parameter	Searching space	Value
Hidden layer sizes	[(100), (500, 50), (1000, 100), (1000, 500, 50)]	(1000, 100)
Activation	[relu, tanh]	relu
Alpha	[0.0001, 0.001, 0.01]	0.01
Learning rate	[0.001, 0.01, 0.1]	0.001
Solver	[adam, sgd]	sgd

References

Akande, O., Okolie, J.A., et al., 2025. A comprehensive review on deep learning applications in advancing biodiesel feedstock selection and production processes. Green Energy Intell. Transp. 4 (3), 100260.

Bellagarda, A., Cesari, S., et al., 2022. Effectiveness of Neural Networks and Transfer Learning for Indoor air-temperature Forecasting, 140, 104314.

Breiman, L.J. M.l., 2001. Random forests, 45 (1), 5–32.

Cheng, X., Li, G., et al., 2023. SAFENESS: a semi-supervised transfer learning approach for sea state estimation using ship motion data, 25 (5), 3352–3363.

Deng, Y.B., Li, Y.F., et al., 2024. Displacement values calculation method for ship multi-support shafting based on transfer learning. J. Mar. Sci. Eng. 12 (1).

Elissaios, S., Nikos, D., et al., 2022. Transfer learning strategies for solar power forecasting under data scarcity, 12 (1).

Fan, A.L., Li, Y.P., et al., 2023. Development trend and hotspot analysis of ship energy management. J. Clean. Prod. 389.

Fan, A.L., Wang, Y.Q., et al., 2024. Comprehensive evaluation of machine learning models for predicting ship energy consumption based on onboard sensor data. Ocean Coast Manag. 248.

Fan, A.L., Wang, Y.F., et al., 2025. Multi-dimensional performance verification of ship fuel consumption prediction model under dynamic operating conditions. Energy 332.

Fan, A.L., Xiong, Y.Q., et al., 2023. Carbon footprint model and low-carbon pathway of inland shipping based on micro-macro analysis. Energy 263.

Fan, A.L., Yang, J., et al., 2022a. Joint optimisation for improving ship energy efficiency considering speed and trim control. Transp. Res., Part D Transp. Environ. 113.

Fan, A.L., Yang, J., et al., 2022b. A review of ship fuel consumption models. Ocean Eng. 264.

Gautam, Y.J. I.t., 2022. Transfer Learning for COVID-19 cases and deaths forecast using LSTM network, 124, 41–56.

Gruetzemacher, R., Paradice, D.J.A.C.S., 2022. Deep transfer learning & beyond: transformer language models in. Inf. Syst. Res. 54 (10s), 1–35.

Han, X., Huang, Z., et al., 2021. Adaptive transfer learning on graph neural networks. Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining, pp. 565–574.
Han, P.X., Liu, Z.B., et al., 2024. A novel prediction model for ship fuel consumption considering shipping data privacy: an XGBoost-IGWO-LSTM-based personalized federated learning approach. Ocean Eng. 302.
He, Q.-Q., Pang, P.C.-I., et al., 2019. Transfer learning for financial time series forecasting. Pacific Rim International Conference on Artificial Intelligence. Springer, pp. 24–36.
Hochreiter, S., Schmidhuber, J.J. N.c., 1997. Long short-term memory, 9 (8), 1735–1780.
Karb, T., Kühl, N., et al., 2020. A Network-based Transfer Learning Approach to Improve Sales Forecasting of New Products.
Lang, X., Wu, D., et al., 2022. Comparison of supervised machine learning methods to predict ship propulsion power at sea. Ocean Eng. 245.
Lang, X., Wu, D., et al., 2024. Physics-Informed Machine Learning Models for Ship Speed Prediction, 238, 121877.
Li, J., Lin, P., et al., 2024. Time prediction in ship block manufacturing based on transfer learning, 12 (11), 1977.
Luo, X., Zhang, M.Y., et al., 2025. Ship fuel consumption prediction based on transfer learning: models and applications. Eng. Appl. Artif. Intell. 141.
Mao, W.G., Rychlik, I., et al., 2016. Statistical models for the speed prediction of a container ship. Ocean Eng. 126, 152–162.
Mavroudis, S., Tinga, T., 2025. Application of transfer learning on physics-based models to enhance vessel shaft power predictions. Ocean Eng. 323.
Milicevic, M., Zubrinic, K., et al., 2018. Data augmentation and transfer learning for limited dataset ship classification, 13 (1), 460–465.
Mohd Razak, S., Cornelio, J., et al., 2022. Transfer learning with recurrent neural networks for long-term production forecasting in unconventional reservoirs, 27 (4), 2425–2442.
Pan, S.J., Yang, Q., et al., 2009. A survey on transfer learning, 22 (10), 1345–1359.
Qiao, D., Liu, G., et al., 2020. Ship target recognition based on transfer learning, 37 (1), 324–325.
Rahman, M.A., Jamal, S., et al., 2024. Remote condition monitoring of rail tracks using distributed acoustic sensing (DAS): a deep CNN-LSTM-SW based model. Green Energy Intell. Transp. 3 (5), 100178.
Ribeiro, M., Grolinger, K., et al., 2018. Transfer learning with seasonal and trend adjustment for cross-building. energy forecasting 165, 352–363.
Shu, Y.Q., Yu, B.S., et al., 2024. Investigation of ship energy consumption based on neural network. Ocean Coast Manag. 254.
Sulaiman, M.H., Mustaffa, Z., 2024. State of charge estimation for electric vehicles using random forest. Green Energy Intell. Transp. 3 (5), 100177.
Sun, B., Saenko, K., 2016. Deep coral: correlation alignment for deep domain adaptation. European Conference on Computer Vision. Springer, pp. 443–450.
Tian, C.L., Liu, Y.C., et al., 2024. Transfer learning based hybrid model for power demand prediction of large-scale electric vehicles. Energy 300.
Wang, K., Hua, Y., et al., 2023. A novel GA-LSTM-based prediction method of ship energy usage based on the characteristics analysis of operational data. Energy 282.
Wilbur, M., Mukhopadhyay, A., et al., 2021. Energy and emission prediction for mixed-vehicle transit fleets using multi-task and inductive transfer learning, 21st joint European conference on machine learning and principles and practice of knowledge discovery in databases (ECML PKDD). Electron. Netw. 502–517.
Wu, Z., Wang, S., et al., 2024. Research on Ship Safety Risk Early Warning Model Integrating Transfer Learning and multi-modal Learning, 150, 104139.
Xi, H.W., Ma, W., et al., 2025. Energy consumption prediction and endurance optimization for underwater gliders based on data-model fusion. Eng. Appl. Artif. Intell. 162.
Xie, D., Jiang, Y., et al., 2025. Full-scene energy consumption prediction for electric vehicles: a knowledge-enhanced hybrid-driven framework. Energy 333.
Yang, K., Yang, T.W., et al., 2021. A transfer learning-based convolutional neural network and its novel application in ship spare-parts classification. Ocean Coast Manag. 215.