https://doi.org/10.1016/j.trc.2022.103924

Efficient and explainable ship selection planning in port state control

*Ran Yan[1], Shining Wu[1], Yong Jin[2], Jiannong Cao[3], Shuaian Wang[1]**
[1]Departmnet of Logistics and Maritime Studies, The Hong Kong Polytechnic University, Hong Kong
[2]School of Accounting and Finance, The Hong Kong Polytechnic University, Hong Kong
[3]Department of Computing, The Hong Kong Polytechnic University, Hong Kong

Abstract

Port state control is the safeguard of maritime transport achieved by inspecting foreign visiting ships and supervising them to rectify the non-compliances detected. One key issue faced by port authorities is optimal allocation of inspection resources with accurate identification of high-risk ships as the premise. This study aims to address the ship selection issue by first developing two data-driven ship risk prediction frameworks using features the same as or derived from the current ship selection scheme. The proposed frameworks can identify up to 60% more deficiencies and nearly 40% more detentions on average compared to the current ship selection method. Like existing ship risk prediction models, the proposed frameworks are of black-box nature whose working mechanism is opaque. To improve model explainability, local explanation of the prediction of individual ships by the Shapley additive explanations (SHAP) with the properties of local accuracy and consistency is provided. Furthermore, we innovatively extend the local SHAP model to a fully-explainable near linear-form global surrogate model of the original black-box data-driven model by deriving feature coefficients and fitting curves of feature values and SHAP values from the SHAP value matrix. This demonstrates that the behavior of black-box data-driven models can be as interpretable as white-box models while retaining their prediction accuracy. Numerical experiments demonstrate that the white-box global surrogate models can accurately show the behavior of the original black-box models, shedding light on model validation, fairness verification, and prediction explanation, and hence promote their acceptance and application among maritime stakeholders. This study makes the very first attempt in the maritime transport area to quantitatively explain the rationale of black-box prediction models from both local and global perspectives, which facilitates the application of data-driven models and promotes the digital transformation of the traditional shipping industry.

Keywords

Port state control (PSC), marine policy, black-box model explanation, Shapley additive explanations (SHAP), linear-form global surrogate model

* Corresponding author. Email addresses: angel-ran.yan@connect.polyu.hk (R Yan), sn.wu@polyu.edu.hk (S Wu), jimmy.jin@polyu.edu.hk (Y Jin), csjcao@comp.polyu.edu.hk (J Cao), wangshuaian@gmail.com (S Wang)

1. Introduction

Maritime transport underpins global supply chain linkages and economic interdependency with shipping and ports estimated to handle over 80% of global merchandise trade by volume and more than 70% by value (Yi et al., 2019; Qi et al., 2021; UNCTAD, 2021; Hellsten, 2021; Trivella et al., 2021). To enhance maritime safety, protect the marine environment, and improve the living and working conditions of seafarers, international and regional maritime conventions and requirements are implemented by the International Maritime Organization (IMO), the International Labour Organization, and the local governments (Karsten et al., 2017; Asadabadi and Miller-Hooks, 2020; Tseng and Ng, 2020; Mokhtari et al., 2021; Wang et al., 2021c; Zhou et al., 2021; Zisi et al., 2021). Along with the joint efforts made by ship owners and operators, classification societies, and flag states, the port state control (PSC) is an effective 'safety net' to catch substandard ships among foreign visiting ships whose conditions are not compliant with the relevant conventions and requirements. After the Paris Memorandum of Understanding (MoU) on PSC was signed in 1982, there are nine current regional MoUs over the world, namely Abuja MoU, Vina del Mar MoU, Black Sea MoU, Caribbean MoU, Indian Ocean MoU, Mediterranean MoU, Paris MoU, Tokyo MoU, and Riyadh MoU. On top of that, the United States Coast Guard (USCG) maintains the tenth inspection regime.

Given a large number of foreign visiting ships and the limited inspection resources at a port, only a small proportion of the ships can be inspected by PSC. For example, only 13.05% of all the foreign ships visiting the Hong Kong Port were inspected in 2019 (Tokyo MoU, 2020). Meanwhile, globally, less than half of the inspected ships were with deficiency detected during 2018 and 2020, while only 2.50% of them were detained (i.e., with very serious deficiency or deficiencies detected) during this period[*] (Marine Department, 2021). This indicates that accurate identification of substandard ships and rational allocation of the scarce inspection resources at port is the key to improve the effectiveness of PSC while reducing the delay of the fast turnover of the maritime logistics systems brought about by non-essential inspections. Moreover, the cost of a PSC inspection can be very high: according to Tokyo MoU, the charge of the first hour of follow-up inspection at the Hong Kong Port is 3,270 HKD (about 420 USD)

[*] The data are derived from the available statistics of different PSC regimes, including Paris MoU, Tokyo MoU, Black Sea MoU, Indian Ocean MoU, Mediterranean MOU, Vina Del Mar, and USCG, by the Marine Department of the government of the Hong Kong Special Administrative Region.

and that of the subsequent hours is 1,115 HKD (about 143 USD) per hour, and the documentation fee is 1,115 HKD (about 143 USD) per hour (Tokyo MoU, 2016). Therefore, correct identification and inspection of high-risk ships can not only improve inspection efficiency, but also save resources and reduce costs.

This study aims to develop and explain state-of-the-art data-driven machine learning (ML) based ship risk prediction models to assist port states in identifying and selecting high-risk foreign visiting ships (Li et al., 2021; Yi et al., 2021; Zhu et al., 2021). We use six years' PSC inspection records at the Hong Kong Port to develop two ship risk prediction frameworks based on gradient boost regression trees (GBRTs) to predict ship deficiency number. Features in the proposed frameworks are either the same as those considered in the current ship selection scheme applied by the Tokyo MoU or the same types of features but with different encoding methods for preprocessing. Post-hoc, model-agnostic, and local explanations are then given by Shapley additive explanations (SHAP) method, aiming to explain the prediction of individual ships. We further extend the local SHAP method to a global explanation method taking a near linear form by calculating the average SHAP values of different states for categorical features and fitting curves of feature values and SHAP values for integer and continuous features. Thorough analysis of model explanations is given to draw policy insights and managerial recommendations for both port authorities as well as ship owners and managers. To be more specific, this study makes the following contributions.

From theoretical perspective, we extend the local SHAP method to a global explanation method in an intuitive and succinct way, showing that the prediction behavior of black-box models (e.g., the GBRT models to predict ship deficiency number in this study) can be presented by white-box models (e.g., the extended SHAP model taking a near linear form) without compromising their prediction performance under arbitrary problem setting. The near linear-form global surrogate model is derived directly from local explanations, and thus can illustrate the average contribution of each feature value to the final prediction on the whole dataset. Such unification of local and global explanations can make the interpretation of black-box model more comprehensive and consistent. Furthermore, we demonstrate that the black-box ship risk prediction models are of satisfactory accuracy, and the difference between the predictions given by the original black-box models and that given by the near linear-

form global surrogate models is minor. In addition, model explanations given by local and global feature importance scores, beeswarm plots, and near linear-form global surrogate models are comprehensible to port authorities and ship owners, operators, and management companies. The explanations are also essential for them to trust and apply the proposed frameworks. Therefore, the explanations can be validated to follow the 'predictive, descriptive, and relevant' framework for black-box model explanation evaluation (Murdoch et al., 2019).

From practical perspective, to the best of the authors' knowledge, this is the very first study that explores explanations of black-box prediction models in maritime transport and thus paves the way of adopting ML models (which is a typical type of black-box model) to address maritime transport problems. The motivation and rationale of doing so are that the maritime industry decision makers are relatively conservative, and many current approaches are based on expert knowledge ~~that~~ and are fully explainable (i.e. in a while-box manner). Therefore, opening up a black-box model of higher efficiency to be in a white-box manner is necessary and significant to promote ~~their~~ its application in maritime transport research. Especially, a critical problem in a major international shipping policy is addressed in this study, i.e., high-risk ship selection in PSC. Only the factors considered in the current ship selection scheme are used for developing ship selection frameworks, making them more applicable to port authorities. Thorough explanations of the black-box prediction models further make them more comprehensible and acceptable by port authorities as well as ship owners and managers. Numerical experiments show that the proposed ship selection frameworks are more efficient in identifying high-risk ships and can help to identify more than 60% more deficiencies and more than 9 out of the total 23 detentions in a hold-out test set compared to the current ship selection scheme.

From policy making point of view, the comprehensive and consistent explanations provided in this study make an initial step to bridge the gap between making a prediction and making a decision in maritime transport area from at least three perspectives: trustworthiness, fairness, and informativeness. Disclosing the inner working mechanism and decision process of a black-box prediction model can help to verify whether the predictions given by the black-box model comply with domain knowledge. If yes, the proposed black-box prediction model can be expected to be more trustable and acceptable by decision makers. Fairness of the recommendations made by

black-box prediction models is a main concern of policy makers, which can also be validated by investigating the coefficients and curves of the features in the global surrogate models developed in this study. Insights extracted from practical data can shed light on policy and decision makings in the future, and thus enhance the informativeness of model explanation.

The remainder of this paper is structured as follows. Section 2 gives a comprehensive review of current literature on ship risk prediction in PSC and explanation of black-box models in the transportation area. Section 3 analyzes the current ship selection scheme at port. Section 4 develops and validates ML based frameworks for ship risk prediction. Section 5 discusses main points of black-box model explanation, especially its importance in maritime transport. Section 6 develops local and global explanations for the proposed black-box ship risk prediction frameworks. Section 7 concludes this research. The overall workflow of the frameworks for high-risk ship selection and explanation algorithm construction is given in Fig. 1.

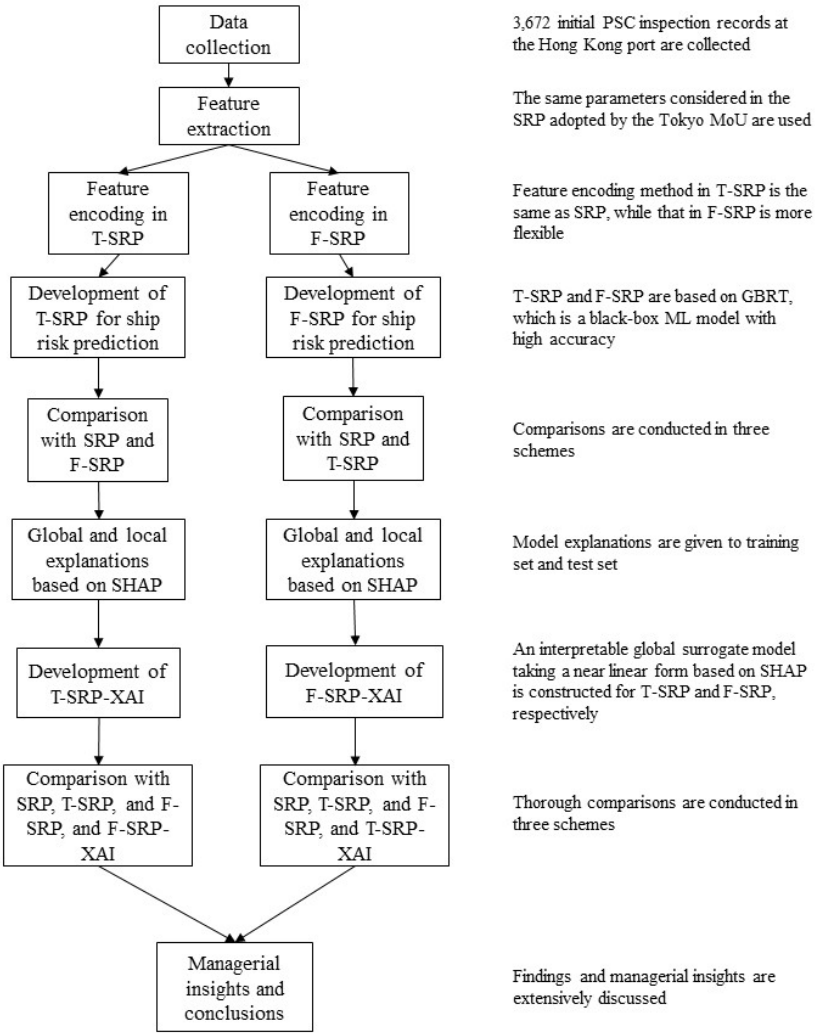


Fig. 1. Overall workflow of the framework~~s~~ for the high-risk ship selection and explanation algorithm construction

2. Literature review

2.1 Ship risk prediction for PSC

In recent 15 years, there has been an increasing number of studies on PSC inspection, which can be roughly classified into four categories: influencing factors of inspection results, high-risk ship selection methods, effects of PSC inspection, and suggestions for MoU management (Yan and Wang, 2019). As this study falls in the second research category, related studies are reviewed in this subsection. Basically, there are two forms of ship risk within the context of PSC in current literature: direct ship risks that can be observed in the current PSC inspection, such as deficiency and detention, and indirect ship risks that cannot be observed in the current inspection, such as future accidents and incident risks. The accuracy of direct ship risk prediction is easy to be verified by leveraging PSC inspection records, while that of indirect ship risk

prediction might be difficult to verify as it can hardly be quantified or observed in the current inspection or near future.

Regarding the literature aiming to predict direct ship risks, Xu et al. (2007a) is the pioneer to use ML model for ship detention prediction by developing a support vector machine (SVM) model using generic and history factors. Later on, Xu et al. (2007b) designed a profile-based web wrapper to extract and incorporate more features into the SVM model. Gao et al. (2007) further combined SVM with k-nearest neighbor (KNN-SVM) to remove noisy training examples. More recently, an SVM model was proposed by Wu et al. (2021) for ship detention prediction. Particularly, input features were selected by analytic hierarchy process (AHP) and grey relational analysis (GRA) to improve prediction accuracy.

Bayesian network (BN) is another popular ship risk prediction method. Yang et al. (2018a) proposed a data-driven BN model based on Tree Augmented Naive Bayes (TAN) learning to predict ship detention probability. The outcomes of a TAN model for ship detention prediction were input to a game model between port authorities and ship owners to determine vessel detention rates (Yang et al., 2018b). Wang et al. (2019) developed a TAN classifier for ship deficiency number prediction to identify high-risk ships. The main reason for the wide application of BN models to PSC related research is its graphic representation: the importance degree of risk factors and their mutual relationship can be visualized, making them partially explainable.

In addition to SVMs and BNs, other types of ML models were also developed for direct ship risk prediction. To address the problems brought by the highly imbalanced dataset due to low detention rate (about 3.55%), Yan et al. (2021b) developed a balanced random forest (BRF) model for ship detention prediction. Ship risk prediction results were also used to address scarce inspection resource allocation problems in current literature. For example, Yan et al. (2020) proposed three random forest models consisting of multi-target regression trees to predict ship deficiency number under each deficiency category. Model outputs were then input to the following PSC officer assignment models to match inspectors' expertise with the predicted ship deficiency condition. Yan et al. (2021a) integrated shipping domain knowledge in an XGBoost model for ship deficiency number prediction, and the output was input to the following PSC officer scheduling models.

Researchers have also attempted to evaluate ship indirect risks in existing studies. Degré (2007) proposed the risk concept to characterize high-risk ships for PSC inspection considering casualty occurrence probability and the potential consequences. Degré (2008) further extended the black-grey-white flag list published by the Paris MoU to black-grey-white ship category list based on ship characteristics and vessel casualty data. Heij and Knapp (2019) proposed a decision support tool based on three logit models for high-risk ship identification considering two risk dimensions: past incidents and detention information. As an extension, Knapp and Heij (2020) proposed five combined models based on logit models and percentile ranks for ship risk prediction. Dinis et al. (2020) input parameters from the ship risk profile (SRP) of Paris MoU as the risk variables to a BN model for static risk profile assessment of individual ships. The dataset used, features considered, risk indicators and prediction model developed, and model explainability of all the reviewed studies are provided in Appendix A.

The above analysis indicates that one of the largest gaps in current literature is the lack of model explainability. On the one hand, except for BN, which is partially explainable, all the other models for direct ship risk prediction are in a total black-box nature. It is also noted that although Naive Bayes, which is the most basic type of BN model, can be viewed as a type of interpretable model (Molnar, 2020), its interpretability is due to the underlying independence assumption, and thus the contribution of each feature towards the prediction target is clearly presented by the conditional probability tables. However, as Naive Bayes models usually oversimplify the reality, their accuracy is highly compromised. Therefore, none of the BN models developed in the abovementioned research is Naive Bayes model. Instead, BNs with more complex structures, especially those taking interdependencies among the variables into account, were developed for ship risk prediction. Consequently, interpretability of these BNs is largely weakened, especially those containing intermediate variables such as Yang et al. (2018a, 2018b) and Dinis et al. (2020). On the other hand, although the statistical models employed for indirect ship risk prediction are interpretable to a certain degree, their predictive power could be weaker than that of the state-of-the-art ML models (Murdoch et al., 2019).

Another gap in existing literature is the features used for ship risk prediction. Table A.1 indicates that except for Dinis et al. (2020) where only factors in the SRP are

considered to predict ship risk, external databases with different degrees of difficulty in obtaining are used by all the other studies. Consequently, these models might be hard to be adopted by the conservative port authorities as such external datasets may not be trusted by them and much more time, efforts, and money might be spent on obtaining and processing the required data. Meanwhile, although only the factors in the SRP were considered to develop ship risk prediction models in Dinis et al. (2020), the prediction target of more detailed ship risk profile (a total of 14 risk levels) is abstract and might hard to be verified.

2.2 Explainable artificial intelligence in transportation research

To make the literature review more comprehensive, we also briefly review the existing literature on exploring explainable artificial intelligence (XAI) in transportation research. ML and deep learning approaches have been adopted by a large number of works in the transportation field, but only quite a few have addressed model explainability issue (Kalatian and Farooq, 2021), and most of the related studies are published in recent three years. These explanation methods can be divided into two types: global explanation, which aims to explain the entire model behavior, and local explanation, which aims to explain an individual prediction. Features' relative importance to the prediction target is the most common way of global explanation, which can be found in Zhang and Haghani (2015) for highway travel time prediction, Hagenauer and Helbich (2017) for travel mode choice prediction, and Chen et al. (2017) for passengers' ridesplitting behavior prediction. In addition, Wang et al. (2020) developed a decision tree as a surrogate of the black-box prediction model for congestion attack prediction. Meanwhile, local explanation is achieved by SHAP in Barredo-Arrieta et al. (2019) for traffic flow prediction, Veran et al. (2020) for crash prediction, and Kalatian and Farooq (2021) for pedestrians' wait time prediction. Other common methods for local explanation, such as partial dependence plot (PDP), individual conditional expectation (ICE), and accumulated local effect (ALE) were used in Khoda Bakhshi and Ahmed (2021) for road crash probability prediction.

Both global and local explanations are provided by some studies via separate approaches. Especially, global explanation was also mainly achieved by deriving feature importance or feature interactions, while local explanation was reached by PDP in Zhao et al. (2018) for travel mode switching behavior prediction, by SHAP in Parmar et al. (2021) for parking duration prediction, by ALE in Kim et al. (2020) for passenger

transit purpose prediction, and in Kim (2021) for travel mode choice prediction, by local interpretable model-agnostic explanations (LIME) in Bukhsh et al. (2019) for rail maintenance need prediction and management, and by PDP and ALE in Xu et al. (2021) for ridesplitting adoption prediction.

The studies covered in this subsection so far mainly adopt existing explanation methods. Besides, researchers have also proposed innovative explanation methods in specific problem settings. Zhao et al. (2019) extended the PDP method to conditional PDP and conditional individual PDP for travel mode switching behavior analysis. The key idea was to group instances into subpopulations first based on some features, and then conduct analysis in each subpopulation. Kim et al. (2020) developed a two-stage framework consisting of a linear regression (LR) part for model interpretability and a long short-term memory (LSTM) part for model accuracy to predict taxi demand. Wang et al. (2020, 2021a, 2021b) tried to explain and extend deep neural networks (DNNs) to analyze travel mode choice. Particularly, Wang et al. (2020) demonstrated that DNNs could provide economic interpretation as complete as classical discrete choice methods (DCMs). Wang et al. (2021a) further substantiated the interpretability of DNN by formulating the function approximation loss to measure interpretation quality. Considering the shared utility interpretation of DCMs and DNNs, Wang et al. (2021b) synergized both models to a unified framework to achieve mutual benefits for travel behavior modeling.

Although some pioneering efforts have been made to disclose the black-box prediction models applied in transportation research, there are still some limitations. First, although some studies aim to provide more comprehensive model explanations by giving both global and local explanations, the two types of explanations are derived from separate methods, and thus inconsistency in explanation might occur. Second, although there are some extensions of current explanation methods, such extensions are problem-specific, making them hard to be applied to other problems. Third, none of these studies are within the context of maritime transport area, where interpretation and explanation of black-box models are urgently needed to facilitate their successful application in the traditional and conservative shipping industries.

To bridge these gaps, two highly accurate ML based ship risk prediction frameworks only using factors from the current ship selection scheme are developed in this study. Local explanation and analysis are then given by SHAP. To go one step

294 further, we extend SHAP to a global method in a near linear form by formulating a
295 global surrogate model of the original ML model, and such extension can be applied to
296 arbitrary problem other than the ship selection problem in maritime transport.
297 Contribution of each feature value to the final prediction can be derived from the
298 parameters in the surrogate model similar to a near linear regression model, and thus
299 the black-box prediction model can be considered as explainable as white-box models
300 while its high accuracy can be fully retained.
301
302 **3. Current ship selection method at port**
303 Unified ship selection method is required to be adopted within one MoU. Currently,
304 two types of ship selection schemes are widely adopted. One is assigning higher
305 inspection priority to ships in certain conditions (e.g., of certain types, with dangerous
306 goods carried, and with unsatisfactory performance in recent inspections) (applied in
307 Vina del Mar, Mediterranean MoU, and Riyadh MoU) or use risk target matrix (applied
308 in Caribbean MoU) based on long-term experience. The other is the New Inspection
309 Regime (NIR) based on the SRP to determine the scope, frequency, and priority of ship
310 inspection (applied in Abuja MoU, Black Sea MoU, Indian MoU, Paris MoU, and
311 Tokyo MoU). Generally, the NIR considers a ship's age, type, flag/recognized
312 organization (RO)/company performances, and recent deficiency and detention
313 conditions in a weighted sum manner to calculate ship risk. It is also noted that the NIR
314 adopted by different MoUs might be slightly different regarding the factors considered
315 and their weighting points, but the basic procedure is similar.
316 This study uses PSC inspection records at the Hong Kong Port within the Asia-
317 Pacific Region under the Tokyo MoU, and thus we mainly introduce the SRP adopted
318 by the Tokyo MoU in detail. The Tokyo MoU starts to adopt the SRP since 1 January
319 2014, which classifies foreign visiting ships into three risk profiles: high risk ship
320 (HRS), standard risk ship (SRS), and low risk ship (LRS). The information sheet to
321 determine ship risk is shown in Table 1 (Tokyo MoU, 2013).

322 Table 1. Information sheet of SRP adopted by the Tokyo MoU

Parameters	Values	Weighting points	Criteria for LRS
Ship type	Chemical tanker, gas carrier, oil tanker, bulk carrier, passenger ship, container ship	2	\

Ship age (calculated based on the keel laid date)	All types with age > 12y	1	\
Flag performance in Black-Grey-White list of Tokyo MoU	Black	1	White, and should be IMO Audit
RO performance evaluated by Tokyo MoU	Low/very low	1	High, and should be an RO recognized by the Tokyo MoU
Company performance evaluated by Tokyo MoU	Low/very low/no inspection within previous 36 months [unknown]	2	High
Deficiencies within previous 36 months	Inspections which recorded over 5 deficiencies	The number of inspections which recorded over 5 deficiencies	All inspections have 5 or less deficiencies and has at least one inspection within previous 36 months
Detentions within previous 36 months	3 or more detentions	1	No detention

Ship risk profile	**Criteria**	**Inspection time window**
HRS	When the sum of weighting points≥4	2 to 4 months
SRS	Neither HRS nor LRS	5 to 8 months
LRS	All the criteria for LRS are met	9 to 18 months

Particularly, the flag Black-Grey-White list and the RO performance list are published by the Tokyo MoU in the annual report considering the inspection and detention history of the vessels under the corresponding flag and RO over the preceding three calendar years. Ship company performance is the performance of a ship's international safety management (ISM) company which is calculated daily on the basis of a running 36-month period considering the detention and deficiency history of the company's fleet.

For a foreign visiting ship attached with a specific risk profile, its inspection priority is determined by the relationship between its last inspection time and the inspection time window attached to its SRP. Especially, there are two levels of inspection priority: ships with the last inspection time beyond the upper bound of the inspection time window are of Priority I and must be inspected; ships with the last inspection time within the inspection time window are of Priority II and may be inspected. Meanwhile, ships with the last inspection time less than the lower bound of the inspection time window have no priority (Tokyo MoU, 2013).

The above description shows that the SRP adopted by the Tokyo MoU considers critical ship risk factors ranging from ship properties to historical inspection performance in an intuitive way for ship risk calculation. Actually, the NIR (and the

SRP) is so far viewed as the most significant change that transforms and modernizes the PSC inspection mechanism in recent year (Yang et al., 2021). It is regarded as an advanced scheme to identify higher risk ships for PSC inspection and has empirically been shown to be able to enhance the effectiveness of PSC (Yang et al., 2018a, 2018b; Dinis et al., 2020; Xiao et al., 2020; Yang et al., 2021). Nevertheless, the main limitation of the SRP is that only an over-simplified weighted sum model with criteria and weights determined by subjective experience is used to calculate ship risk, which largely simplifies the practical situation. In addition, it totally ignores the correlations among parameters, and hence weakens its ability to distinguish ships under extremely poor conditions. Furthermore, only rough inspection priority largely depending on domain knowledge is given, while no specific risk level of each ship within the same inspection priority is provided. This further weakens its effectiveness as an indicator of ship risk to guide ship selection in PSC.

4. Development of ML based ship risk prediction frameworks for PSC

As the domain knowledge based on SRP applied by the Tokyo MoU has several drawbacks which reduce its effectiveness in high-risk ship identification, data-driven ship risk prediction frameworks based on ML models are developed in this study to achieve efficient ship selection. Data sources and features used for model calibration are first overviewed, the data-driven frameworks for ship risk prediction are then introduced, and finally the prediction performance is comprehensively compared and analyzed.

4.1 Data

A total of 3,672 initial PSC inspection records at the Hong Kong Port from 1 January 2015 to 31 December 2020 constitute the case dataset of this study. The inspection records are searched from the public PSC database of Tokyo MoU[†]. The whole dataset is randomly split into training set (80%, 2,937 samples) and test set (20%, 735 samples). To make the ship risk prediction frameworks developed more consistent with the current ship selection scheme at the Hong Kong Port, i.e., the SRP, and to avoid imposing extra burden of data acquisition and model understanding on the model users, we adopt the same parameters considered in the SRP within the Tokyo MoU as the features to develop the ML models. Particularly, we develop two ship selection

[†] http://www.tokyo-mou.org/inspections_detentions/psc_database.php

frameworks with different feature processing methods. One framework uses the same criteria as the SRP to decode each parameter considered (denoted by T-SRP), and the other uses more flexible decoding method to process the parameters (denoted by F-SRP). Detailed parameter decoding methods in the two frameworks are presented in Table 2.

Table 2. Feature processing methods in T-SRP and F-SRP

Type	Parameters in SRP	Criteria in SRP	SRP	Ship selection frameworks based on ML models			
				T-SRP		F-SRP	
			Weighting points	Feature decoding method	Feature type after decoding	Feature decoding method	Feature type after decoding
1	Ship type	Chemical tanker, gas carrier, oil tanker, bulk carrier, passenger ship, container ship	2	If ship type within the criteria of SRP: 1_ship_type_concerned = 1; else: 1_ship_type_concerned = 0	Binary	If ship type within the criteria of SRP, set the corresponding binary variable to 1: ship_type_Bulk carrier ship_type_Chemical tanker ship_type_Container ship ship_type_Gas carrier ship_type_Oil tanker ship_type_Passenger ship else: set ship_type_other = 1	Binary
2	Ship age	All types with age > 12 years	1	If ship age more than 12: 2_ship_age_12+ = 1; else: 2_ship_age_12+ = 0	Binary	Ship age value calculated based on keel laid date and inspection date	Continuous
3	Flag performance in Black-Grey-White list of Tokyo MoU	Black	1	If ship flag performance black: 3_flag_black = 1; else: 3_flag_black = 0	Binary	If ship flag performance available, set the corresponding binary variable to 1: flag_White flag_Grey flag_Black else: set flag_Undefined = 1	Binary*
4	RO performance in Tokyo MoU	Low/very low	1	If ship RO performance low or very low: 4_RO_low = 1; else: 4_RO_low = 0	Binary	If ship RO performance available, set the corresponding binary variable to 1: RO_High RO_Medium RO_Low** RO_Very Low** else: set RO_Undefined = 1	Binary*
5	Company performance in Tokyo MoU	Low/very low/no inspection within previous 36 months [unknown]	2	If ship company performance low, very low, or unknown: 5_company_low = 1; else: 5_company_low = 0	Binary	If ship company performance is available, set the corresponding binary variable to 1: company_High company_Medium company_Low company_Very Low else: set company _Unknown = 1	Binary*
6	Number of deficiencies recorded in each inspection within previous 36 months	How many inspections were there which recorded over 5 deficiencies?	Number of inspections which recorded over 5 deficiencies	The number of inspections with over 5 deficiencies in previous 36 months: 6_deficiency_no_last_36	Integer	The average deficiency number per inspection within previous 36 months: avg_deficiency_last_36_months	Continuous
7	Number of detentions within previous 36 months	3 or more detentions	1	If involved in 3 or more detentions within previous 36 months: 7_deficiency_last_36 = 1; else: 7_deficiency_last_36 = 0	Binary	The detention rate within previous 36 months: detention_rate_last_36_months	Continuous
8	If there is any inspection within previous 36 months	\	\	\	\	If there is any inspection within previous 36 months: whether_inspection_last_36_month = 1; else: whether_inspection_last_36_month = 0	Binary***

Note*: Except for the state 'unknown', all the other states are sequential. For example, ship flag performance gets worse from white, grey, to black. To capture different influence degrees of the states on ship risk in the downstream analysis, we use one-hot encoding to process their values.

Note**: There is no such record in the dataset in this study.

Note***: This variable is only in the F-SRP model as a complimentary to variables in types 6 and 7 to distinguish between no inspection and zero average deficiency or detention rate within previous 36 months.

4.2 Introduction of GBRT

Boosting is one of the most powerful learning methods in the ML community (Friedman et al., 2001). The basic idea of boosting is to develop a procedure that combines the outputs of many less accurate but diverse weak learners in an additive manner to produce a power ensembled model (Friedman et al., 2001). Flexible regression and classification trees (CARTs) are popular weak learners in boosting models. In GBRT for regression tasks, one CART is fit on the negative gradient value of the given loss function in each iteration. Denote a dataset with n samples and m features by $D=\{(\mathbf{x}_i, y_i), i=1,...,n\}, \mathbf{x}_i \in R^m, y_i \in R$, and the prediction of sample $(\mathbf{x}_i, y_i)$ by $f(\mathbf{x}_i)$. If the squared loss in Eq. (1) is used as the loss function, least squares is applied,

$$L(y_i, f(\mathbf{x}_i)) = \frac{1}{2}[y_i - f(\mathbf{x}_i)]^2, \tag{1}$$

and the negative gradient value of the loss function for sample i is the ordinary residual represented by

$$-g_i = -\frac{\partial L(y_i, f(\mathbf{x}_i))}{\partial f(\mathbf{x}_i)} = y_i - f(\mathbf{x}_i). \tag{2}$$

The main hyperparameters of a GBRT model are listed in Table 3.

Table 3. Main hyperparameters of GBRT

Hyperparameter	Meaning	Value space
n_estimators (K)	The number of iterations (weak learners) constituting a GBRT model	integer, $[1,+\infty)$
learning_rate (ε)	This hyperparameter aiming to shrink the contribution of each tree to the whole ensemble model to reduce overfit	decimal, $(0,1]$
max_depth	The maximum depth of each regression tree	integer, $[1,+\infty)$
min_samples_leaf	The minimum number of training samples required to be at a leaf node	integer, [1, the number of samples]
sub_sample	The fraction of training samples to be randomly selected to construct each regression tree	decimal, $(0,1]$
sub_feature	The fraction of features to be randomly selected to construct each regression tree	decimal, $(0,1]$

The detailed procedure to construct a GBRT model is presented in Procedure 1 (Friedman et al., 2001).

Procedure 1. Construction of a GBRT model	
Input	Training set D; the number of iterations/regression trees K; the loss function L; *max_depth* and *min_samples_leaf* as the stopping criteria of one tree; learning rate ε; *sub_sample*; *sub_feature*
Output	A GBRT model denoted by $f(\mathbf{x})$
Step 1	Initialize $f_0(\mathbf{x}) = \arg\min_{c_0} \sum_{i=1}^{n} L(y_i, c_0)$, where c_0 is the initial predicted target value.
Step 2	for $k = 1,...,K$: Randomly select $n' = sub\_sample \times n$ training samples and $m' = sub\_feature \times m$ features to construct the kth tree.
Step 2.1	for $i = 1,...,n'$: Calculate the residual of sample i in iteration k by $r_{ki} = -g_{ki} = y_i - f_{k-1}(\mathbf{x}_i)$. Set r_{ki} as the new prediction target value for sample i by updating the ith sample to $(\mathbf{x}_i, r_{ki})$.
Step 2.2	Use the new training set $D' = \{(\mathbf{x}_i, r_{ki}), i = 1,...,n'\}$ with m' features to train an ordinary regression tree using the CART algorithm as the kth tree in the ensemble. Especially, all the m' features and their corresponding values should be traversed to select the feature value pair leading to the minimum sum of losses in the left and right child nodes when splitting one node in the tree. The tree grows in a depth-first and recursive manner and stops growing if either of the stop criteria evaluated by *max_depth* and *min_samples_leaf* is reached. Denote the total number of leaf nodes contained in the constructed regression tree by J_k, with one leaf node denoted by $R_{kj}, j = 1,...,J_k$.
Step 2.3	for $j = 1,...,J_k$: Calculate the optimal output value of leaf j denoted by c_{kj} by $c_{kj} = \arg\min_{\tilde{c}_{kj}} \sum_{\mathbf{x}_i \in R_{kj}} L(y_i, f_{k-1}(\mathbf{x}_i) + \tilde{c}_{kj})$. Under our problem setting, c_{kj} is the mean of r_{ki} falling in this leaf node.
Step 2.4	Update the current GBRT model to $f_k(\mathbf{x}) = f_{k-1}(\mathbf{x}) + \varepsilon \sum_{j=1}^{J_k} c_{kj} I(\mathbf{x} \in R_{kj})$.
Step 3	The final GBRT model can be expressed by $f(\mathbf{x}) = f_K(\mathbf{x}) = f_{K-1}(\mathbf{x}) + \varepsilon \sum_{j=1}^{J_K} c_{Kj} I(\mathbf{x} \in R_{Kj})$.

In the first step, the optimal c_0 can be obtained by calculating the derivative of $\sum_{i=1}^{n} L(y_i, c_0)$ regarding c_0 and then set it to zero, i.e.

$$\sum_{i=1}^{n} \frac{\partial L(y_i, c_0)}{\partial c_0} = \sum_{i=1}^{n} \frac{\partial [\frac{1}{2}(y_i - c_0)]^2}{\partial c_0} = \sum_{i=1}^{n} (c_0 - y_i) = 0, \quad (3)$$

and we can have $c_0 = \frac{\sum_{i=1}^{n} y_i}{n}$, which is the average target value of all the samples. Similarly, c_{kj} is the average target value of all the samples contained in leaf j in the kth iteration. Recall that the target value of sample i in the k th iteration is the residual r_{ki} instead of the original target value y_i.

After the construction of the GBRT models, three typical regression model performance metrics are used to demonstrate model performance: mean squared error (MSE), root mean squared error (RMSE), and mean absolute error (MAE) (Li et al., 2020; Guo et al., 2022; Liu et al., 2022; and Xiao et al., 2022). The definitions of the metrics are given as follows.

$$MSE = \frac{1}{n}\sum_{i=1}^{n}[f(\mathbf{x}_i) - y_i]^2 , \tag{4}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}[f(\mathbf{x}_i) - y_i]^2} , \tag{5}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}| f(\mathbf{x}_i) - y_i | . \tag{6}$$

4.3 Ship risk prediction frameworks based on GBRT

GBRT models are developed for the T-SRP and F-SRP frameworks for ship risk prediction using the features shown in Table 2. The searching spaces of the hyperparameters are given in Table 4, and they are tuned based on 5-fold cross-validation on the training set with MSE as the metric.

Table 4. Hyperparameter tuning in T-SRP and F-SRP

Hyperparameter	T-SRP		F-SRP	
	Searching space	Value adopted	Searching space	Value adopted
n_estimators	[200, 1000] with 200 as the interval	200	[200, 1000] with 200 as the interval	200
learning_rate	{0.01,0.02,0.05,0.1,0.2}	0.02	{0.01,0.02,0.05,0.1,0.2}	0.02
max_depth	[3, 13] with 2 as the interval	9	[3, 13] with 2 as the interval	5
min_samples_leaf	[1, 9] with 2 as the interval	7	[1, 9] with 2 as the interval	3
sub_sample	{0.4,0.5,0.6,0.7,0.8}	0.4	{0.4,0.5,0.6,0.7,0.8}	0.4
sub_feature	{0.3,0.4,0.5,0.6,0.7}	0.3	{0.3,0.4,0.5,0.6,0.7}	0.3

Both frameworks are finally developed using the hyperparameter values found by hyperparameter tuning on the whole training set, and the model performance is validated on the test set. The MSE of T-SRP and F-SRP is 17.9821 and 17.2895, the RMSE is 4.2405 and 4.1581, and the MAE is 2.7564 and 2.6478, respectively. It can be seen that given the same types of features considered and using the same prediction model as well as the hyperparameter tuning method, the performance of the F-SRP framework with more flexible feature processing is much better than the T-SRP framework with rough feature processing. The results indicate that compared to data

processing based on simple threshold, more fine-grained data processing method can lead to more accurate ML models.

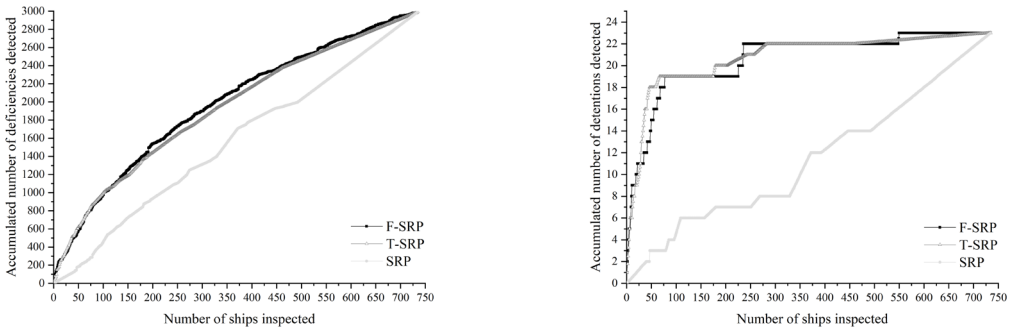
4.4 Comparison of the new frameworks and the SRP for ship risk prediction

We compare the newly proposed T-SRP framework and F-SRP framework, and the SRP framework under three comparison schemes. In scheme I, ship inspection priority in the SRP is ignored. In other words, the ship inspection sequence is purely dependent on the ship risk scores generated in each framework and the ships are inspected from high risk score to low risk score. In scheme II, ship inspection priority in the SRP is considered. Specifically, ship inspection priority from the highest to the lowest is as follows: ships with no previous inspection (P1), ships with the last inspection time beyond the upper bound of the time window (where the time window attached to each risk profile is specified in Table 1) (P2), ships with the last inspection time within the time window (P3), and ships with the last inspection time below the lower bound of the time window (P4). In comparison scheme I and scheme II, we use the total number of deficiencies and detentions detected after inspecting a certain number of ships as the performance metrics. In scheme III, we first divide the ships in the test set into high-risk, standard-risk, and low-risk types considering their predicted risk scores in T-SRP and F-SRP with the same ratios as those generated by the SRP. Specifically, the number of ships belonging to HRS, SRS, and LRS is 225, 337, and 173 in the test set, respectively. Then, we calculate the average ship deficiency number and detention within each risk type.

The ship risk scores given by the T-SRP and F-SRP are represented by the number of deficiencies predicted by the corresponding GBRT models. The ship risk score given by the SRP is calculated using the risk calculation matrix presented in Table 5 (Wang et al., 2019). As there might be ties in ship risk scores, we run each framework in each comparison scheme 1,000 times and use the mean as the result. The performance of each framework in comparison schemes I, II, and III are shown in Fig. 2, Fig. 3, and Fig. 4. The overall comparison of SRP, T-SRP, and F-SRP under each comparison scheme is summarized in Table 6.

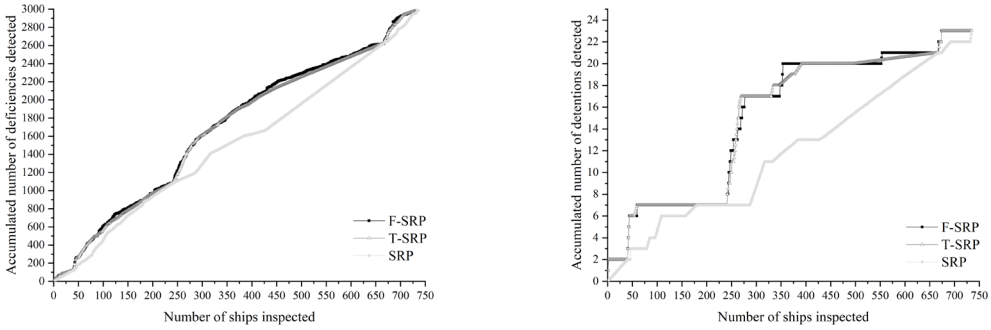
462 Table 5. Calculation of ship risk score in SRP

SRP	Time window (months)	Relationship between the last inspection time (T_l) and the time window		
		T_l beyond the upper bound of the time window	T_l within the time window	T_l below the lower bound of the time window
LRS	9 to 18	$\frac{T_l}{18}$	$\frac{T_l-9}{18-9}$	$\frac{T_l}{9}$
SRS	5 to 8	$\frac{T_l}{8}$	$\frac{T_l-5}{8-5}$	$\frac{T_l}{5}$
HRS	2 to 4	$\frac{T_l}{4}$	$\frac{T_l-2}{4-2}$	$\frac{T_l}{2}$



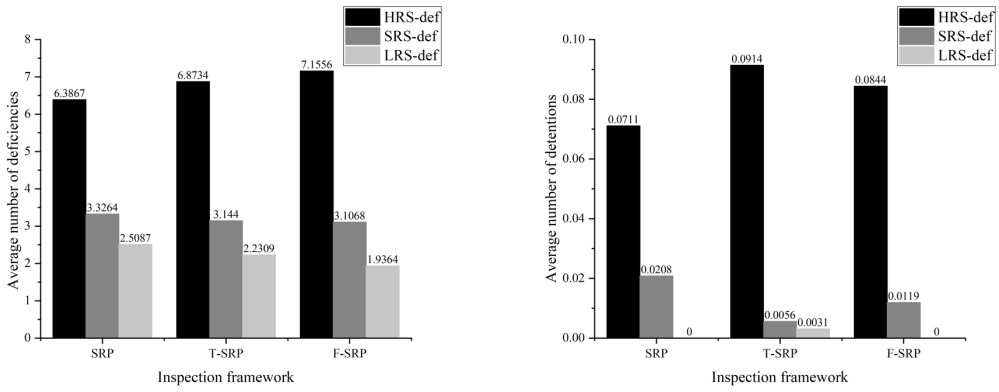
(a) Comparison of deficiency (b) Comparison of detention

463 Fig. 2. Comparison results in scheme I



(a) Comparison of deficiency (b) Comparison of detention

464 Fig. 3. Comparison results in scheme II



(a) Comparison of deficiency (b) Comparison of detention

465 Fig. 4. Comparison results in scheme III

Table 6. Summary of the comparison of SRP, T-SRP, and F-SRP

Comparison scheme	Improvement regarding the number of deficiencies detected (represented by ratio, a total of 2,992 deficiencies)		Improvement regarding the number of detentions detected (represented by absolute value, a total of 23 detentions)	
Scheme I				
T-SRP over SRP	63.71 %		9.36	
F-SRP over SRP	67.44 %		9.25	
F-SRP over T-SRP	2.24%		–0.11	
Scheme II				
T-SRP over SRP	17.48%		3.36	
F-SRP over SRP	18.48 %		3.41	
F-SRP over T-SRP	0.97%		0.05	
Scheme III	**SRP**	**T-SRP**		**F-SRP**
Average no. of deficiencies among 'HRS'	6.3867	6.8734		7.1556
Average no. of deficiencies among 'SRS'	3.3264	3.1440		3.1068
Average no. of deficiencies among 'LRS'	2.5087	2.2309		1.9364
Average no. of detentions among 'HRS'	0.0711	0.0914		0.0844
Average no. of detentions among 'SRS'	0.0208	0.0056		0.0119
Average no. of detentions among 'LRS'	0	0.0031		0

It is shown that when evaluating the performance of the frameworks by the accumulated number of deficiencies detected, the F-SRP has the best performance under comparison schemes I and II. Meanwhile, although the T-SRP performs slightly worse than the F-SRP, the T-SRP is also much better than the SRP. To be more specific, Fig. 2 (a) shows that in case of ignoring the inspection priority, when the number of inspected ships is no more than 120, sometimes the F-SRP performs moderately better than the T-SRP but sometimes not. When 120 or more ships are inspected, the F-SRP performs better than the T-SRP in most cases. The superiority of the F-SRP becomes the most obvious when the number of inspected ships is between 190 and 350. Nevertheless, both newly proposed frameworks are much better than the SRP no matter how many ships are inspected. Fig. 3 (a) indicates that when the inspection priority is considered, the performance of the T-SRP and the F-SRP is similar, while it is also obvious that after inspecting a certain number of ships, e.g., from 430 to 480, the F-SRP performs better than the T-SRP.

Even though both newly proposed frameworks aim to predict ship deficiency number, they are much more efficient than the SRP regarding the number of detentions identified after inspecting a certain number of ships as indicated by Fig. 2 (b) and Fig. 3 (b) by identifying more than 9 and 3 more detentions on average.

By comparing Fig. 2 (a) with Fig. 3 (a), it can be found that the slope of the lines in Fig. 2 (a), especially the ones representing the performance of the newly proposed frameworks, gradually reduces as the number of inspected ships increases. This

indicates that ships with a larger deficiency number can be distinguished from those with less deficiencies in both new frameworks. The lines of the newly proposed frameworks in Fig. 3 (a) are divided into four segments with 40, 241, and 667 inspected ships as the splitting points, which are the thresholds of ship inspection priorities from P1 to P4. The slope gradually decreases in each segment, which also shows that the newly proposed frameworks are effective within each inspection priority, although their effectiveness is highly compromised when considering such inspection priority. In addition, they are always much better than the SRP ship selection scheme. A similar pattern can be found in Fig. 2 (b) and Fig. 3 (b). Among all the 23 detentions, 19 of them can be identified after inspecting 66 and 77 of the 735 ships by the T-SRP and the F-SRP, while all the detentions can be found after inspecting 550 ships by the F-SRP in Scheme I, and both of which are much more effective than the SRP. In contrast, in scheme II, segmentations of the new frameworks also exist, and the number of detentions identified increases significantly at the beginning of each segment, as is the case in Fig. 3 (b).

Finally, Fig. 4 and Table 6 show that all the three frameworks are effective in classifying ships into HRS, SRS, and LRS types, as the average numbers of deficiencies and detentions gradually decrease from HRS to LRS in all frameworks. Particularly, F-SRP has the highest average number of deficiencies among the HRS at 7.2, which is larger than that in T-SRP at 6.9 and much larger than that in SRP at 6.4. Meanwhile, the F-SRP has the least average deficiency number in LRS at 1.9, while the SRP has the largest average deficiency number in LRS at 2.5. Regarding the number of detentions detected, the T-SRP has the highest average detention rate in HRS at 0.09, while the SRP has the lowest at 0.07. It is also noted that none of the LRS is detained in the SRP and the F-SRP, while the average detention rate is 0.0031 in the LRS selected by the T-SRP framework.

To conclude, Fig. 2 to Fig. 4 and Table 6 show that given a set of features with identical processing methods, a data-driven ship risk prediction model (i.e., the T-SRP) can be much more efficient than the current SRP ship selection scheme to identify ship deficiency and detention. Meanwhile, when the set of features are processed using different methods in data-driven models, the model with finer feature encoding (i.e., the F-SRP) can generate more accurate output than the model with coarser feature processing (i.e., the T-SRP). Therefore, data-driven prediction model with finer grained

feature processing has the best performance, followed by data-driven prediction model with more casual feature processing, and then by framework purely based on expert judgement and simple working mechanism. Moreover, it is evident that ships with higher priority indicated by the SRP do not necessarily have larger number of deficiencies or higher probability of detention, and vice versa. Therefore, the efficiency of the two new frameworks is compromised when considering the inspection priority.

5. XAI and its importance in maritime transport

In addition to developing highly-efficient ML based ship risk prediction frameworks for high-risk ship selection in PSC, we further try to explain the predictions given by them from various aspects. In this section, we first clarify the definition of XAI and common approaches to achieve it as well as its benefits. Factors making XAI essential in marine policy making as well as in PSC are then analyzed.

5.1 Introduction of XAI

Despite the success of ML models to address real-world problems, the most significant drawback of ML models is their lack of transparency (Du et al., 2020). As a matter of fact, ML models do not explicitly show its internal mechanisms and cannot be understood by looking at their parameters. In addition, the intermediate computation process of the output is opaque. To make the black-box ML models understandable by humans, the area of XAI gained a rapid development in recent years (Doshi-Velez and Kim, 2017). One widely used definition of XAI is given by Arrieta et al. (2020): "Given a certain audience, explainability refers to the details and reasons a model gives to make its functioning clear or easy to understand." This definition covers three key points. First, explainability should be presented to 'a certain audience', as different audiences pose different requirements for an explainable system due to different background knowledge and communication styles. Second, 'details' should answer a 'how' question: the more explainable a model is, the more detailed information about its internal structure and working process should be disclosed to an audience. Third, 'reasons' should answer a 'why' question: the more explainable a model is, the easier for a human to understand why certain predictions or recommendations have been made.

Using XAI models brings several advantages. For model developers, XAI models can help verify model accuracy and robustness with the assist of domain knowledge. If any irrationality is found, the XAI models can further contribute to model debugging.

For model users, XAI models are more likely to be trusted and accepted than black-box models with similar accuracy. Actually, model explainability is even considered as a prerequisite for the adoption of AI systems in high stakes or traditional and conservative domains where reliability, safety, and fairness are required. In addition, explicit decision rules can be extracted from the explanations, and thus shed light on future judgments and decisions for the users.

XAI techniques can be divided into two categories depending on the time when explainability is obtained: one is to develop an interpretable ML model directly, and the other is to use post-hoc explainability techniques after developing a (usually uninterpretable) ML model. Particularly, interpretable ML models are by themselves understandable, such as linear/logistic regression, decision trees, k-nearest neighbors, rule-based learning, general additive models, and Naive Bayes models. In contrast, post-hoc explainability techniques are used to explain the output of an ML model, which are further divided into model-agnostic techniques that can be applied to any ML model disregarding its inner structures and mechanisms, and model-specific techniques that are designed to explain certain ML models considering their internal structure. Popular model-agnostic techniques include PDP, ICE, ALE plot, and SHAP (Molnar, 2020). Particularly, the first two are global methods considering all samples and give a global relationship between a feature and the predicted outcome in one explanation. The others are local methods, where only part of the instances is covered in one explanation. Model-specific techniques have been designed for neural networks and tree-based models.

The major advantage of interpretable ML models is that their explainability is inherent and the prediction and explainability are consistent as both are derived from the ML model directly. However, model accuracy and interpretability need to be balanced. Usually, the higher the prediction accuracy achieved, the lower the model interpretability (Du et al., 2019; Arrieta et al., 2020; Burkart and Huber, 2021). In contrast, post-hoc explanation developed after model construction can help to ease this problem by using a white-box surrogate model of the black-box prediction model to gain explanation while keeping its high accuracy. Nevertheless, the post-hoc surrogate models might cause inconsistency due to their approximation nature (Du et al., 2020; Babic et al., 2021).

5.2 The necessity of XAI to facilitate marine policy making

When black-box ML models are used to assist policy making, a detailed understanding of the prediction model and its output are as important as the prediction accuracy. According to Doshi-Velez and Kim (2017), explainability of black-box model can only be omitted in two situations: 1) no significant consequences will be caused by unacceptable prediction results, and 2) the problem is sufficiently well-studied and the system's decision ~~are~~ is trusted even if it is not perfect. Unfortunately, neither condition is satisfied in the context of critical marine policy making. This is mainly because there are several heterogeneous and conservative stakeholders involving and the decisions are heavily dependent on long-term experience while seldom on recommendations given by data-driven models. Consequently, policy recommendations generated by black-box models without convincing explainability provided are seldom accepted, even if they could be much more efficient than recommendations made from naive but transparent rules or expert systems. One example is ship selection models in PSC: although various accurate and efficient ship selection models for substandard ship identification are proposed in several studies, they are rarely adopted by any port authority at the moment. Instead, intuitive and comprehensible ship selection schemes based on domain knowledge are preferred.

In sum, the main reasons for requiring XAI models applied to assist marine policy making are as follows:

a) Trust: conservative practitioners in the traditional maritime industry are reluctant to trust any black-box model to guide policy making. Only when they understand and verify the prediction model's internal schemes, working processes, and strengths and weaknesses, can they trust and thus use the model.

b) Transferability: Only when the policy makers know how well the prediction model generalizes, or in which context it generalizes well, can this prediction model be put in charge of policy making.

c) Fairness: As various stakeholders, such as ship owners, operators, management companies, port authorities, and shipping service providers, are influenced by maritime conventions, fairness is the key to the successful implementation of any critical marine policy. Explanations generated by XAI can help to verify the recommendations given by black-box models to be fair and compliant to ethical standards.

d) Extensibility: On the one hand, XAI enables the developers to improve the prediction model by adjusting its parameters and hyperparameters and by integrating domain knowledge. On the other hand, policy makers can extract new knowledge from massive data by the XAI models and thus to obtain insights for future decision making.

The above mentioned points are also essential for developing XAI models for ship selection in PSC (Adadi and Berrada, 2018). Similar to the situation discussed by Kleinberg et al. (2015) and Athey (2017), black-box models for ship risk prediction without explanation provided are not enough, as they cannot answer more complex question of why a certain ship should be given a higher inspection priority or what properties would increase ship risk. With the assistance of the tailored explanations given to these black-box models, the above question can be addressed to a large extent, making the models more likely to be adopted in practice and thus a larger number of substandard ships can be inspected by PSC. Therefore, the ports as well as the PSC inspection can better fulfill their responsibility to enhance the maritime safety, to protect the marine environment, and to guarantee decent living and working conditions of seafarers. For ship owners, operators, and managers, they will be more willing to accept explainable ship selection methods as both time and monetary costs can be high if their ships are frequently involved in PSC inspections. Meanwhile, fair ship selection can in turn motive them to keep their ships in satisfactory condition to reduce future inspections. For shipping service providers, they can provide tailored services by considering a ship's PSC inspection results, and thus to reduce maritime risks and pollutions.

6. Black-box model explanation using SHAP

Predictions given by the black-box GBRT models in the T-SRP and F-SRP frameworks are explained from both local and global perspectives in this section based on SHAP. We first introduce the concept of SHAP and give local explanations for the prediction of individual ships given by the GBRT models. SHAP values in both frameworks are also visualized and analyzed. Then, we go one step further to extend the local SHAP method to a global method by formulating near linear-form global surrogate models that can closely approximate the outputs of the GBRT models with full explainability. Validation of the explanation performance of the global surrogate models is finally presented.

6.1 Introduction of SHAP

SHAP is proposed by Lundberg and Lee (2017) aiming to explain the output of an individual prediction (i.e., local method) of any machine learning model (i.e., model agnostic) and it is applied after model construction (i.e., post-hoc). SHAP is based on the Shapley value from coalition game theory first developed by Shapley (1953). It assigns an additive importance value (which can be negative, zero, or positive) to each feature as its contribution to the prediction. Therefore, the prediction is similar to a near linear model by summing the base value, which is the mean of the outputs in the training set denoted by $\bar{y}$, and the contributions of all the features. In the context of XAI, the 'game' refers to the prediction task of a sample, the 'players' are the features included in the model, and the 'gain' is the difference between the actual prediction and the base value.

The basic idea of applying Shapley value to XAI is that the marginal contribution of a single feature concerned is determined by the differences in the outputs of the possible combinations of features with and without this feature. There are several algorithms to calculate SHAP values with different tricks to reduce the computational burden. Here we briefly introduce a basic but easy-to-digest one. To begin with, a power set of features with different feature coalitions ranging from no feature contained to all features contained presented by a tree structure are formulated as shown in Fig. 5, where each node represents a coalition of features, and each edge indicates adding a feature excluded in the coalition at the head to the coalition at the tail. l is the depth of the tree. Given the dataset in our problem with m features, we can have a total of 2^m coalitions of features, and thus 2^m nodes in the tree.

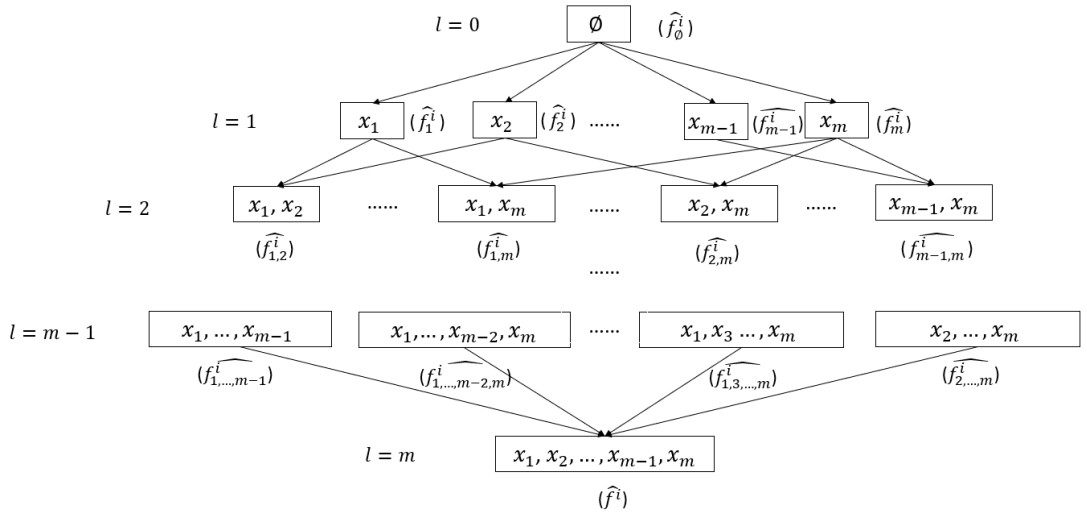


Fig. 5. An illustration of feature coalitions

Suppose we want to explain the prediction of sample D_i given by the developed ML model. After deciding the feature coalitions, the next step is to decide the predicted target value of D_i given by the ML model using the feature coalition contained in each node. The feature(s) contained in each node is(are) input to the developed ML model, while the absent feature(s) is(are) replaced by a random feature value from the data. The predicted target value of each of the 2^m feature coalitions is presented on the right of or below the corresponding node in Fig. 5. Particularly, the output of the node containing no feature at the root of the tree is $\hat{f}_{\varnothing}$, which is the average target values in the training set called the base value. As shown in Fig. 5, the difference between two nodes lies in just one feature. Therefore, the prediction difference between these two nodes connected by an edge can be regarded as the effect, or the marginal contribution, brought by that additional feature (Lundberg and Lee, 2017). For example, if we only consider the first two layers, the marginal contribution of feature x_1 regarding sample D_i can be presented by $\hat{f}_1^i - \hat{f}_{\varnothing}$.

One last question is how to combine the marginal contribution of each feature presented by different node pairs connected by the edges where the feature is not contained in the node at the head but is contained in the node at the tail in Fig. 5. The weights connecting all the node pairs in consecutive layers l and $l+1$, $l \in [0, m-1]$, are required to be equal and are denoted by $w_{l,l+1}$. For feature x_1, the overall effect of its marginal contribution, which is also called the SHAP value of feature x_1, is denoted by ϕ_1^i and can be calculated by

$$\begin{aligned}\phi_1^i = {} & w_{0,1} \times (\hat{f}_1^i - \hat{f}_{\varnothing}) + \\ & [w_{1,2} \times (\hat{f}_{1,2}^i - \hat{f}_2^i) + \ldots + w_{1,2} \times (\hat{f}_{1,m}^i - \hat{f}_m^i)] + \\ & [w_{2,3} \times (\hat{f}_{1,2,3}^i - \hat{f}_{2,3}^i) + \ldots + w_{2,3} \times (\hat{f}_{1,m-1,m}^i - \hat{f}_{m-1,m}^i)] + \\ & \ldots \\ & + [w_{m-1,m} \times (\hat{f}^i - \hat{f}_{2,\ldots,m}^i)]\end{aligned}, \tag{7}$$

where the sum of all the weights is 1. The sum of weights connecting each two consecutive layers is further required to be equal, and thus the weights connecting layer l and $l+1$ is $w_{l,l+1} = [(l+1) \times C_m^{l+1}]^{-1}, l \in [0, m-1]$, where $C_m^{l+1} = \binom{m}{l+1} = \frac{m!}{(l+1)!(m-l-1)!}$.

The SHAP value or the feature importance of a feature m', $m' \in [1, m]$ regarding sample i, can therefore be calculated by

$$\phi_{m'}^{i} = \sum_{S \subseteq M \setminus \{m'\}} \frac{|S|!(m-|S|-1)!}{m!} [\hat{f}_{S \cup \{m'\}}^{i} - \hat{f}_{S}^{i}], \tag{8}$$

where M is the set of all features. Finally, according to the 'local accuracy' property of SHAP indicated by Lundberg and Lee (2017), summing the Shapley values of all features of sample D_i yields the difference between its predicted output and the base value, where the sum of Shapley values can be regarded as the effects of all the features on the output of this sample. Therefore, the predicted output of sample D_i can also be represented in an additive linear function form as follows:

$$f(\mathbf{x}_i) = \overline{y} + \sum_{m'=1}^{m} \phi_{m'}^{i} . \tag{9}$$

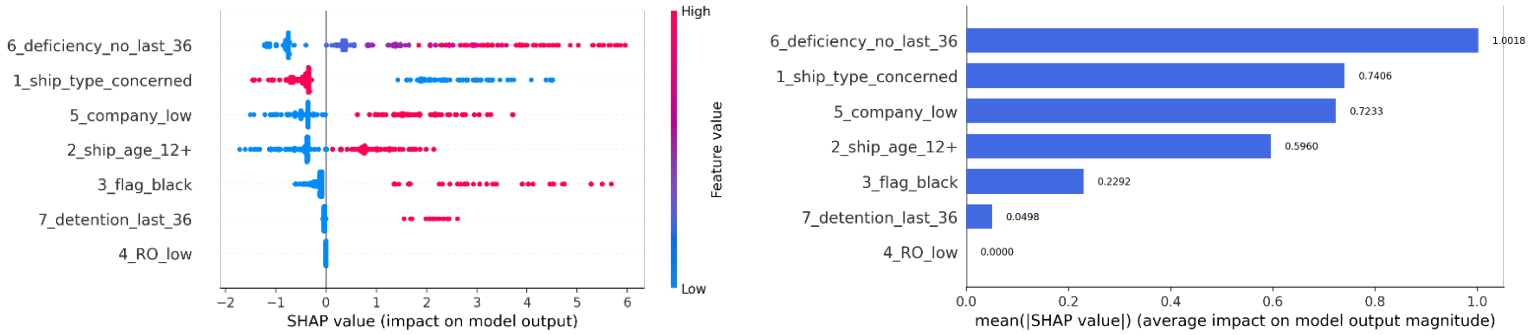
The above algorithm for SHAP value calculation is computationally expensive as its needs to predict the targets for a total of 2^m times using different feature coalitions. Fortunately, efficient implementations to calculate SHAP values are proposed by several studies such as Lundberg and Lee (2017) and Lundberg et al. (2019) which can be found from the SHAP API for Python (Lundberg, 2021). The SHAP values are calculated based on the implementation of Lundberg et al. (2019) for tree-based models in this study.

6.2 Explanation of GBRT via SHAP

SHAP is originally designed for local explanation, which aims at knowing the reasons for a specific prediction (such as why a particular ship is predicted to have a certain number of deficiencies). In the following subsections, we first give an overview of local feature effects in the T-SRP and F-SRP frameworks and the global feature importance derived from the local explanations. Then, we explain specific predictions in the test set.

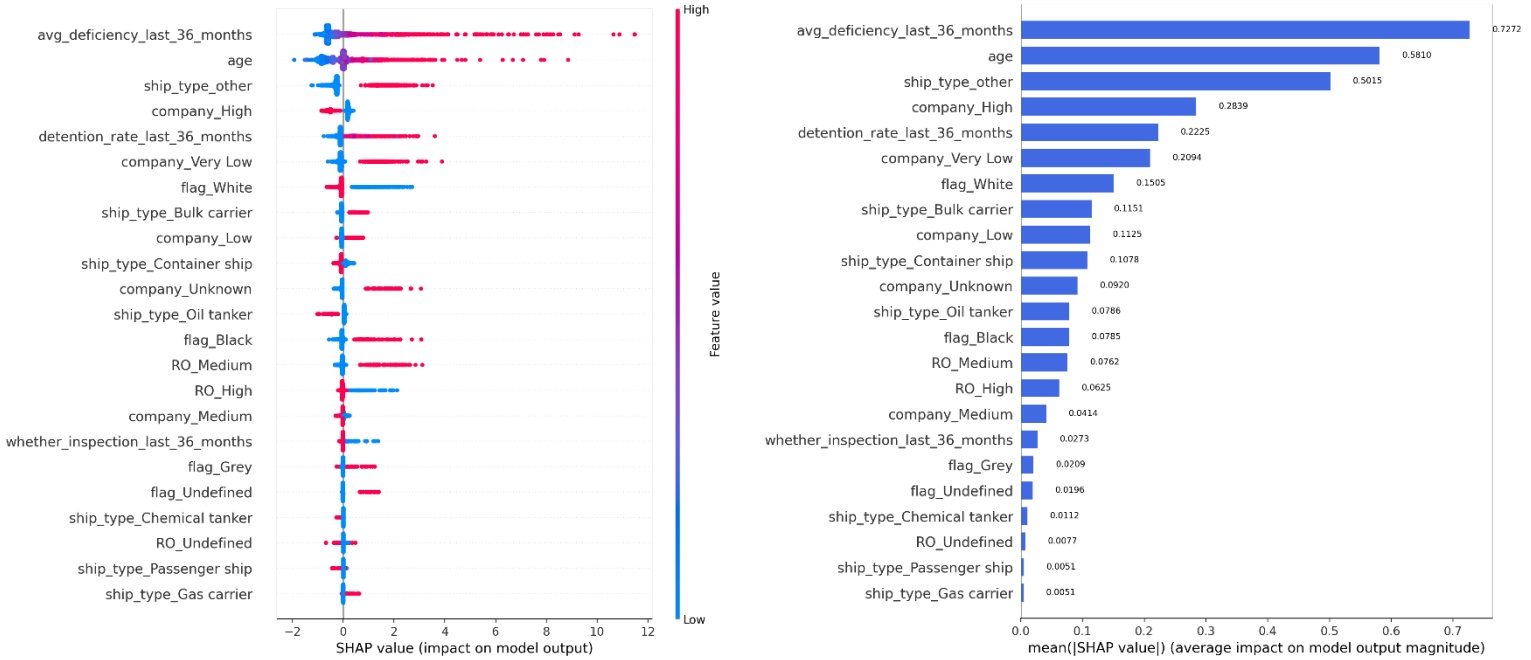
6.2.1 Model explanation based on the training set

Feature SHAP values in the training set and the global feature importance in the T-SRP and F-SRP frameworks are presented in Fig. 6 and Fig. 7.



(a) Local explanation summary (b) Global feature importance

Fig. 6. Local explanation summary and global feature importance in the T-SRP framework



(a) Local explanation summary (b) Global feature importance

Fig. 7. Local explanation summary and global feature importance in the F-SRP framework

Fig. 6 (a) and Fig. 7 (a) are a set of beeswarm plots with *y*-axis representing each feature and *x*-axis representing the features' SHAP values, while each dot in a figure represents a single ship in the training set. Feature values from low to high are shown by gradient colors as illustrated by the chromatographic on the right side, and the dot's position on the *x*-axis shows the impact that feature value has on the ship's predicted deficiency number given by the GBRT model, i.e., the SHAP value of the feature value for each ship. When multiple dots land at the same *x* position, they pile up to show the density. Fig. 6 (b) and Fig. 7 (b) are bar charts showing the importance of each feature calculated by the mean absolute SHAP values of a feature among all the samples in the

741 training set. The larger a feature's mean absolute SHAP value, the greater influence the
742 feature has on the prediction as it can change the predicted target more.
743 Fig. 6 (b) shows that in the T-SRP framework, the only integer variable, i.e., the
744 number of inspections with over 5 deficiencies within previous 36 months, has the
745 highest feature importance. Fig. 6 (a) indicates that a larger value of this feature leads
746 to a larger predicted deficiency number. Especially, the highest feature values (e.g.,
747 more than 20) can increase the final prediction by more than 6. In contrast, if there is
748 no inspection with over 5 deficiencies in previous 36 months, the final prediction will
749 be reduced by 0 to 2. Among the binary features, whether a ship is of a certain ship type
750 concerned has the largest feature importance, followed by whether a ship has low
751 performance management company. It is interesting to find that if a ship is of a type of
752 concern, less deficiencies will be found; otherwise, much more deficiencies ranging
753 from 1 to 5 will be found. This finding shows that the other features override the feature
754 of ship type when deciding ship risk level.
755 Moreover, if the performance of a ship's management company is evaluated to be
756 low, very low, or its performance is not listed by the Tokyo MoU, up to 4 more
757 deficiencies can be found compared to the base value. Regarding ship age, it is not
758 surprising to find that if ship age is more than 12, much more deficiencies will be
759 detected; if not, up to 2 less deficiencies will be found compared to the base value. Fig.
760 6 also indicates that although features 3_flag_black and 7_detention_last_36 are less
761 important in the T-SRP framework, if a ship's flag is on the black-list or it is detained 3
762 times or more within the previous 36 months, its predicted number of deficiencies will
763 be increased by 1 to 6 and 1 to 3, respectively. Finally, as there is no ship with low or
764 very low RO performance in the training set, i.e., 4_RO_low is 0 for all the samples,
765 this feature will not influence the prediction results and thus it has zero feature
766 importance.
767 As shown in Fig. 7, similar pattern regarding continuous variables
768 (avg_deficiency_last_36_months, age, and detention_rate_last_36_month) is shown in
769 the F-SRP framework, where all of them are among the 5 most important features. Fig.
770 7 (a) shows that generally, the larger values of the three features, the more deficiencies
771 will be detected. Particularly, large values of avg_deficiency_last_36_months and age,
772 which are the two most important features, can result in nearly 12 and 10 more
773 deficiencies, respectively. For most binary variables, the influence of feature value (0

or 1) on the predicted deficiency number is clear. For example, ships belonging to types 'other' or 'bulk carriers', with very low or unknown company performance, with flag performance on black-list or unknown, and with RO of medium performance would increase the predicted deficiency number. In addition, ship flag not on white-list, with RO performance not high, and without inspection within last 36 months would also increase the predicted number of deficiencies. Nevertheless, it is also found that the influence of binary values is unclear in some features. For example, low company performance or grey-list flag performance would increase the predicted deficiency number in most cases; however, it would also decrease the predicted deficiency number sometimes as they could also be overriden by other features considered in some cases.

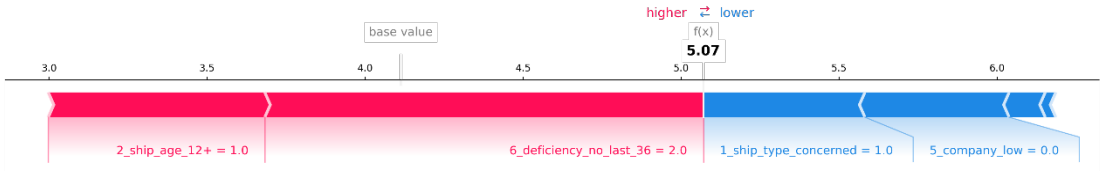
The explanations based on feature SHAP values in both the T-SRP and the F-SRP frameworks indicate that port authorities should pay more attention to ships with worse performance in the last 36 months, especially those with larger deficiency numbers detected. In addition, older ships, ships of certain types (e.g., bulk carrier, other type, and gas carrier), and ships with worse performance management organizations especially the ISM company should also receive more attention.

6.2.2 Model explanation in the test set

This section aims to explore feature contributions in specific samples. The feature values and the SHAP values as well as the prediction results of two samples in the test set are shown in Tables 7 and 8. Visualization of major features' contribution in the T-SRP and the F-SRP is given in Fig. 8 to Fig. 11.

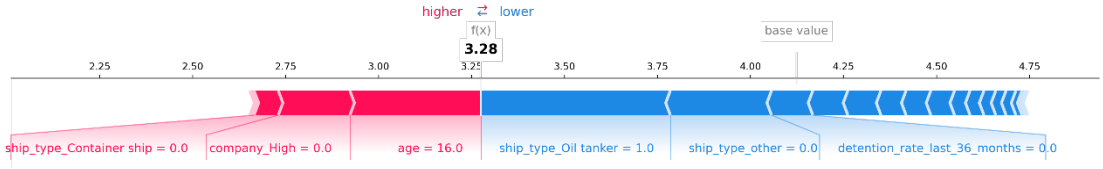
795 Table 7. Feature values and the corresponding SHAP values of sample ship 1

Parameters	Ship feature	Feature in T-SRP	Feature value in T-SRP	SHAP value in T-SRP	Feature in F-SRP	Feature value in F-SRP	SHAP value in F-SRP
Ship type	Oil tanker	1_ship_type_concerned	1	–0.511285	ship_type_Bulk carrier	0	–0.063852
					ship_type_Chemical tanker	0	0.007691
					ship_type_Container ship	0	0.065138
					ship_type_Gas carrier	0	–0.002802
					ship_type_Oil tanker	1	–0.509014
					ship_type_Passenger ship	0	0.002955
					ship_type_other	0	–0.272122
Ship age	16	2_ship_age_12+	1	0.685112	age	16	0.352818
Flag performance in Black-Grey-White list of Tokyo MoU	White	3_flag_black	0	–0.114320	flag_White	1	–0.066574
					flag_Grey	0	–0.004192
					flag_Black	0	–0.026481
					flag_Undefined	0	–0.008269
RO performance in Tokyo MoU	High	4_RO_low	0	0	RO_High	1	–0.031784
					RO_Medium	0	–0.037375
					RO_Undefined	0	0.000339
Company performance in Tokyo MoU	Medium	5_company_low	0	–0.456442	company_High	0	0.192526
					company _Medium	1	–0.027626
					company _Low	0	–0.094147
					company _Very Low	0	–0.089074
					company _Unknown	0	–0.041666
Number of deficiencies in each inspection within previous 36 months	0 0 8 0 7 3	6_deficiency_no_last_36	2	1.387823	avg_deficiency_last_36_months	3	–0.067552
Detention condition in inspections within previous 36 months	No no no no no no	7_detention_last_36	0	–0.033159	detention_rate_last_36_months	0	–0.111055
If there is any inspection in the last 36 months	Yes	\	\	\	whether_inspection_last_36_months	1	–0.016488
Real deficiency number		4					
Base value		4.112735			4.123770		
Sum of feature SHAP values		0.957728			–0.848605		
Predicted deficiency number		5.070463			3.275165		
Difference between real and predicted deficiency number		–1.070463			0.724835		



796

797 Fig. 8. Major feature contribution of sample ship 1 in the T-SRP framework



798

799 Fig. 9. Major feature contribution of sample ship 1 in the F-SRP framework

800 It is noted that the base value in the T-SRP framework is slightly different from
801 that in the F-SRP framework. This is because a subsample is randomly selected to
802 construct each tree in a GBRT model, and the average output value in the training set
803 is calculated based on the selected subsamples, and thus leading to some variations in
804 the base value obtained. Specifically, in the T-SRP framework, compared to the base
805 value at 4.11, the increase of the predicted deficiency number is mainly caused by 2

inspections with 5 or more deficiencies in the last 36 months and ship age more than 12 by 2.07, while the final prediction is mainly reduced by being the type of ship concerned and with ship company performance not low, very low, or undefined by 0.97. The sum of all the feature SHAP values is 0.96, and thus the final prediction is 5.07. In the F-SRP framework where the base value is 4.12, the main contributors to increase the final predicted deficiency number are ship age at 16, company performance not high, and not a container ship with the total increasing effect at 0.61. Feature values reducing the final prediction are being an oil tanker and not of other ship type, as well as with zero detention rate in the last 36 months with the contribution at –0.89. Overall, the final predicted deficiency number is 3.23.

Table 8. Feature values and the corresponding SHAP values of sample ship 2

Parameters	Ship feature	Feature in T-SRP	Feature value in T-SRP	SHAP value in T-SRP	Feature in F-SRP	Feature value in F-SRP	SHAP value in F-SRP
Ship type	Bulk carrier	1_ship_type_concerned	1	–0.349309	ship_type_Bulk carrier	1	0.469987
					ship_type_Chemical tanker	0	0.008032
					ship_type_Container ship	0	0.140246
					ship_type_Gas carrier	0	–0.004188
					ship_type_Oil tanker	0	0.048352
					ship_type_Passenger ship	0	0.002348
					ship_type_other	0	–0.240308
Ship age	6	2_ship_age_12+	0	–0.377828	age	6	–0.587908
Flag performance in Black-Grey-White list of Tokyo MoU	White	3_flag_black	0	–0.084827	flag_White	1	–0.058509
					flag_Grey	0	–0.009244
					flag_Black	0	–0.031942
					flag_Undefined	0	–0.009384
RO performance in Tokyo MoU	High	4_RO_low	0	0	RO_High	1	–0.027627
					RO_Medium	0	–0.036413
					RO_Undefined	0	0.001882
Company performance in Tokyo MoU	Medium	5_company_low	0	–0.357464	company_High	0	0.206370
					company _Medium	1	–0.021865
					company _Low	0	–0.062560
					company _Very Low	0	–0.089145
					company _Unknown	0	–0.044613
Number of deficiencies in each inspection within previous 36 months	0 0 3	6_deficiency_no_last_36	0	–0.747952	avg_deficiency_last_36_months	1	–0.615608
Detention condition in inspections within previous 36 months	no no no	7_detention_last_36	0	–0.028229	detention_rate_last_36_months	0	–0.133514
If there is any inspection in the last 36 months	Yes	\	\	\	whether_inspection_last_36_months	1	–0.007005
Real deficiency number		3					
Base value		4.112735			4.123771		
Sum of feature SHAP values		–1.945609			–1.102616		
Predicted deficiency number		2.167127			3.021154		
Difference between real and predicted deficiency number		0.832873			–0.021154		

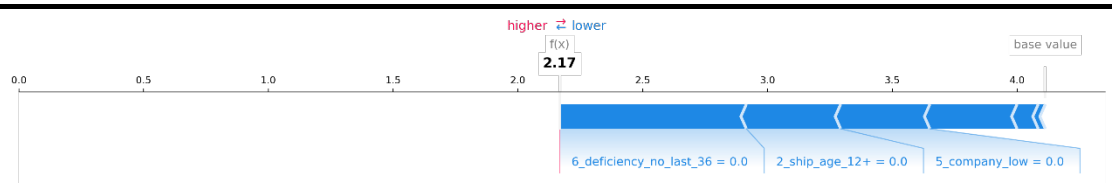


Fig. 10. Major feature contribution of sample ship 2 in the T-SRP framework

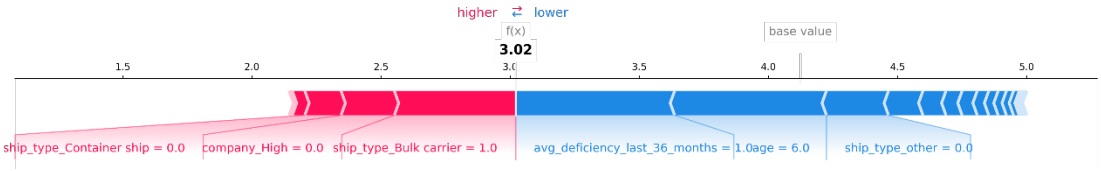


Fig. 11. Major feature contribution of sample ship 2 in the F-SRP framework

Table 8 indicates that in the T-SRP framework, there is no feature with increasing effects on the predicted number of deficiencies compared to the base value. Especially, ship features of no inspection with over 5 deficiencies in previous 36 months, ship age less than 12, and with company performance not low, very low, or undefined (actually medium) contribute the most to the difference between the final prediction and the base value. The total contribution of these features is −1.95, and hence the final predicted deficiency number is 2.17 given the base value 4.11. In the F-SRP framework, features with positive effects on the prediction results outnumber those with negative effects. Nevertheless, Fig. 11 also clearly indicates that features of ship type as bulk carrier, ship company performance not high, and ship type not as container ship increase the final prediction by 0.82. These increasing effects are weakened by features decreasing the predicted deficiency number, especially the average deficiency number within last 36 months as only 1, young age at 6, and not of other ship type. Consequently, the predicted deficiency number is decreased by −1.10 by all contributors, and the final prediction is 3.02 given the model base value 4.12.

Several findings can be drawn after analyzing sample ships 1 and 2. First, explaining the final prediction of a single ship using feature SHAP values makes the decision making process of the black-box GBRT models transparent. Such explanation makes the new ship selection frameworks more convinced by the PSC officers. Second, the same feature value can have quite different effects on different samples, and the determinant features of the final prediction are varied among different samples. This is mainly because the features considered interact with each other. As the number of features increases, such interaction effects become more complex. Third, although SHAP facilitates the explanation of a single sample in a white-box manner, the explanation within the context of T-SRP with much less features can be more intuitive than that within the context of F-SRP. Moreover, decoding the features into binary variables in the T-SRP model makes its explanation more understandable. However, such processing simplifies the original features, and thus there will be many samples

with the same feature values in the SRP even if the original samples are very different from each other. Consequently, the black-box model's predictive power is mitigated.

6.3 Development of an interpretable global surrogate model based on SHAP: one step further

The above analysis is focused on generating local model explanation, which is the original aim of SHAP. To explain the overall performance of the GBRT models from a global perspective, we innovatively extend the local SHAP method to a global method by fitting a near linear-form global surrogate model where the parameters are derived from the SHAP value matrix of the samples in the training set.

6.3.1 Main parts of the interpretable global surrogate model

As shown in Table 2, the T-SRP framework contains 6 binary features and 1 integer feature, while the F-SRP framework contains 20 binary features and 3 continuous features. For each binary feature, we calculate the average SHAP values when it takes the value 0 or 1 in the training set as its coefficient in the surrogate model. Specifically, denote a binary feature by $b_{\hat{m}}$, the value of $b_{\hat{m}}$ in sample i is $b_{\hat{m}}^{i}$ and the corresponding SHAP value is $\phi_{\hat{m}}^{i}$. The average Shapely value of $b_{\hat{m}}$ when it takes 1 (denoted by $\phi_{\hat{m}\_1}$) and when it takes 0 (denoted by $\phi_{\hat{m}\_0}$) in the whole training set can be calculated by Eq. (10) and Eq. (11), respectively:

$$\phi_{\hat{m}\_1} = \frac{\sum_{i=1}^{n} \phi_{\hat{m}}^{i} \times b_{\hat{m}}^{i}}{\sum_{i=1}^{n} b_{\hat{m}}^{i}}, \tag{10}$$

$$\phi_{\hat{m}\_0} = \frac{\sum_{i=1}^{n} \phi_{\hat{m}}^{i} \times (1 - b_{\hat{m}}^{i})}{\sum_{i=1}^{n} (1 - b_{\hat{m}}^{i})}. \tag{11}$$

For each integer and continuous feature, we fit its feature values and the corresponding SHAP values using three types of curves: linear curve, quadratic curve, and the mean squared root (sqrt) curve. Specifically, denote a continuous feature by $c_{\hat{m}}$ and the SHAP value calculated by the three modes by $\phi_{c_{\hat{m}}}^{\text{linear}}$, $\phi_{c_{\hat{m}}}^{\text{quadratic}}$, and $\phi_{c_{\hat{m}}}^{\text{sqrt}}$ which are presented by Eq. (12) to Eq. (14):

$$\phi_{c_{\hat{m}}}^{\text{linear}} = a^{\text{linear}} + b^{\text{linear}} \times c_{\hat{m}}, \tag{12}$$

$$\phi_{c_{\hat{m}}}^{\text{quadratic}} = a^{\text{quadratic}} + b^{\text{quadratic}} \times c_{\hat{m}} + c^{\text{quadratic}} \times c_{\hat{m}}^{\ 2}, \tag{13}$$

$$\phi_{c_{\hat{m}}}^{\text{sqrt}} = a^{\text{sqrt}} + b^{\text{sqrt}} \times \sqrt{c_{\hat{m}}}. \tag{14}$$

As quadratic curve has the most complex form (three parameters in contrast to two parameters in linear and sqrt modes) and thus is more likely to overfit the data, it will be selected only when its R^2 is higher than that of linear mode and sqrt mode by no less than 0.1. Otherwise, linear or sqrt curve with a higher R^2 will be selected. Finally, the prediction of sample i by the global surrogate model can be presented by

$$\hat{y}_i' = \overline{y} + \sum_{b_{\hat{m}} \in B} [\phi_{\hat{m}\_1} \times b_{\hat{m}}^i + \phi_{\hat{m}\_0} \times (1 - b_{\hat{m}}^i)] + \sum_{c_{\hat{m}} \in C} \sum_{\text{mode} \in \{\text{linear, quadratic, sqrt}\}} z_{c_{\hat{m}}}^{\text{mode}} \times \phi_{c_{\hat{m}}^i}^{\text{mode}}, \quad (15)$$

where B and C are the set of binary features and the set of integer or continuous features, respectively, $c_{\hat{m}}^i$ is the feature value of $c_{\hat{m}}$ of sample i, $z_{c_{\hat{m}}}^{\text{mode}} \in \{0,1\}$ indicates the fitting mode of feature $c_{\hat{m}}$ and $\sum_{\text{mode} \in \{\text{linear, quadratic, sqrt}\}} z_{c_{\hat{m}}}^{\text{mode}} = 1, \forall c_{\hat{m}} \in C$. It should also be mentioned that Eq. (15) can easily be extended to contain classification features taking more than 2 values by treating them as continuous or integer values and then fitting the curves of feature values and the corresponding SHAP values. Alternatively, the values can also be treated separately by calculating the average SHAP value of each feature value of the classification feature.

6.3.2 Construction of an interpretable global surrogate model for T-SRP

The average feature effects of the T-SRP framework are shown in Table 9. The relationship between the feature values and the SHAP values of the integer feature 6_deficiency_no_last_36 is shown in Fig. 12. The fitting curve form and the fitting performance are summarized in Table 10.

Table 9. Average SHAP values of the binary features in T-SRP

Binary feature	Average SHAP of value 1	Average SHAP of value 0
1_ship_type_concerned (x_1^T)	–0.4454	2.3722
2_ship_age_12+ (x_2^T)	0.8501	–0.4540
3_flag_black (x_3^T)	3.1336	–0.1342
4_RO_low (x_4^T)	0	0
5_company_low (x_5^T)	1.7802	–0.4434
7_detention_last_36 (x_7^T)	2.1108	–0.0307

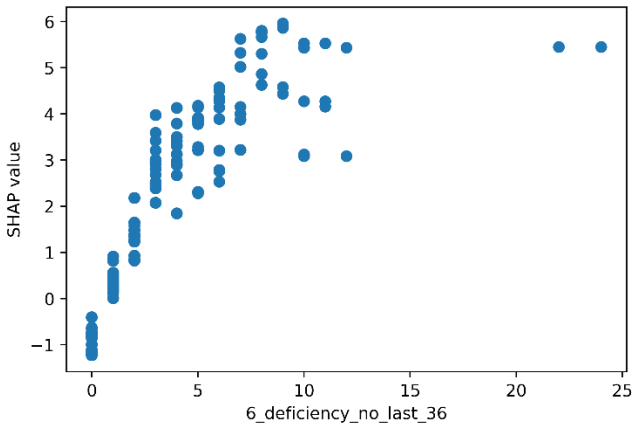


Fig. 12. Relationship between feature value and SHAP value of feature '6_deficiency_no_last_36'

Table 10. Curve fitting performance of feature '6_deficiency_no_last_36'

Integer feature	linear mode	quadratic mode	sqrt mode
6_deficiency_no_last_36 (x_6^T)	$\phi_{x_6^T}^{\text{linear}} = -0.6036 + 0.7462 \times x_6^T$	$\phi_{x_6^T}^{\text{quadratic}} = -0.7553 + 1.0864 \times x_6^T - 0.0403 \times (x_6^T)^2$	$\phi_{x_6^T}^{\text{sqrt}} = -0.8871 + 1.6953 \times \sqrt{x_6^T}$
R^2	0.8383	0.9477	**0.9200**

The sqrt mode is selected to fit the curve of the feature values and their SHAP values of 6_deficiency_no_last_36 in the T-SRP. As the base value of the T-SRP is 4.112735, the near linear form global surrogate model of the T-SRP framework, which is denoted by T-SRP-XAI, can be presented by

$$\hat{y}_i^T = 4.112735 + \left\{ \begin{array}{l} x_1^{T,i} \times (-0.4454) + (1 - x_1^{T,i}) \times 2.3722 + x_2^{T,i} \times 0.8501 + (1 - x_2^{T,i}) \times (-0.4540) + \\ x_3^{T,i} \times 3.1336 + (1 - x_3^{T,i}) \times (-0.1342) + x_5^{T,i} \times 1.7802 + (1 - x_5^{T,i}) \times (-0.4434) + \\ x_7^{T,i} \times 2.1108 + (1 - x_7^{T,i}) \times (-0.0307) + \left[-0.8871 + 1.6953 \times \sqrt{x_6^T} \right] \end{array} \right\}, \quad (16)$$

where the term in the curly brackets is the sum of feature effects. Eq. (16) is a fully white-box model in a near linear form‡ showing the decision process of the T-SRP framework, which is basically consistent with shipping domain knowledge, i.e., older ships, ships with flag on the black-list, low/very low RO performance, low/very low/undefined company performance, larger number of deficiencies and more detentions in recent inspections are more likely to lead to a larger number of

‡ We are fully noted that strictly speaking, $\hat{y}_i^T$ does not take a linear form as it contains a squared root item. Nevertheless, the squared root item can easily be transformed to a linear item by converting all the values of x_6^T into their arithmetic squared root and then feeding to Eq. (16).

deficiencies in the current inspection. This verification has greatly increased the transparency and credibility of the T-SRP framework, and thus makes it more acceptable by shipping practitioners. However, it is also noted that the only difference between the T-SRP and the SRP is that ships of certain types of concern, i.e., chemical tanker, gas carrier, oil tanker, bulk carrier, passenger ship, and container ship are instead with much smaller deficiency number than other types.

Furthermore, the T-SRP-XAI also offers insights into high-risk ship identification from a qualitative perspective. For example, ships with flag on black-list, not of the type concerned, and with no less than 3 detentions in the last 36 months should receive more attention. Finally, it is interesting to find that different values of the same binary feature can have different effects on the final prediction. For example, 0.85 more deficiency will be detected if a ship is more than 12 years old, while 0.45 less deficiency will be detected, otherwise. The absolute difference between the two average SHAP values is 1.3. In contrast, for binary feature such as 3_flag_black, the difference reaches 3.0, indicating that the situations of ships with flag performance not on the black-list is complex and hence their effects can be divergent, that is, ships with flag on the white-list and grey-list can be quite different.

We then apply Eq. (16) to predict the deficiency number of the samples in the test set. The MSE, RMSE, and MAE on the test is 18.4831, 4.2992, and 2.7909. Compared to the T-SRP framework, whose MSE, RMSE, and MAE is 17.9821, 4.2405, and 2.7564, the accuracy of the T-SRP-XAI is lower due to the approximation of feature effects. However, the sacrifice of model accuracy results in a globally fully-interpretable model presented in a near linear form, enabling the recommendations given by the black-box GBRT model of the T-SRP framework totally transparent and verifiable. Further experiments show that the MSE and MAE between the prediction of T-SRP and the prediction of T-SRP-XAI are only 0.9262 and 0.6233, respectively.

6.3.3 Construction of an interpretable global surrogate model for F-SRP

The situation is a little more complex in the F-SRP model, as it has much more features. The average feature effects in the F-SRP framework are shown in Table 11. The relationship between the feature values and the SHAP values of the continuous features, i.e., age, avg_deficiency_last_36_months, and detention_rate_last_36_months, are shown in Fig. 13 to Fig. 15. The fitting curves and the fitting performance are summarized in Table 12.

Table 11. Average SHAP values of the binary features in F-SRP

Binary feature	Average SHAP of value 1	Average SHAP of value 0
ship_type_Bulk carrier ($x^F_{1\_1}$)	0.444786	–0.069018
ship_type_Chemical tanker ($x^F_{1\_2}$)	–0.134718	0.006498
ship_type_Container ship ($x^F_{1\_3}$)	–0.093260	0.126920
ship_type_Gas carrier ($x^F_{1\_4}$)	0.171025	–0.002811
ship_type_Oil tanker ($x^F_{1\_5}$)	–0.470995	0.046185
ship_type_Passenger ship ($x^F_{1\_6}$)	–0.084661	0.003549
ship_type_other ($x^F_{1\_7}$)	1.628127	–0.297624
flag_White ($x^F_{3\_1}$)	–0.087689	1.005036
flag_Grey ($x^F_{3\_2}$)	0.286741	–0.010743
flag_Black ($x^F_{3\_3}$)	1.115679	–0.044236
flag_Undefined ($x^F_{3\_4}$)	1.081154	–0.010516
RO_High ($x^F_{4\_1}$)	–0.030582	0.775022
RO_Medium ($x^F_{4\_2}$)	1.590331	–0.039581
RO_Undefined ($x^F_{4\_3}$)	–0.096012	0.003828
company_High ($x^F_{5\_1}$)	–0.538600	0.190329
company _Medium ($x^F_{5\_2}$)	–0.029329	0.052223
company _Low ($x^F_{5\_3}$)	0.512541	–0.060889
company _Very Low ($x^F_{5\_4}$)	1.649439	–0.047817
company _Unknown ($x^F_{5\_5}$)	1.477573	–0.111640
whether_inspection_last_36_months (x^F_8)	–0.015759	0.137883

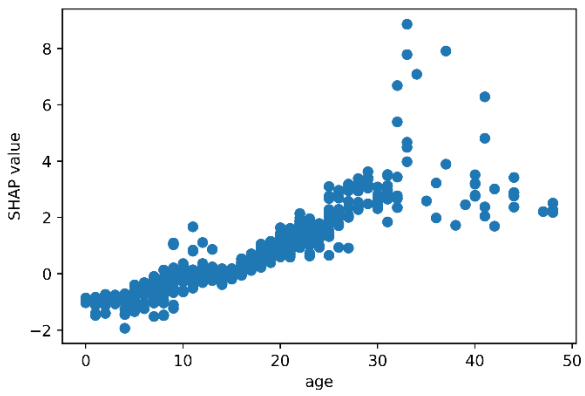


Fig. 13. Relationship between feature value and SHAP value of feature ‘age’

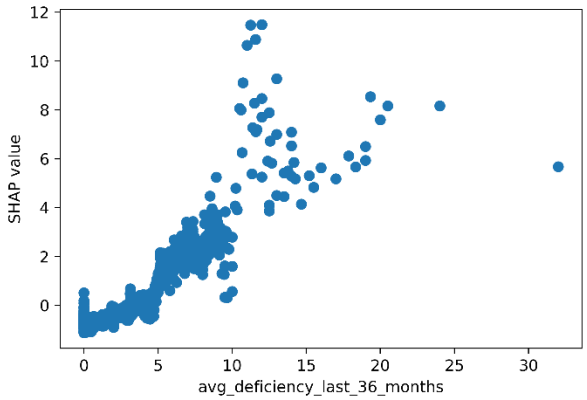


Fig. 14. Relationship between feature value and SHAP value of feature 'avg_deficiency_last_36_months'

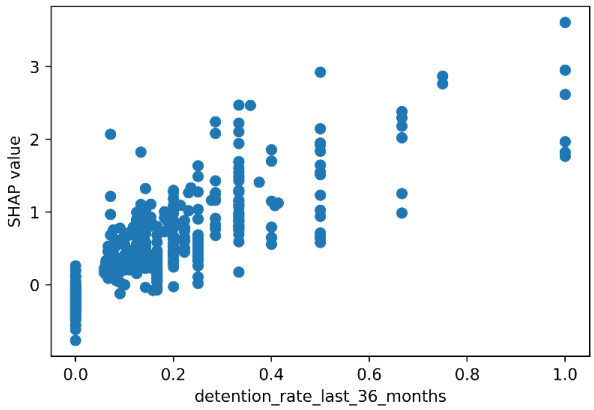


Fig. 15. Relationship between feature value and SHAP value of feature 'detention_rate_last_36_months'

Table 12. Curve fitting performance of continuous features in the F-SRP

Continuous feature	linear mode	quadratic mode	sqrt mode
age (x_2^F)	$\phi_{x_2^F}^{\text{linear}} = -1.2776 + 0.1156 \times x_2^F$	$\phi_{x_2^F}^{\text{quadratic}} = -1.1594 + 0.0938 \times x_2^F + 0.0007 \times (x_2^F)^2$	$\phi_{x_2^F}^{\text{sqrt}} = -2.2400 + 0.7123 \times \sqrt{x_2^F}$
R^2	**0.8369**	0.8419	0.7441
avg_deficiency_last_36_months (x_6^F)	$\phi_{x_2^F}^{\text{linear}} = -0.9837 + 0.4198 \times x_6^F$	$\phi_{x_2^F}^{\text{quadratic}} = -0.9469 + 0.3913 \times x_6^F + 0.0025 \times (x_6^F)^2$	$\phi_{x_6^F}^{\text{sqrt}} = -1.3440 + 1.0641 \times \sqrt{x_6^F}$
R^2	**0.8020**	0.8035	0.5536
detention_rate_last_36_months (x_7^F)	$\phi_{x_7^F}^{\text{linear}} = -0.1263 + 3.4134 \times x_7^F$	$\phi_{x_7^F}^{\text{quadratic}} = -0.1364 + 4.3012 \times x_7^F - 1.7303 \times (x_7^F)^2$	$\phi_{x_7^F}^{\text{sqrt}} = -0.1470 + 1.9878 \times \sqrt{x_7^F}$
R^2	0.7879	0.8075	**0.7889**

We choose linear curve, linear curve, and sqrt curve to fit the relationship between age values and their SHAP values, avg_deficiency_last_36_months values and their SHAP values, and detention_rate_last_36_months values and their SHAP values in the

F-SRP, respectively. The global surrogate model of the F-SRP is denoted by F-SRP-XAI. As too many terms are contained in the F-SRP-XAI while it takes a similar form to T-SRP-XAI, we omit its concrete expression.

Compared to the T-SRP-XAI, more detailed findings can be drawn from the F-SRP-XAI. Regarding ship type, it is clearly shown that if a ship is of other type, bulk carrier, and gas carrier, the predicted deficiency number will be increased by 1.63, 0.44, and 0.17 respectively from the base value. Meanwhile, being an oil tanker, chemical tanker, container ship, and passenger ship would decrease the number of deficiencies from the base value. Regarding flag performance, only when a ship has its flag on the white-list can its deficiency number decrease from the base value. In contrast, flag on the grey-list and black-list would increase the predicted deficiency number by 0.29 and 1.12 from the base value. When the flag is not defined on the list, which means that the vessels flying this flag are involved in less than 30 PSC inspections over the previous 3-year period, the predicted deficiency number would be increased by1.08 from the base value. Similar to the effects of flag performance, as the performance of ship management company gets worse from high to very low, -0.54 to 1.65 more deficiencies will be detected. If the company performance is unknown (i.e., with no inspection within previous 36 months), 1.48 more deficiencies will be detected. The situation of RO performance is a little different: when it is not defined by the Tokyo MoU, which means that it has less than 60 inspections in the previous three years, the predicted deficiency number would be slightly decreased by 0.10 from the base value. Similarly, high RO performance would decrease the deficiency number by 0.03 while medium RO performance would increase the deficiency number by 1.59 from the base value. If a vessel is not inspected within the previous 36 months, 0.14 more deficiency is likely to be detected compared to the base value.

The fitting curves of the continuous features in the F-SRP-XAI show that as the age of a ship increases by one year and one more deficiency is detected on average in the previous 36 months, 0.12 and 0.42 more deficiency will be detected compared to the base value, respectively. Meanwhile, when the squared root of the detention rate within the previous 36 months increases by 1, 1.99 more deficiencies will be detected. Therefore, the policy implication generated from the F-SRP-XAI is that the port states should pay more attention to ships of other type, with flag on black-list or not defined,

with medium performance RO, with company performance low, very low, or undefined, as well as ships of older age and worse performance in recent inspections.

The above analysis shows that like the T-SRP-XAI, the F-SRP-XAI makes the whole prediction process and the internal working mechanism of the black-box F-SRP framework transparent, while the decision process is also verified to comply with shipping domain knowledge. Furthermore, as the features are more finely processed, their effects on the final prediction can be described in a more detailed manner, which can guide the ship selection procedure more effectively. However, one drawback of the F-SRP-XAI is that as it has much more features than the T-SRP-XAI, the interpretation of the F-SRP-XAI is more complicated, and thus it is not as intuitive as the F-SRP-XAI.

We then apply the F-SRP-XAI to predict the deficiency number of the samples in the test set. The MSE, RMSE, and MAE on the test is 18.4691, 4.2976, and 2.7408. Although the performance of the F-SRP-XAI is slightly worse than the original F-SRP framework, whose MSE, RMSE, and MAE are 17.2895, 4.1581, and 2.6478, it can fully disclose the decision process of the black-box F-SRP in a white-box manner, with the MSE and MAE between the prediction of F-SRP and the prediction of F-SRP-XAI as 0.8661 and 0.5427, respectively.

6.3.4 Comparison of the SRP and the global surrogate models

We compare the performance of the SRP, T-SRP-XAI, and F-SRP-XAI regarding the number of deficiencies detected and the ship detentions identified on the test set. The results are summarized in Table 13.

Table 13. Comparison results of SRP, T-SRP-XAI, and F-SRP-XAI

Comparison scheme	Improvement regarding the number of deficiencies detected represented by ratio (a total of 2,992 deficiencies)		Improvement regarding the number of detentions detected represented by absolute value (a total of 23 detentions)
Scheme I			
T-SRP-XAI over SRP	65.19%		9.27
F-SRP-XAI over SRP	63.64%		9.13
F-SRP-XAI over T-SRP-XAI	0.79%		–0.14
Scheme II			
T-SRP-XAI over SRP	17.46%		3.35
F-SRP-XAI over SRP	17.61%		3.37
F-SRP-XAI over T-SRP-XAI	0.28%		0.01
Scheme III	**SRP**	**T-SRP-XAI**	**F-SRP-XAI**
Average no. of deficiencies among 'HRS'	6.3867	6.8207	7.1822
Average no. of deficiencies among 'SRS'	3.3264	3.1610	3.0148
Average no. of deficiencies among 'LRS'	2.5087	2.2665	2.0809
Average no. of detentions among 'HRS'	0.0711	0.0911	0.0933
Average no. of detentions among 'SRS'	0.0208	0.0045	0.0030
Average no. of detentions among 'LRS'	0	0.0058	0.0058

Tables 6 and 13 indicate that the difference in the performance between the near linear form global surrogate model and the corresponding black-box model regarding the number of deficiencies and detentions detected is minor, even though the original black-box model is more accurate than its global surrogate model. It is also shown that the F-SRP-XAI is a little more efficient than the T-SRP-XAI regarding the number of deficiencies detected in Schemes I and II, while the F-SRP-XAI performs slightly worse than the T-SRP-XAI regarding the detentions detected in Scheme I. Regarding the deficiency and detention conditions of the ships in three risk levels, results of comparison Scheme III show that F-SRP-XAI can identify ships in 'HRS' the most efficiently as evaluated by both deficiency and detention conditions. Furthermore, both global surrogate models perform similarly to their original black-box models. Based on the above findings, it can be concluded that the near linear-form global surrogate models are almost as efficient as their original black-box models regarding the ability to identify high-risk ships, although their accuracy is slightly worse than the original models. Therefore, it is justifiable to go one step further to extend the local SHAP method to a near linear-form global surrogate model.

7. Discussion

Several managerial implications and insights can be generated from the prediction models developed and the prediction results obtained. First, the extensive numerical experiments have clearly shown that the data-driven ML models are more efficient than the current method based on domain knowledge for ship risk prediction and high-risk ship selection. Furthermore, even if the same set of features ~~are~~ is used, finer grained feature encoding can improve model performance. Therefore, decision makers in port authorities can rely more on the recommendations given by these data-driven models when deciding which ships should be inspected. In particular, as it is also shown that the global surrogate model taking a near linear form can almost mimic the performance of the original black-box model, decision makers can also verify the reliability and rationality of the original black-box models, and then safely trust and rely on them more if ~~they~~ these models are consistent with their expert knowledge. Second, as can be ~~seem~~ seen from the parameters in the global surrogate models, ~~a~~the same feature value can have different effects on the risk level for different ships. When identifying ships with a higher risk, decision makers should pay more attention to ships with worse management organizations (especially flags and companies), followed by worse historical inspection performance, and then older ships.

Data property plays an important role in the development of efficient and explainable data-driven models for ship selection in PSC. Especially, data quality is the most important factor that determines model prediction performance and explanation reliability, as they are 没看懂“数据质量是最重要的决定模型性能和解释性的，因为性能和解释性是模型最根本的。这里的因果关系没看懂” the underlying of data-driven model development. In addition, data should be diverse, i.e., features from comprehensive aspects for the prediction of the target should be collected from various sources, as data diversity determines the prediction accuracy and can help to alleviate the problem of over-fitting. Another important factor is data quantity, which is also a determinant of model performance and ~~high-quality~~more voluminous data can improve model generalization ability. If model explanation should be provided, data quantity becomes more important, as non-representative or even inaccurate explanations might be given if too few (and biased) data are used.

In addition to the prediction of ship risk level in PSC inspection, the proposed framework involving accurate prediction and efficient explanation can also be used to

address other predictive problems in maritime transport where explanations for model working mechanism and prediction results are needed, such as ship trajectory prediction, ship energy efficiency prediction, ocean freight market condition prediction, and ship destination and arrival time prediction~~, etc.~~, in addition to other typical examples discussed in Yan et al.~~,~~ (2021c). Furthermore, it can also be applied to address prediction problems in other transportation modes and the entire supply chain, where typical applications can be found in Simroth and Zähle (2010), Ilie-Zudor et al. (2015), and Iovan (2017).

There are some limitations in the ~~current~~ data-driven ship risk prediction framework~~s~~ proposed in this study. For example, we only use the inspection records at the Hong Kong port in the case dataset. In future research, inspection records from more ports in the Tokyo MoU or even in different MoUs can be used to improve the applicability and robustness of the prediction model. In addition, we only use three types of curves to fit the values of continuous features and their corresponding SHAP values and the most suitable one is selected largely by experience. In future study, more types of curves should be fit, and the most suitable curve~~s~~ should be selected in a more systematic way (e.g., using cross validation or a hold-out validation set).

8. Conclusion

PSC is an effective safety net to catch substandard ships to enhance maritime safety, protect the marine environment, and guarantee seafarers' rights. To identify high-risk ships more efficiently, two ship risk prediction frameworks based on state-of-the-art GBRTs, namely T-SRP and F-SRP, are developed and validated in this study using six years' inspection records at the Hong Kong Port. To make the new frameworks more comprehensible and acceptable by the port authority part and the ship part, features used and their processing methods in the T-SRP are the same as those in the SRP, while features in the F-SRP are the same as those in the SRP but with more sophisticated processing. In addition, predictions given by the black-box models are thoroughly explained from both local and global perspectives using the post-hoc, model-agnostic, and local SHAP method by explaining the prediction of individual ships, calculating global feature importance scores, and formulating white-box global surrogate models in near linear form of the original ML models denoted by T-SRP-XAI

(for T-SRP) and F-SRP-XAI (for F-SRP). The analysis of model explanations is given, and policy implications are drawn from various perspectives.

Comprehensive numerical experiments show that the predictions given by the T-SRP and the F-SRP are accurate. When applying them to predict ship risk and identify high-risk ships, more than 60% more deficiencies and nearly 40% more detentions can be detected by both new frameworks when ignoring ship inspection priority compared to the current SRP. When the inspection priority is considered, nearly 20% more deficiencies and over 10% more detentions can be detected compared to the SRP. In both cases, the F-SRP has better performance than the T-SRP. The new frameworks are also more efficient in identifying the type of HRS ship compared to the SRP. Meanwhile, their while-box global surrogate models taking a near linear form follow the PDR model explanation evaluation framework and can provide accurate and comprehensive explanations to decisions makers and practitioners in the shipping industry, so as to enhance their applicability to the conservative maritime transport area.

Our main contributions and novelties are summarized as follows. First, to the best of the authors' knowledge, we make the first few attempts in the maritime transport area to open up a black-box data-driven model based on ML techniques for high-risk ship selection in PSC inspection. Second, we extend the local SHAP method to a global explanation method in an intuitive and succinct way by deriving a near linear-form global surrogate model. We also demonstrate that the performance of the global explanation method based on SHAP is almost as accurate as the original black-box prediction model. Third, comprehensive and consistent explanations provided in this study can shed light on future policy and decision making in ship selection for PSC, which is one of the most important international marine policies. It can help to fulfill the IMO's goal of realizing 'safe, secure and efficient shipping on clean oceans'.

References

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE Access 6, 5213852160.

Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila R. Herrera, F., 2020. Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82–115.

Asadabadi, A., Miller-Hooks, E., 2020. Maritime port network resiliency and reliability through co-opetition. Transportation Research Part E: Logistics and Transportation Review 137, 101916.

Athey, S., 2017. Beyond prediction: Using big data for policy problems. Science 355(6324), 483–485.

Babic, B., Gerke, S., Evgeniou, T., Cohen, I., 2021. Beware explanations from AI in health care. Science 373(6552), 284–286.

Barredo-Arrieta, A., Laña, I., Del Ser, J., 2019. What lies beneath: a note on the explainability of black-box machine learning models for road traffic forecasting. In Proceedings of 2019 IEEE Intelligent Transportation Systems Conference, 2232–2237.

Bukhsh, Z., Saeed, A., Stipanovic, I., Doree, A., 2019. Predictive maintenance using tree-based classification techniques: a case of railway switches. Transportation Research Part C: Emerging Technologies 101, 35–54.

Burkart, N., Huber, M., 2021. A survey on the explainability of supervised machine learning. Journal of Artificial Intelligence Research 70, 245–317.

Chen, X., Zahiri, M., Zhang, S., 2017. Understanding ridesplitting behavior of on-demand ride services: an ensemble learning approach. Transportation Research Part C: Emerging Technologies 76, 51–70.

Degré, T., 2007. The use of risk concept to characterize and select high risk vessels for ship inspections. WMU Journal of Maritime Affairs 6(1), 37–49.

Degré, T., 2008. From black-grey-white detention-based lists of flags to black-grey-white casualty-based lists of categories of vessels? The Journal of Navigation 61(3), 485–497.

Dinis, D., Teixeira, A., Soares, C., 2020. Probabilistic approach for characterising the static risk of ships using Bayesian networks. Reliability Engineering & System Safety 203, 107073.

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

Du, M., Liu, N., Hu, X., 2019. Techniques for interpretable machine learning. Communications of the ACM 63(1), 68–77.

Friedman, J., Hastie, T., and Tibshirani, R., 2001. The Elements of Statistical Learning. Berlin: Springer Publisher.

Gao, Z., Lu, G., Liu, M., Cui, M., 2008. A novel risk assessment system for port state control inspection. In Proceedings of 2008 IEEE International Conference on Intelligence and Security Informatics, 242–244.

Guo, Y., Lu, Y., Liu, R. W., 2022. Lightweight deep network-enabled real-time low-visibility enhancement for promoting vessel detection in maritime video surveillance. The Journal of Navigation 75(1), 230–250.

Ilie-Zudor, E., Ekárt, A., Kemeny, Z., Buckingham, C., Welch, P. and Monostori, L., 2015. Advanced predictive-analysis-based decision support for collaborative logistics networks. Supply Chain Management 20(4), 369–388.

Iovan, S., 2017. Predictive analytics for transportation industry. Journal of Information Systems & Operations Management 11(1), 1–14.

Hagenauer, J., Helbich, M., 2017. A comparative study of machine learning classifiers for modeling travel mode choice. Expert Systems with Applications 78, 273–282.

Heij, C., Knapp, S., 2019. Shipping inspections, detentions, and incidents: an empirical analysis of risk dimensions. Maritime Policy & Management 46(7), 866–883.

Hellsten, E., Koza, D. F., Contreras, I., Cordeau, J. F., Pisinger, D., 2021. The transit time constrained fixed charge multi-commodity network design problem. Computers & Operations Research 136, 105511.

Kalatian, A., Farooq, B., 2021. Decoding pedestrian and automated vehicle interactions using immersive virtual reality and interpretable deep learning. Transportation Research Part C: Emerging Technologies 124, 102962.

Karsten, C. V., Brouer, B. D., Desaulniers, G., Pisinger, D., 2017. Time constrained liner shipping network design. Transportation Research Part E: Logistics and Transportation Review 105, 152–162.

Khoda Bakhshi, A., Ahmed, M., 2021. Utilizing black-box visualization tools to interpret non-parametric real-time risk assessment models. Transportmetrica A: Transport Science 17(4), 739–765.

Kim, E., 2021. Analysis of travel mode choice in Seoul using an interpretable machine learning approach. Journal of Advanced Transportation, in press.

Kim, E., Kim, Y., Kim, D., 2021. Interpretable machine-learning models for estimating trip purpose in smart card data. In Proceedings of the Institution of Civil Engineers-Municipal Engineer 174(2), 108–117.

Kim, T., Sharda, S., Zhou, X., Pendyala, R., 2020. A stepwise interpretable machine learning framework using linear regression (LR) and long short-term memory (LSTM): City-wide demand-side prediction of yellow taxi and for-hire vehicle (FHV) service. Transportation Research Part C: Emerging Technologies 120, 102786.

Kleinberg, J., Ludwig, J., Mullainathan, S., Obermeyer, Z., 2015. Prediction policy problems. American Economic Review 105(5), 491–95.

Knapp, S., Heij, C., 2020. Improved strategies for the maritime industry to target vessels for inspection and to select inspection priority areas. Safety 6(2), 18–39.

Li, Q., Garg, S., Nie, J., Li, X., Liu, R. W., Cao, Z., Hossain, M. S., 2020. A highly efficient vehicle taillight detection approach based on deep learning. IEEE ~~t~~Transactions on I~~i~~ntelligent ~~t~~Transportation S~~s~~ystems 22(7), 4716–4726.

Li, C., Xie, Y., Wang, G., Zeng, X., Jing, H., 2021) Lateral stability regulation of intelligent electric vehicle based on model predictive control. Journal of Intelligent and Connected Vehicles 4(3), 104–114.

Liu, R. W., Liang, M., Nie, J., Lim, W. Y. B., Zhang, Y., Guizani, M., 2022. Deep learning-powered vessel trajectory prediction for improving s~~S~~mart traffic services in maritime internet of things. IEEE Transactions on Network Science and Engineering, ~~early access~~in press.

Lundberg, S., Lee, S., 2017. A unified approach to interpreting model predictions. In Proceedings of the 31st International Conference on Neural Information Processing Systems, 4768–4777.

Lundberg, S., Erion, G., Lee, S., 2019. Consistent individualized feature attribution for tree ensembles. arXiv preprint arXiv:1802.03888.

Lundberg, S., 2021. API Reference. Accessed 10 May 2021. https://shap-lrjball.readthedocs.io/en/latest/api.html.

Lundberg, S., Erion, G., Chen, H., DeGrave, A., Prutkin, J., Nair, B., Katz, R., Himmelfarb, J., Bansal, N., Lee, S., 2020. From local explanations to global

understanding with explainable AI for trees. Nature Machine Intelligence 2(1), 56–67.

Marine Department, 2021. PSC Information Related to Hong Kong Ships. Accessed 10 August 2021. https://www.mardep.gov.hk/en/faq/pscinfo.html.

Mokhtari, K., Rahman, N. S. F. A., Soltani, H. R., Al Rashdi, S. A., Al Balushi, K. A. A. M., 2021. Security risk management: a case of Qalhat liquefied natural gas terminal. Maritime Business Review 6(4), 318–338.

Molnar, C., 2020. Interpretable Machine Learning. Accessed 10 May 2021. https://christophm.github.io/interpretable-ml-book/.

Murdoch, W., Singh, C., Kumbier, K., Abbasi-Asl, R., Yu, B., 2019. Definitions, methods, and applications in interpretable machine learning. Proceedings of the National Academy of Sciences 116(44), 22071–22080.

Parmar, J., Das, P., Dave, S., 2021. A machine learning approach for modelling parking duration in urban land-use. Physica A: Statistical Mechanics and its Applications 572, 125873.

Qi, J., Wang, S., Psaraftis, H., 2021. Bi-level optimization model applications in managing air emissions from ships: A review. Communications in Transportation Research, 1, 100020.

Shapely, L., 1953. A value for n-person games. Contributions to the theory of games. Annals of Mathematics Studies (2), 307–318.

Simroth, A., Zähle, H., 2010. Travel time prediction using floating car data applied to logistics planning. IEEE Transactions on Intelligent Transportation Systems 12(1), 243–253.

Tokyo MoU, 2013. Information sheet of the new inspection region (NIR). Accessed 15 November 2018. http://www.tokyo-mou.org/doc/NIR-information%20sheet-r.pdf.

Tokyo MoU, 2016. Information on fees and charges by authorities for follow-up PSC inspection. Accessed 17 August 2021. http://www.tokyo-mou.org/doc/INFORMATION%20FOR%20PSC%20INSPECTION%20FEES%20AND%20CHARGES%20BY%20AUTHORITIES-ver16-r.pdf.

Tokyo MoU, 2020. Annual report on port state control in the Asia-Pacific region 2019. Accessed 17 July 2020. http://www.tokyo-mou.org/doc/ANN19-f.pdf.

Trivella, A., Corman, F., Koza, D. F., Pisinger, D., 2021. The multi-commodity network flow problem with soft transit time constraints: Application to liner

shipping. Transportation Research Part E: Logistics and Transportation Review 150, 102342.

Tseng, P. H., Ng, M.W., 2020. Assessment of port environmental protection in Taiwan. Maritime Business Review 6(2), 188–203.

UNCTAD, 2021. COVID-19 and maritime transport: Impact and responses. Accessed 10 April 2021. https://unctad.org/webflyer/covid-19-and-maritime-transport-impact-and-responses.

Veran, T., Portier, P., Fouquet, F., 2020. Crash prediction for a French highway network with an XAI-informed Bayesian hierarchical model. In Proceedings of 2020 IEEE International Conference on Big Data,1256–1265.

Xiao, Z., Fu, X., Zhang, L., Zhang, W., Liu, R. W., Liu, Z., Goh, R. S. M., 2022. Big data driven vessel trajectory and navigating state prediction with adaptive learning, motion modeling and particle filtering techniques. IEEE Transactions on Intelligent Transportation Systems 23 (4), 3696–3709.

Wang, S., Wang, Q., Zhao, J., 2020. Deep neural networks for choice analysis: Extracting complete economic information for interpretation. Transportation Research Part C: Emerging Technologies 118, 102701.

Wang, S., Wang, Q., Bailey, N., Zhao, J., 2021a. Deep neural networks for choice analysis: a statistical learning theory perspective. Transportation Research Part B: Methodological 148, 60–81.

Wang, S., Mo, B., Zhao, J., 2021b. Theory-based residual neural networks: a synergy of discrete choice models and deep neural networks. Transportation Research Part B: Methodological 146, 333–358.

Wang, S., Psaraftis, H.N., Qi, J., 2021c. Paradox of international maritime organization's carbon intensity indicator. Communications in Transportation Research, 1, 100005

Wang, S., Yan, R., Qu, X., 2019. Development of a non-parametric classifier: Effective identification, algorithm, and applications in port state control for maritime transportation. Transportation Research Part B: Methodological 128, 129–157.

Wang, X., Xiang, Y., Niu, W., Tong, E., Liu, J., 2020, December. Explainable congestion attack prediction and software-level reinforcement in intelligent traffic signal system. In Proceedings of 2020 IEEE 26th International Conference on Parallel and Distributed Systems, 667–672.

Wu, S., Chen, X., Shi, C., Fu, J., Yan, Y., Wang, S., 2021. Ship detention prediction via feature selection scheme and support vector machine (SVM). Maritime Policy & Management, in press.

Xiao, Y., Wang, G., Lin, K. C., Qi, G., Li, K., 2020. The effectiveness of the new inspection regime for port state control: application of the Tokyo MoU. Marine Policy 115, 103857.

Xu, Y., Yan, X., Liu, X., Zhao, X., 2021. Identifying key factors associated with ridesplitting adoption rate and modeling their nonlinear relationships. Transportation Research Part A: Policy and Practice 144, 170–188.

Xu, R., Lu, Q., Li, W., Li, K. X., Zheng, H., 2007a. A risk assessment system for improving port state control inspection. In Proceedings of 2007 International Conference on Machine Learning and Cybernetics, 818–823.

Xu, R., Lu, Q., Li, K. X., Li, W., 2007b. Web mining for improving risk assessment in port state control inspection. In Proceedings of 2007 International Conference on Natural Language Processing and Knowledge Engineering, 427–434.

Yan, R., Wang, S., 2019. Ship inspection by port state control—Review of current research. Smart Transportation Systems 2019, 233–241.

Yan, R., Wang, S., Fagerholt, K., 2020. A semi-"smart predict then optimize"(semi-SPO) method for efficient ship inspection. Transportation Research Part B: Methodological 142, 100–125.

Yan, R., Wang, S., Cao, J., Sun, D., 2021a. Shipping domain knowledge informed prediction and optimization in port state control. Transportation Research Part B: Methodological 149, 52–78.

Yan, R., Wang, S., Peng, C., 2021b. An artificial intelligence model considering data imbalance for ship selection in port state control based on detention probabilities. Journal of Computational Science 48, 101257.

Yan, R., Wang, S., Zhen, L., Laporte, G., 2021c. Emerging approaches applied to maritime transport research: Past and future. Communications in Transportation Research 1, 100011.

Yang, Z., Yang, Z., Yin, J., 2018a. Realising advanced risk-based port state control inspection using data-driven Bayesian networks. Transportation Research Part A: Policy and Practice 110, 38–56.

Yang, Z., Yang, Z., Yin, J., Qu, Z., 2018b. A risk-based game model for rational inspections in port state control. Transportation Research Part E 118, 477–495.

Yang, Z., Wan, C., Yang, Z., Yu, Q., 2021. Using Bayesian network-based TOPSIS to aid dynamic port state control detention risk control decision. Reliability Engineering & System Safety 213, 107784.

Yi, W., Wang, H., Jin, Y., Cao, J., 2021. Integrated computer vision algorithms and drone scheduling. Communications in Transportation Research, 1, 100002.

Yi, W., Wang, W., Hu, Y., Li, M., Zhen, L., 2019. Collaborative stowage planning problem for a liner ship. International Journal of Shipping and Transport Logistics, 11(2-3), 176–195.

Zhang, Y., Haghani, A., 2015. A gradient boosting method to improve travel time prediction. Transportation Research Part C: Emerging Technologies 58, 308–324.

Zhao, X., Yan, X., Yu, A., Van Hentenryck, P., 2018. Modeling stated preference for mobility-on-demand transit: a comparison of machine learning and logit models. arXiv preprint arXiv:1811.01315.

Zhao, X., Yan, X., Van Hentenryck, P., 2019. Modeling heterogeneity in mode-switching behavior under a mobility-on-demand transit system: an interpretable machine learning approach. arXiv preprint arXiv:1902.02904.

Zhou, C., Xu, J., Miller-Hooks, E., Zhou, W., Chen, C. H., Lee, L. H., Chew, P., Li, H., 2021. Analytics with digital-twinning: A decision support system for maintaining a resilient port. Decision Support Systems 143, 113496.

Zhu, W., Wu, J., Fu, T., Wang, J., Zhang, J., Shangguan, Q., 2021. Dynamic prediction of traffic incident duration on urban expressways: a deep learning approach based on LSTM and MLP. Journal of Intelligent and Connected Vehicles 4(2), 80–91.

Zisi, V., Psaraftis, H. N., Zis, T., 2021. The impact of the 2020 global sulfur cap on maritime CO_2 emissions. Maritime Business Review 6(4), 339–357.

Appendix A. Summary of literature on ship risk prediction

Table A.1 Summary of studies on ship risk prediction for PSC inspection

Literature	Dataset	Features considered	Risk indicator	Risk prediction model	Explainability
Xu et al. (2007a)	5,000 ships with more than 4 inspection records in the Paris MoU from January 2003 to January 2007	Generic factors: ship age, type, tonnage, flag, classification society, company, History factors: the number of deficiencies, outstanding deficiencies, duplicate deficiencies, duplicate outstanding deficiencies, and detentions in past 4 inspections, time since last initial inspection	Ship detention	SVM	No
Xu et al. (2007b)	The same as Xu et al. (2007a)	Numbers of non-lasting and lasting equipment/operation deficiencies, number of outstanding non-lasting and lasting equipment/operation deficiencies, numbers of deficiencies/outstanding deficiencies in areas 1 to 8 in past 4 inspections and the features considered by Xu et al. (2007a)	Ship detention	SVM	No
Gao et al. (2007)	140,000 inspection records in the Tokyo MoU	15 features including ship generic factors, dynamic factors, and history factors	Ship detention	KNN-SVM	No
Wu et al. (2021)	Inspection records of general cargo ship from 2014 to 2018 in the Tokyo MoU	Ship age, number of deficiencies, and 5 types of deficiencies selected by AHP and GRA	Ship detention	SVM	No
Yang et al. (2018a)	Inspection records of bulk carriers from 2005 to 2008 in the Paris MoU	Ship flag, RO, deadweight tonnage, age, inspection type, inspection port, and the number of deficiencies detected	Ship detention	BN model	Partially explainable, presented by conditional probability
Yang et al. (2018b)	Inspection records of bulk carriers from 2015 to 2017 in the Paris MoU	Ship flag, age, company performance, inspection type, inspection port, inspection date, and the number of deficiencies detected	Ship detention	BN model	Partially explainable, presented by conditional probability
Wang et al. (2019)	Inspection records in 2017 at the Hong Kong Port in the Tokyo MoU	Ship age, gross tonnage (GT), type, flag performance, company performance, RO performance, last inspection time, the number of deficiencies in last inspection, the number of previous detentions, and the number of times of changing flag	Ship deficiency number	BN model	Partially explainable, presented by conditional probability
Yan et al. (2020)	Inspection records from 2016 to 2018 at the Hong Kong Port in the Tokyo MoU	Ship age, GT, length, depth, beam, type, the number of times of changing flag, total detention times, casualties in last five years, ship flag, RO, and company performance, last inspection time, last deficiency number, follow-up inspection rate	Ship deficiency number under each deficiency category	RF models consisting of multi-target regression trees	No
Yan et al. (2021)	Inspection records from 2016 to 2018 at the Hong Kong Port in the Tokyo MoU	Ship age, GT, type, depth, length, beam, the number of times of changing flag, casualties in the last 5 years, total detentions, ship flag, RO, and company performance, last inspection time, last deficiency number, and follow-up inspection rate	Ship detention	BRF model	No
Yan et al. (2021c)	Inspection records from 2016 to 2018 at the Hong Kong Port in the Tokyo MoU	Ship age, GT, length, depth, beam, type, ship flag, RO, and company performance, last inspection date, last deficiency number, total detentions, the number of flag changes, and casualty in last 5 years	Ship deficiency number	XGBoost model	No
Degré (2007)	IMO casualty records from 1998 to 2003	Ship type, size, and age	Ship risk evaluated by the probability of the occurrence of casualties and their potential consequences	A statistical model	Yes
Degré (2008)	Casualty descriptive statistics and world merchant fleet descriptive statistics	Ship type, size, and age	Black-grey-white lists of categories of ships	A binomial calculation method	Yes

Heij and Knapp (2019)	IHS Markit for ship-particular data, ship incident database from 2010 to 2014, and ship inspection database from 2010 to 2014	A total of thirty factors with more than 500 variables, such as flag, owner, engine designer and builder are contained in the initial model, while only significant variables are contained in sub-models	Ship inspections, detentions, and very serious and serious incidents	A logit model	Yes
Knapp and Heij (2020)	IHS Markit for ship-particular data, ship incident database from 2010 to 2014, and ship inspection database from 2010 to 2014 for estimating risk formulas and probabilities, and quarterly data of incidents, inspection and ship particular data for estimating probabilities	Over 500 variables are contained in the initial model, and 16 to 172 variables are contained in the sub-models	Ship inspections, detentions, and very serious and serious incidents	A combination of logit model and percentage rank model	Yes
Dinis et al. (2020)	Inspection records of 136 ships at the port of Lisbon in the Paris MoU in 2018, and AIS data of 25 ships that have entered the same port	Ship type, age, flag, RO, company, deficiency and detention within the last 3 years	Ship risk profile with more detailed states	BN	Partially explainable, presented by conditional probability