https://doi.org/10.1016/j.jocs.2020.101257

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

An Artificial Intelligence Model Considering Data Imbalance for Ship Selection in Port State Control Based on Detention Probabilities

Abstract

Port state control inspection is seen as the safety net to guard marine safety, protect the marine environment, and guarantee decent onboard working and living conditions for seafarers. A substandard ship can be detained in an inspection if serious deficiencies are found onboard. Ship detention is regarded to be a severe result in port state control inspection, however, developing accurate prediction models for ship detention based on ship's generic factors (e.g. ship age, ship type, and ship flag), dynamic factors (e.g. times of changing ship flag), and inspection historical factors (e.g. total previous detentions in PSC inspection, last PSC inspection time, and last deficiency number in PSC inspection) before an inspection is conducted is not a trivial task as the low detention rate leads to a highly imbalanced inspection records dataset. To address this issue, this paper develops a classification model called balanced random forest (BRF) to predict ship detention by using 1,600 inspection records at the Hong Kong port for three years. Numerical experiments show that the proposed BRF model can identify 81.25% of all the ships with detention in the test set which contains another 400 inspection records. Compared with the currently used ship selection method at the Hong Kong port, the BRF model is much more efficient and can achieve an average improvement of 73.72% for the 400 ships in the test set.

Keywords

Port state control inspection, ship detention, machine learning in maritime transportation, artificial intelligence in maritime transportation, imbalanced data

1. Introduction

Maritime transport is the backbone of global trade (Rekik and Elkosantini, 2019; Abbassi et al., 2019). Although maritime transport is relatively safe, accidents and casualties involving maritime vessels can bring about huge losses to the shipping industry and the whole society. As reported by European Maritime Safety Agency (EMSA, 2019), from 2011 to 2018, 25,614 ships were involved in marine casualties, causing 7,694 persons injured and 696 fatalities. Meanwhile, as the vessels are mainly powered by heavy fuel oil, heavy environmental footprint is created by emission of greenhouse gas and pollutants from vessels (Ramdin et al., 2016). To guarantee maritime safety and protect the marine environment, numerous international regulations and conventions are proposed and implemented by the International Maritime Organization (IMO), such as the International Convention for the Safety of Life at Sea (SOLAS) and the International Convention for the Prevention of Pollution from Ships (MARPOL). In recent years, onboard living and working conditions of the crew members have gained much attention from the International Labour Organization (ILO), and international agreements such as Maritime Labour Convention (MLC) was implemented to set out seafarers' rights to decent conditions of work. These international instruments provide comprehensive standards serving as the base for regulating the design, manning, equipment, operation, management, maintenance, and disposal of ships (Graziano et al., 2018).

Ships with hull, machinery, equipment or operational safety substantially below the international standards or whose crew is not in conformance with the safe manning document are called "substandard ships" (IMO, 2017).The flag state of a ship, which is the jurisdiction under whose laws the ship is registered and licensed and is deemed as the nationality of the ship, is the first line of defense against substandard shipping. However, it is widely believed that the flag states cannot perform their duty well (Cariou et al., 2008; Li et al., 2014; Wang et al., 2019; Yan et al., 2020). Under this circumstance, port state control (PSC), which is an inspection to verify that the foreign visiting ships are manned and operated in compliance with the international rules, is established as a complement to flag state control and it is regarded as the second line of defense against substandard shipping (Cariou et al., 2009; Heij et al., 2011). To allow information exchange and avoid multiple inspections in a certain region over a period of time, as well as to standardize inspection criteria and processes, the regional

48 Memorandum of Understandings on port state control (i.e. MoUs on PSC) are signed and established. As of 1
49 July 2020, nine MoUs on PSC have been established all over the world.

50 Before the ships coming to the port state, the PSC officers (PSCOs) would first select the ships with higher
51 risk to inspect. Different decision support systems are used in different MoUs to target high-risk ships. For
52 example, the Paris MoU and Tokyo MoU adopt ship risk profile (SRP) to select ships that are more likely to
53 have larger number of deficiencies and to be detained (Paris MoU, 2014; Tokyo MoU, 2014). The results of an
54 inspection mainly contain identified deficiencies and ship detention (IMO, 2017). A ship deficiency is a
55 condition found not to be in compliance with the requirements of the relevant convention, whereas ship
56 detention is an intervention action taken by the port state when the ship is unseaworthy (IMO, 2017). Ship
57 deficiencies are clearly classified and listed by the MoUs. For example, 17 deficiency codes are listed by Tokyo
58 MoU regarding ship safety, management, condition and structure, and communication and navigation (Tokyo
59 MoU, 2017b). If serious deficiencies which make the ship unsafe to sail at sea are identified, the PSCO can
60 detain the ship and require the ship to rectify the deficiencies before departing. Ship detention is the most
61 important decision generated during an inspection and can be regarded as the most severe result of PSC
62 inspection. Ship detention not only indicates poor ship condition and higher probability of involvement in future
63 incidents and accidents but may also delay ship schedule. Besides, ship detention can adversely influence the
64 reputation of its flag state, recognized organization, and company and thus can lead to higher inspection rate of
65 their ships. Generally, ship detention rate is low in Tokyo MoU. The years from 2009 to 2018 have witnessed a
66 decrease in the detention rate in the Tokyo MoU from 5.78% to 2.96%. In 2019, there is a slight increase in ship
67 detention rate to 3.13% (Tokyo MoU, 2020). Although ship detention is the most importance decision in PSC
68 inspection and the detention rate is usually low, several steps in a PSC inspection, ranging from ship selection
69 to ship inspection and final decision making, might lead to inaccuracy and inefficiency. First, several studies
70 have reported that the SRP ship selection scheme is inefficient for targeting high-risk ships (Xu et al., 2007a,
71 2007b; Gao et al., 2008; Wang et al., 2019). As a result, ships selected for inspection by the port states using
72 SRP are not necessarily the ones with the largest number of deficiencies and highest probability of detention. In
73 other words, ships that should be detained may be ignored in the process of ship selection. Second, even if the

74 SRP ship selection scheme can pick out the high-risk ships for inspection, no specific risk scores can be
75 generated for the ships in the same risk profile. Consequently, to what extent the ships should be inspected is
76 highly dependent on the judgement of the PSCO. For those PSCOs who are lack of expertise, they may let the
77 substandard ships go without further inspection. Third, to the best of our knowledge, no detainable deficiencies
78 are specifically illustrated by the IMO or the MoUs. Instead, only rough description of the deficiencies
79 warranting detention is given in the documents (IMO, 2017). Therefore, the decision of detention is also highly
80 dependent on the expertise and judgement of PSCOs. If too excessive PSC inspections are conducted, the
81 competitiveness of the ports is harmed and the burden of the ship owners is increased. On the contrary, a loose
82 inspection policy cannot guarantee the implementation of effective PSC inspections and thus increase the
83 possibilities of marine accidents and casualties' occurrence (Yang et al., 2018b).

84 To improve the accuracy and efficiency of PSC inspection, this paper aims to propose a ship detention
85 prediction model serving as the decision support tool for ship selection and inspection for the port states. The
86 model takes ship generic factors (i.e. ship age, gross tonnage, type, depth, length, beam, flag performance,
87 recognized organization performance, and company performance), ship dynamic factors (i.e. times of changing
88 flag and casualties in last 5 years), and ship inspection historical factors (i.e. total previous detention, last
89 inspection time, last deficiency number, and PSC follow-up inspection rate) into account to predict ship
90 detention probability. It addresses the imbalanced distribution of ships with and without detention (i.e. ships
91 without detention significantly outnumbers ships with detention) by using a revised version of random forest
92 classifier. The probability of detention of each ship can be generated and thus the inspection sequence can also
93 be given. A comparison of the working processes between the proposed model and the currently implemented
94 ship risk profile ship selection scheme which is based on expert knowledge is shown in Figure 1.

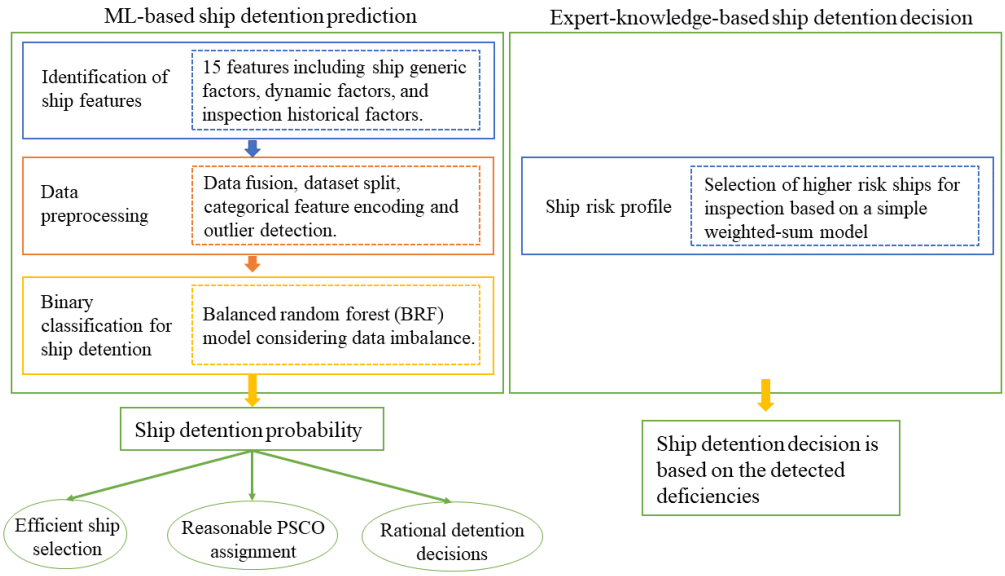

Figure 1. Comparison between the proposed and currently implemented ship detention prediction/decision models

2. Literature review

There is a large volume of published studies on PSC inspection. Several studies have investigated how to improve the efficiency and effectiveness of PSC inspection, including how to select ships for inspection (Xu et al., 2007a, 2007b; Gao et al., 2008; Degré, 2007, 2008; Yang et al., 2018a, 2018b; Wang et al., 2019; Heij and Knapp, 2019; Knapp and Heij, 2020), how to assign PSCOs considering their background and expertise and the deficiency conditions of the ships (Yan et al., 2020), and how to determine the onboard inspection sequence (Tsou, 2018; Yan et al., 2020). Besides, various studies have assessed the impact of PSC inspection on reducing incidents and accidents (Hänninen et al., 2014; Li et al., 2014; Fan et al., 2020), on future PSC inspection results (Cariou et al., 2008), on ship behavior (Cariou and Wolff, 2011; Fan et al., 2014), and on protecting the marine environment (Heij et al., 2011). In addition, several studies have provided general comments on MoU management, such as the development process of the MoUs (Mansell, 2009), the future developing directions of the MoUs (Liou et al., 2011; Grazianoa et al., 2017), and the critical challenges faced by port states and the MoUs (Grazianoa et al., 2017; Graziano et al., 2018). Factors influencing the final inspection results (i.e. identified deficiencies and detention) have also been analyzed and discussed from different perspectives (Cariou et al., 2007, 2009; Knapp and Franses, 2007; Cariou and Wolff, 2015; Ravira and Piniella, 2016; Chen et al.,

2019; Xiao et al., 2020; Şanlıer, 2020). In this section, studies on how to select ships for inspection and the factors influencing the final inspection results are reviewed as they are highly related to ship detention. For comprehensive review of the studies on PSC inspection, readers are referred to Yan and Wang (2019).

2.1 Ship selection for inspection

Before a foreign ship comes to the port state, the PSCOs need to first decide whether to inspect the ship based on its risk calculated by a decision system. In Tokyo MoU and Paris MoU, SRP ship selection scheme is used which gives different weighting points to ship generic factors and previous inspection factors. However, this simple weighted-sum method may not be efficient enough to identify the substandard ships. In this circumstance, several studies have proposed ship selection methods to improve ship selection efficiency. As the results of PSC inspection contain the identified ship deficiencies and ship detention, many studies aim to select ships with larger number of deficiencies or with higher probability to be detained from the numerous visiting ships. For studies aim to predict ship detention, Xu et al. (2007a) developed a risk assessment system based on support vector machine to identify high-risk ships that are highly likely to be detained. The performance of the system was improved by combining new input features extracted by web mining technology (Xu et al., 2007b). The system was further improved by combing support vector machine with K-nearest neighbor and bag-of-words model (Gao et al., 2008). Numerical experiments of the three studies suggested that among all the ships predicted to be detained, only about 20% of them were actually detained. Bayesian networks were developed by Yang et al. (2018a) to predict the detention probabilities of bulk carriers in Paris MoU considering ship generic factors and PSC inspection factors. Based on the model, a risk-based game model was constructed to figure out the optimal inspection policy at a certain port (Yang et al., 2018b). For studies predicting ship deficiencies, Wang et al. (2019) proposed a Bayes classifier called TAN to help the port state to select ships with large number of deficiencies for inspection. Apart from ship detention and deficiency, marine casualties are also considered for ship selection. Risk concept combining the occurrence of casualties and the potential consequences of such occurrences was proposed by Degré (2007) to select high-risk vessels for PSC inspection. In addition, a black-grey-white list of ships based on their observed casualties in a given period was proposed by Degré (2008) to help Paris MoU to target high-risk ships. Risk score of ships, which combined ship detentions

and incidents was employed by Heij and Knapp (2019) and Knapp and Heij (2020) to help port state authorities to identify high-risk ships for inspection and to guide the onboard inspection priorities.

Although ship detention is the major consequence of a PSC inspection which guarantees its effectiveness, there are few studies aiming to propose accurate prediction models for ship detention. Among all the visiting ships to the port state, ships without detention significantly outnumber ships with detention as ship detention rate is very low. According to the annual report of Tokyo MoU, 24 ships were detained among the total 716 PSC inspections conducted at the Hong Kong port in 2017 and thus the detention rate was 3.35%. The overall detention rate within Tokyo MoU in 2017 was only 2.96% (Tokyo MoU, 2018a). The imbalance between the number of ships of the two classes (i.e. with and without detention) makes the prediction problem difficult. However, the issue is not addressed in the current literature. Even if there are some machine learning models in the studies that directly make prediction on ship detention, the accuracy of identifying ships with detention was low. In addition, these studies do not generate a specific inspection sequence considering the risk level of the ships for the port state's reference.

2.2 Factors influencing PSC inspection results

The results of a PSC inspection contain ship deficiencies identified and ship detention decision. The literature on analyzing PSC inspection results has highlighted several influencing factors including ship generic factors (e.g. ship age, ship type, and ship flag) and inspection resource related factors (such as inspection authorities and background of PSCOs). Ship deficiency is mainly influenced by ship generic factors. Cariou et al. (2007) concluded that ship age, flag of registry, and ship type were the main determinants of the reported number of deficiencies in PSC inspection. As for factors influencing the decision of ship detention, both ship generic factors and the detected deficiencies can have impacts. Regarding the deficiency types that were highly likely to lead to ship detention, it was found that deficiencies on International Safety Management (ISM) incompliance, emergency system, and fire-safety measures were the main factors leading to ship detention (Chen et al., 2019). Xiao et al. (2020) developed binary logistic regression and decision tree models and suggested that ship age, type, performance of flag state, and deficiency number had significant impact on detention decision. Ship generic factors and inspection related factors can influence both detected deficiencies

and detention. Knapp and Franses (2007) reported that differences in the use of deficiencies towards detention among the port states were the main reason for different decisions on ship detention, given that basic ship profiles did not vary significantly across the regimes. Cariou et al. (2009) suggested that ship age, recognized organization, and inspection authority were the determinants of deficiency number and detention probability. Based on Probit and count data models, Cariou and Wolff (2015) reported that factors influencing the probability of detention and the number of deficiencies were quite similar. Besides, professional profiles of PSCOs might also impact the identified deficiencies and detention (Ravira and Piniella, 2016). Şanlıer (2020) also indicated that ship generic factors such as ship age, type, flag of registry, and recognized organization as well as inspection authority would influence the detected deficiencies and detention.

A summary of current literature on analyzing factors influencing PSC inspection results is presented in Table 1. Overall, these studies are mainly focused on ship generic factors and inspection resource related factors. Although ship generic factors were taken into account, factors regarding ship structure, such as ship length, ship depth, and ship beam were rarely analyzed. In addition, the influence of historical inspection factors, such as previous deficiencies, previous detentions, and follow-up inspections were rarely considered. Meanwhile, factors related to ship dynamic information, e.g. ship flag change, which may be due to the bad performance in previous PSC inspections, were seldom considered. To bridge this gap, this study aims to develop a prediction model of ship detention in PSC inspection which considers a wider range of related features by combing different databases. The proposed model considers the imbalanced distribution of the ships with and without detention and would also generate specific inspection sequence considering ship detention probability.

Table 1. Summary of literature on analyzing factors influencing PSC inspection results

Author(s)	Methods/Models	Main influencing factors	Analysis target
Cariou et al. (2007)	Count data models	Ship age, flag of registry, and ship type	Ship deficiency
Chen et al. (2019)	Grey rational analysis model	Deficiencies on International Safety Management (ISM) incompliance, emergency system, and fire-safety measures	Ship detention
Xiao et al. (2020)	Binary logistic regression model and decision tree model	Ship age, type, performance of flag state, and deficiency number	Ship detention
Knapp and Franses (2007)	Binary logistic regression	Differences in port states across several regions	Ship deficiency and detention
Cariou et al. (2009)	Probit models and count data models	Ship age, recognized organization, and inspection authority	Ship deficiency and detention
Cariou and Wolff (2015)	Probit model and count data models	Ship age, gross tonnage, type, flag of registration, recognized organization	Ship deficiency and detention
Ravira and Piniella (2016)	Survey and basic statistical analysis	professional profiles of PSCOs	Ship deficiency and detention
Şanlıer (2020)	Basic statistical analysis	Ship age, type, flag of registry, and recognized organization as well as PSC inspection authority	Ship deficiency and detention

3. Data

3.1 Introduction of imbalanced dataset

The ship detention rate in PSC inspection is usually quite low. Actually, the detention rate of our whole case dataset is only 3.55%, which means that on average, there are only 3.55 ships with detention (denoted by minority class or class "1") among 100 inspected ships while 96.45 ships are without detention (denoted by majority class or class "0"). As the majority examples significantly outnumber the minority examples, the distribution is highly imbalanced (Galar et al., 2012; Wu et al., 2020). Imbalanced class distribution in a dataset would bring about a serious difficulty in most classifier learning algorithms which assume a relatively balanced distribution (i.e. class size ratio of about 1:1) (Sun et al., 2009). Therefore, many widely-used classical classifiers, such as support vector machine, decision tree, and logistic regression may not be suitable if applied directly to the imbalanced dataset as they are highly likely to ignore examples in minority class (Galar et al., 2012).

Addressing class imbalanced problem is significant as it is present in many real-world classification problems. Generally, there are three main approaches to learning from imbalanced data: (a) Data-level methods, which aims to create balanced distribution by modifying the data collection methods (e.g. sub-sampling and

over-sampling) in training set. Standard learning algorithms are then applied to the new balanced dataset. (b) Algorithm-level methods, which modify the existing learning algorithms directly to reduce the bias towards majority examples and adapt them to data mining with imbalanced distributions. (c) Hybrid methods, which combines methods (a) and (b).

3.2 Overview of databases and data preprocessing

In this study, PSC inspection records at the Hong Kong port from 1 Jan 2016 to 31 Dec 2018 are used to calibrate and validate the ship detention prediction model. The data are collected from two databases: the web-based Asia Pacific Computerized Information System (APCIS) provided by the Tokyo MoU[1] and Word Register of ships (WRS). APCIS provides detailed inspection records at the Hong Kong port and PSC-related information of the inspected ships within the Tokyo MoU, while WRS provides information of ship-related factors.

Based on the current literature and shipping domain knowledge, we select 15 input features that are regarded to be highly related to ship detention from the APCIS database and WRS database. The features from the two databases are combined by ship's IMO number, which is a ship unique identifier. The 15 features comprise ship generic factors (e.g. ship age, gross tonnage, length, depth, beam, and type), ship historical factors (e.g. times of changing flag and casualties in the last five years), and PSC-related factors (e.g. total detentions in previous PSC inspections, ship flag performance, recognized organization performance, and company performance evaluated by Tokyo MoU, last inspection time, last deficiency number, and follow-up inspection rate within Tokyo MoU). The prediction target is ship detention in the current inspection. The 16 variables and their explanation are shown in Table 2. After deleting the inspection records with missing values (no more than 5%), we form the whole dataset which contains 2,000 inspection records with 71 records with detention and 1,929 records without detention. We randomly divide the whole dataset into training set (80%) and test set (20%). The training set contains 1,600 records in total with 55 records with detention, and the test set contains 400 records in total with 16 records with detention. The proposed ship detention prediction model is constructed

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

using the training set and its performance is validated by the test set.

Table 2. Variables in the model

Variable	Explanation
(a) Detention (prediction target)	Whether a ship is detained (set the state as "1") or not detained (set the state as "0") in the current inspection. The detention rate over the whole dataset is 3.55%.
(b) Age	The time interval (in years) between the keel laid date and the current PSC inspection date.
(c) GT (Gross tonnage)	Ship GT is a nonlinear measure of a ship's internal volume, with 100 cubic feet as the unit.
(d) Type	Ship type (container ship, general cargo/multipurpose, bulk carrier, passenger ship, tanker, and other).
(e) Depth	The vertical distance (in meters) measured from the top of the keel to the underside of the upper deck at side.
(f) Length	The overall maximum length of a ship (in meters).
(g) Beam	The width of the hull (in meters).
(h) Flag-changing-times	The total times of ship flag changing from keel laid date to the current PSC inspection date.
(i) Casualties-in-the-last-five years	Whether the ship is involved in casualties or not in the last five years.
(j) Total-detentions	Total detentions of the ship in all previous PSC inspections.
(k) Ship-flag-performance	Ship flag performance is calculated based on the flag Black-Grey-White list provided by Tokyo MoU (Tokyo MoU, 2018b). It gets worse from "white", "grey" to "black". If the flag is not listed, the value for this variable is set to be "not listed".
(l) Ship-RO (recognized organization)-performance	Ship RO performance is calculated based on RO performance list provided by Tokyo MoU (Tokyo MoU, 2018b). It gets worse from "high", "medium", "low" to "very low". If the RO is not listed, the value for this variable is set to be "not listed".
(m) Ship-company-performance	Ship company performance is calculated based on company performance list provided by Tokyo MoU (Tokyo MoU, 2018b). It gets worse from "high", "medium", "low" to "very low". If the company is not listed, the value for this variable is set to be "not listed".
(n) Last-inspection-time	The time of last PSC inspection within Tokyo MoU (in months).
(o) Last-deficiency-number	The number of deficiencies identified in last PSC inspection within Tokyo MoU.
(p) Follow-up-inspection-rate	The total number of follow-up inspections divided by the total number of inspections within Tokyo MoU.

3.3 Feature encoding and outlier detection

Among all the 15 features considered in the detention prediction model, three features are related to historical PSC inspection: last-inspection-time, last-deficiency-number, and follow-up-inspection-rate. For the ships that are inspected for the first time, the states of the three features are not available and we set them to be "−1" in both training set and test set. Besides, there are five categorical features that need to be encoded, namely

type, casualties-in-the-last-five-years, ship-flag-performance, ship-RO-performance, and ship-company-performance. Ship type and casualties-in-the-last-five-years are nominal data and thus one-hot encoding is applied. Ship-flag-performance, ship-RO-performance, and ship-company-performance are ordinal data and thus label encoding is applied. After encoding categorical features, we have 10 categorical features and 10 numerical features. To detect the outliers in the numerical features in the training set, we further analyze the distribution of the numerical features in the training set in boxplot. The results are shown in Appendix A.

Although some "outliers" are indeed detected using the boxplot, one thing needs to be mentioned is that all the data is collected from official website and database and the data quality can be guaranteed. Therefore, the detected "outliers" are mainly due to the variation in feature values instead of data inaccuracy or noise. As shown in Appendix A, five features do not contain outliers: beam, depth, length, GT, and follow-up-inspection-rate. Meanwhile, the outliers contained in the other features can be justified as follows. The oldest ship in the training set is 42 years old, while the oldest sailing ship still afloat over the world is more than 200 years old (Babamail, 2020), and thus the outliers detected for feature "age" are reasonable in our dataset. Although Figure A-1(d) shows that inspection records of ships with more than four ship flag changes are regarded to be outliers, there can be several reasons for the ship owners to decide to change ship flag, such as the policy on crew nationality requirements, trade flexibilities, and lower operating costs (Luo et al., 2013), and the value can be larger for older ships. Figure A-1(g) and Figure A-1(j) indicate that there are a lot of outliers for features "last-deficiency-no" and "total-detentions" as more than 16 deficiencies in last inspection and more than 4 times of total detentions are regarded to be outliers. This is because in practice, the average number of deficiencies in a PSC inspection is about 2 and the regional detention rate is about 3% in Tokyo MoU in 2018 (Tokyo MoU, 2019), which means that most of the ships are in satisfactory condition with few deficiencies and rare previous detentions. Therefore, ships with large number of deficiencies in last PSC inspection and several previous detentions should be paid more attention to as the features of these ships can provide valuable information on ship detention prediction. A lot of outliers also occur in the feature "last-inspection-time" for the records with the last inspection time more than 30.8 months ago. The reason is that as required by Tokyo MoU, the longest inspection time window for the ships is 18 months, out of which the ships are required to be inspected.

Nevertheless, some inspected ships only visit the Hong Kong port as well as other ports in the region of Tokyo MoU occasionally and some of them may have gone through ship repair, which means that the last inspection time can be much longer than 18 months (and even 30.8 months) and thus those values are also plausible. Therefore, although there are "outliers" detected by boxplot among the numerical features, we neither process them nor delete them based on the above domain knowledge. The distribution of the variables is shown in Appendix B in the supplementary materials.

To summarize, the data preprocessing scheme used in this study is summarized in Table 3.

Table 3. Summary of data preprocessing scheme

Data preprocessing method	Task
Feature selection	Selection of 15 features consisting of ship generic factors, dynamic factors, and inspection historical factors that are regarded to be highly related to ship detention
Data fusion	Combination of PSC inspection records from the public database provided by Tokyo MoU and ship related factors from World Register of Ships using ships' unique identifier
Dataset split	Randomly splitting the whole dataset into training set (80% of data) and test set (20% of data)
Categorical feature encoding	One-hot encoding for ship type and casualties in the last five years, and label encoding for ship flag performance, RO performance, and company performance
Outlier detection and analysis	Outlier detection and analysis is applied to numerical features

4. Ship detention prediction model

4.1 Introduction of decision tree (DT)

In this study, we adopt a classifier named balanced random forest (BRF) implemented in *imblearn* library in Python based on the framework proposed by Chen et al. (2004). BRF is a data-level method which is based on classification decision tree and random forest to address imbalanced classification problem. Before introducing the BRF model, we first introduce DT model, which is a popular supervised machine learning model for both regression and classification tasks. BRF contains several classification DTs based on classification and regression tree (CART) algorithm. All the training examples are first stored in the root node, and the root node is further split into successive nodes which contain subsets of the training examples in order to reduce node impurity. For each split, a feature and one of its values are selected for splitting. The criteria used to evaluate a split is Gini index (Breiman et al., 1984). CART algorithm requires recursively and binarily splitting the nodes

to build a binary DT. Originally, the split stops when all the nodes contain examples of the same output value. However, this may lead to too complicated trees that suffer from overfitting. Therefore, hyperparameters can be preset to control tree dimension. In this study, two hyperparameters for a single DT are used:

(a) *$max depth$*: the depth of a leaf node is the number of splits taken from the root node to that leaf node. The criterion of *$max depth$* requires the depth of all the leaf nodes in the DT not to exceed the value of *$max depth$*. The value of *$max depth$* is an integer, and too large value results in a complicated tree while too small value results in a too simple tree. Therefore, the value for *$max depth$* needs to be tuned.

(b) *$min samples leaf$*: the minimum number of examples that is required to be contained in a leaf node. The value of *$min samples leaf$* is an integer, and too small value results in a complicated tree while too large value results in a too simple tree. Therefore, the value for *$min samples leaf$* also needs to be tuned.

The procedure to generate a classification DT based on CART algorithm is presented in Appendix C in the supplementary materials (Breiman et al., 1984).

4.2 Introduction of random forest (RF) and balanced random forest (BRF)

Although DTs are interpretable and intuitive, a single DT is easy to overfit and is of high variance. To improve the performance of DTs, ensemble models based on DTs are proposed. An ensemble model contains multiple weak learners, which are prediction models performing a little better than random guessing. Random forest (RF) consisting of multiple DTs as weak learners is based on bootstrap aggregating (bagging) and is a state-of-the-art learning model that performs well in many applications (Breiman, 2001; Liaw and Wiener, 2002; Biau and Scornet, 2016). Compared to the construction process of a single DT, two layers of randomness are incorporated in RF construction process to reduce the dependence among the DTs: a bootstrap sample is used to construct each DT and a subset of features are considered for each split in a DT. Therefore, apart from the two hyperparameters in a single DT (i.e. *$max depth$* and *$min samples leaf$*), RF model has two more hyperparameters:

(c) *$n estimators$*: the number of DTs contained in the RF model. Averaging/voting of more trees is generally believed to better alleviate variance, and thus the value for this hyperparameter should be set as large as possible.

(d) *$max features$*: the number of features considered in each split. The value for *$max features$* is an integer

and the maximum value is the total number of input features. If the value is set to be too small, the performance of a single DT is negatively affected, whereas if the value is too large, the correlations of the DTs are increased. Therefore, the value for *$max features$* needs to be tuned.

If classical RF models are directly applied to imbalanced dataset, it is highly likely that bootstrap data contain few or even none of the minority samples, resulting in a tree with poor performance for predicting the minority class (Chen et al., 2004). This issue is addressed by the BRF model based on the idea of RF implemented by *imblearn* library in Python (Imbalanced-learn API, 2020). The only difference between the BRF model and the classical RF model when applied to binary classification is that when sampling for each single tree, BRF first draws all examples in the minority class and the same number of examples from the majority class without replacement to create a balanced dataset. Then, a bootstrap sample is drawn on the new balanced dataset before feeding to a single tree. For example, in our training set which contains 55 minority examples and 1,545 majority examples, the BRF would first draw all the 55 minority examples and 55 majority examples without replacement to form a new dataset containing 110 samples for a single tree. Then, a bootstrap sample of the new dataset is drawn for this tree. Like the classical RF, it only considers a subset of all the features for each split. All the other settings and the hyperparameters in BRF are the same as those in the RF models: *$n estimators$* DTs are contained; the depth of each DT should not exceed *$max depth$* and the minimum number of examples contained in a leaf node is *$min samples leaf$*. The number of features considered for each split is *$max features$*, and Gini index is used to evaluate a split.

4.3 Evaluation metrics

The evaluation metrics for binary classification problems are defined based on Table 4. Traditionally, model accuracy, which is represented by

$$accuracy = \frac{TP + TN}{TP + FP + FN + TN},$$

is the most commonly used measure to evaluate the performance of classifiers. However, it is not suitable to be applied to evaluate the performance of classifiers developed for imbalanced data. Under this condition, other effective evaluation metrics should be used.

Table 4. Confusion matrix for binary classification problem

	Predicted to be in class "1" (with detention)	Predicted to be in class "0" (without detention)
Actual in class "1" (with detention)	True positives (TP)	False negative (FN)
Actual in class "0" (without detention)	False positive (FP)	True negative (TN)

In this study, we adopt four popular metrics to evaluate the performance of the classifiers developed for imbalanced data (He and Garcia, 2009; Sun et al., 2009; Galar et al., 2012): *recall*, *precision*, *F-measure* and area under a *Receiver Operating Characteristics* (*ROC*) curve (*ROC AUC*) as our main focus is on the detained ships. The four metrics are defined based on Table 4 and the reasons for choosing them in this study is summarized in Table 5.

Table 5. Summary of metrics

Metric	Definition	Reason for choosing the metric
recall	$recall = \frac{TP}{TP + FN}$	It shows the percentage of detained ships that are correctly identified by the proposed model, which is also called true positive rate.
precision	$precision = \frac{TP}{TP + FP}$	It shows the percentage of detained ships among all the ships that are predicted to be detained by the proposed model.
F-measure	$F\text{-}measure = \frac{2}{\frac{1}{recall} + \frac{1}{precision}}$	It shows how accurate and robust a classifier is, especially when dealing with imbalanced data.
ROC AUC	Area under the curve composed by pairs of (FP_{rate}, TP_{rate}), where $FP_{rate} = \frac{FP}{FP + TN}$ and $TP_{rate} = recall$	It is an expectation that a uniformly drawn ship with detention is ranked before a uniformly drawn random ship without detention as predicted by the proposed classifier.

5. Model evaluation and results

5.1 Model performance

As mentioned in Section 4.2, a hyperparameter tuple containing 3 hyperparameters in the BRF model needs to be tuned: *$max depth$*, *$max features$*, and *$min samples leaf$*. We use grid search with 5-fold cross-validation and *ROC AUC* as the metric on the training set to tune the three hyperparameters. We fix *$n estimators$* to be 200, and if more than half of the trees in the BRF model (i.e. more than 100 trees) vote a ship to be detained, the ship is predicted to be detained as the final output. Otherwise, the ship is predicted not to be

detained. The search value space for *$max depth$*, *$max features$*, and *$min samples leaf$* and the optimal values found are shown in Table 6.

Table 6. Hyperparameter tuning in the BRF model

Hyperparameter	Search boundary	Optimal value
$max depth$	from 5 to 11	7
$max features$	from 4 to 9	9
$min samples leaf$	from 1 to 7	3

Both *$max depth$* and *$min samples leaf$* are used to control tree complexity: a deeper tree with smaller minimum number of samples required to be in a leaf node would fit the training data better and reduce bias to a larger extent. Meanwhile, as a tree gets more complex, the model variance becomes higher. As we are in the context of RF which contains a certain number of estimators to reduce variance, we can try deeper trees with smaller minimum number of samples required to be in a leaf node. Therefore, we let the search space for *$max depth$* contain large values from 5 to 11, and let the search space for *$min samples leaf$* contain small values from 1 (which is the minimum allowable value for this hyperparameter) to 7. As for *$max features$*, as the recommended value in regression problem is about $n\_features / 3$ (Friedman et al., 2001), which is between 6 and 7 in this problem. We extend the recommended value by 2 from two sides to form the search value space ranging from 4 to 9. It should also be mentioned that if we apply the BRF model to other problems with imbalanced datasets and even when dealing with the same problem with different datasets (e.g. the inspection records from other ports or other MoUs), the optimal hyperparameter values found in this problem may not be directly applied, as in practice the best values for these hyperparameters will depend on the problem, and should be treated as tuning parameter (Friedman et al., 2001). We then use the optimal hyperparameters to construct the BRF classifier on the whole training set and validate its performance on the test set. The confusion matrix of the test set is shown in Table 7. The performance of the BRF model is shown in Table 8.

Table 7. Confusion matrix of the test set by using BRF

	No. of predicted samples with detention	No. of predicted samples without detention	Total
No. of actual samples with detention	13	3	16
No. of actual samples without detention	46	338	384
Total	59	341	400

Table 8. Model performance on test set by using BRF

Metric	Average* *precision*	Average *recall*	Average *F-measure*	*ROC AUC*
Score	0.61	0.85	0.64	0.85

Note*: Average here means the arithmetic mean of the metric for class "1" and class "0"

The *precision* of class "0" (i.e. without detention) and class "1" (i.e. with detention) is 0.99 and 0.22, respectively, which means that among all the ships predicted not to be detained, 99% will not be detained. Meanwhile, 22% of the ships predicted to be detained will actually be detained. The arithmetic mean of the *precision* scores for class "0" and class "1" is 0.61. The *recall* of class "0" (i.e. without detention) is 0.88 and that of class "1" (i.e. with detention) is 0.81, which indicates that 88% of the ships without detention are accurately predicted whereas 81% of the ships with detention are accurately predicted by the BRF model. The arithmetic mean of the *recall* scores for class "0" and class "1" is 0.85. Given the *precision* and *recall* scores, *F-measure* for class "0" is 0.93 and *F-measure* for class "1" is 0.35, and thus the arithmetic mean of *F-measure* is 0.64. The *ROC AUC* of the BRF model is 0.85, which shows that the proposed model performs 70% better than random guessing.

In the above analysis, a ship is predicted to be detained if over half of the trees contained in the BRF predict it to be detained. We can further adjust the threshold (i.e. the number of trees that predict a ship to be detained divided by the total number of trees) if the decision maker would like to inspect less ships when the resources are limited (setting a higher threshold) or the decision maker would like to capture more ships that might be detained (setting a lower threshold). The decisions under different thresholds are shown in Table 9.

Table 9. Decisions for detention under different thresholds

Threshold	0.3	0.4	0.5 (benchmark)	0.6	0.7	0.8
No. of ships without detention & predicted to be detained	107	64	46	28	21	12
No. of ships with detention & predicted to be detained	15	13	13	10	6	6
Total no. of ships predicted to be detained	122	77	59	38	27	18

Table 9 indicates that if the inspection resources are quite limited, e.g. only no more than 30 ships can be inspected among all the 400 ships, then about 22.22% of the inspected ships will be detained. Considering the actual detention rate is about 3.55%, the proposed model can identify the detained ships about six times more efficiently. On the contrary, if the port state authority would like to identify more ships that will actually be detained, after inspecting 122 ships, 15 ships that are detained can be accurately identified and the detention rate is 12.30%, which is about 3.5 times as efficient as the currently implemented scheme.

5.2. Comparison with other machine learning models and Ship Risk Profile

5.2.1 Comparison with current literature

To the best of our knowledge, there are three studies that aim to predict detention rate of all ship types visiting a port in current literature: Xu et al. (2007a), Xu et al (2007b), and Gao et al. (2008). The three studies all use the *precision* of ships in class "1" (i.e. with detention) as the evaluation metric: the models first make prediction on the ships that are highly likely to be detained (i.e. of high-risk) and inspect them. If the ship is detained, the prediction is regarded to be accurate. The highest *precision* scores in the test sets of the three studies are 13.44%, about 20%, and 20.93%, respectively. In our model, 13 ships are detained among the 59 ships that are predicted to be high-risk, and thus the *precision* score is 22.03%, which is higher than that in the current literature.

5.2.2 Comparison with other machine learning models

In this section, the proposed BRF model is compared with the performance of several popular supervised and unsupervised machine learning models. For supervised machine learning models, we consider random forest (RF) (Breiman, 2001), gradient boosting decision tree (GBDT) (Breiman, 1997), RF with synthetic minority over-sampling technique (SMOTE) (Chawla et al., 2002), GBDT with SMOTE, RF in groups, and GBDT in groups. The RF and GBDT models are implemented by *sklearn* library in Python (Pedregosa et al., 2011) while fixing the number of trees to be 200, and the hyperparameters (i.e. *$max depth$*, *$max features$,* and *$min samples leaf$*) are tuned by grid search. The SMOTE algorithm is implemented by *imblearn* library (Imbalanced-learn API, 2020). More specifically, in RF with SMOTE and GBDT with SMOTE, balanced training set by over-sampling of the samples in minority class is first generated by using SMOTE algorithm

413 before feeding into the RF and GBDT models. In RF in groups and GBDT in groups, the training set are
414 randomly split into several sub-training sets with each containing all the 55 records with detention and 55
415 records without detention (i.e. by sub-sampling the majority class). To reduce overfit, records without detention
416 contained in the sub-training sets are mutually exclusive with each other, and thus we can form a total of 28
417 sub-training sets. In each sub-training set, a record is predicted to be detained if more than half of the decision
418 trees in the RF/GBDT model predict it to be detained. The final prediction result is then voted by all sub-training
419 sets: a record is predicted to be detained if more than 14 sub-training sets predict it to be detained; otherwise,
420 the record is predicted not to be detained.

421 Apart from using supervised machine learning models to predict ship detention, it is also interesting to treat
422 ship detention as anomaly and thus unsupervised anomaly detection methods can be applied. We adopt two
423 popular anomaly detection models for comparison: isolation forest (denoted by iForest) (Liu et al., 2008) which
424 is implemented by *sklearn* library and the number of trees contained in the model is set to be 200, and auto-
425 encoder neural network (denoted by auto-encoder NN) (Ballard, 1987) which is implemented by *PyOD* library
426 in Python (Zhao et al., 2019). It should be mentioned that as auto-encoder NN requires feature standardization,
427 we re-encode the missing values of the features related to previous PSC inspection, i.e. “last-inspection-time”,
428 “last-deficiency-number”, and “follow-up inspection rate” for the inspection records without PSC inspection
429 before. The values of “last inspection time” and “follow-up-inspection-rate” for ships without detention in both
430 training set and test set are filled by the mean values in the training set and “last-deficiency-number” is filled
431 by the median value in the training set. Then another feature called “first-time-inspection” is added, which is
432 set to 1 if a ship is not inspected before and 0, otherwise. Therefore, we have a total of 21 features, and thus
433 both the input layer and output layer of an auto-encoder NN should have 21 nodes. We consider three auto-
434 encoder NNs with one or two hidden layers as we only have limited number of features and training samples.
435 More specifically, we consider two auto-encoder NNs of one hidden layer with 5 nodes and 10 nodes
436 respectively (denoted by auto-encoder NNs(a), auto-encoder NNs(b)), and one auto-encoder NN of two hidden
437 layers with 8 nodes in each layer (denoted by auto-encoder NNs(c)). We further set the ratio of anomalous
438 samples to be 15%, which means that a sample is regarded to be anomalous if and only if its generated score is

in top 15%. All the other settings are in default. The performance of the supervised and unsupervised machine learning models is shown in Table 10.

Table 10. Model performance and comparison

Metric/Model	Average *Precision*	Average *recall*	Average *F-measure*	*ROC AUC*
BRF	0.61	**0.85**	0.64	**0.85**
RF	0.52	0.50	0.51	0.50
GBDT	**0.68**	0.56	0.58	0.56
RF+SMOTE	0.58	0.58	0.58	0.58
GBDT+SMOTE	0.62	0.58	0.60	0.58
RF in groups	0.61	0.82	**0.65**	0.82
GBDT in groups	0.60	0.82	0.64	0.82
iForest	0.59	0.71	0.62	0.71
auto-encoder NNs(a)	0.56	0.71	0.58	0.71
auto-encoder NNs(b)	0.56	0.71	0.58	0.71
auto-encoder NNs(c)	0.56	0.71	0.58	0.71

Table 10 indicates the BRF model performs best among all the models listed if evaluated by average *recall* and *ROC AUC*, while GBDT performs best if evaluated by average *precision* and RF in groups performs best if evaluated by average *F-measure*. If evaluated by *ROC AUC*, two models combining random sub-sampling of majority class with classical RF and GBDT models (i.e. RF in groups and GBDT in groups) perform second best, followed by the unsupervised anomaly detection models: iForest and auto-encoder NNs in three different structures. Meanwhile, applying the classical RF model and GBDT model directly to the imbalanced dataset has the worst performance, and the performance of RF model is even no better than random guessing. The results indicate that classical machine learning models for classification could not perform well on imbalanced dataset, even for the state-of-the-art models such as RF and GBDT, which is consistent with current literature (Sun et al., 2009). Besides, the performance of the models which combine over-sampling or generation of synthetic samples in minority class (such as SMOTE) with traditional machine learning models neither perform well as a result of overfitting and the inaccuracy brought about by the generation process of synthetic samples. Although balanced dataset can be formed by using random sub-sampling methods in majority class before feeding into classical machine learning models, such as RF in groups and GBDT in groups, model performance can be adversely impacted as only a very small number of samples in majority class are used in the model of each group. The samples in minority class can be regarded as anomalies, which refer to the patterns in the data

459 that do not conform to a well-defined notion of normal behavior. Within this context, anomaly detection models
460 can also be applied to the classification task of highly imbalanced dataset. Anomaly detection refers to the
461 problem of finding patterns in data that do not conform to expected behavior (Chandola et al., 2009). As ship
462 detention is mainly determined by whether detainable (fatal) deficiencies are detected in the current PSC
463 inspection (which can only be observed during the inspection process), ships with detention may not have
464 distinct and abnormal features (including ship-related features and PSC inspection related features) compared
465 to ships without detention, and thus the performance of applying anomaly detection methods to predict ship
466 detention may not have supreme performance. Meanwhile, data of normal samples can often contain features
467 that tend to be similar to the actual anomalies and hence makes the anomalies difficult to be distinguished and
468 removed (Chandola et al., 2009).

469 The main reason for us to choose BRF model as the prediction model of ship detention is threefold. First,
470 it actually makes a compromise between sub-sampling and over-sampling, which are both popular methods to
471 deal with imbalanced dataset. To be more specific, it randomly formulates balanced datasets and bootstraps on
472 the datasets to construct each tree (sub-sampling in majority class). Meanwhile, as several trees contain in a
473 BRF model, over-sampling of minority class and several times of sub-sampling of majority class can be realized
474 as the forest grows, and thus the problem of overfitting and ignoring too many majority samples can be reduced.
475 Besides, it does not require the samples in minority class to have remarkable abnormal features compared with
476 the samples in majority class and thus is more suitable to the task of ship detention detection. On top of that, as
477 BRF model is based on decision tree which are explainable and can be visualized, the working process of the
478 BRF model is more comprehensible and acceptable for the experts in shipping industry.

479 **5.2.3 Comparison with Ship Risk Profile**

480 The currently implemented ship selection scheme at the Tokyo MoU is Ship Risk Profile (SRP). SRP
481 assigns different weighting points to different states of ship type, ship age, ship flag, RO and company
482 performance, and the number of deficiencies and detentions in previous PSC inspections (Tokyo MoU, 2014).
483 Based on the total points, ships are divided into three risk categories: low risk ship (LRS), standard risk ship
484 (SRS), and high risk ship (HRS). Inspection time windows are attached to the profiles, and thus the inspection

priority can be determined as presented in Wang et al. (2019). Based on the SRP ship selection scheme, the PSC inspection rate (No. of individual ships inspected/No. of individual ships visited) at the Hong Kong port is 12.02% (621 out of 5,165) in 2016, 11.97% (632 out of 5,280) in 2017, and 13.52% (708 out of 5,235) in 2018, and the average inspection rate over the three years is 12.50% (Tokyo MoU, 2017a, 2018a, 2019). As there are 400 ships in our test set which are actually inspected, the total number of visiting ships can be estimated to be 3,200 and 16 of them are detained. We compare the ship selection efficiency of the SRP and the proposed BRF model by calculating the number of detentions identified after inspecting 0.5%, 1%, 1.5%, 2%, …, 12.5% of the total 3,200 foreign visiting ships. The inspection sequence generated by the SRP is calculated by using the formulas proposed by Wang et al. (2019). The inspection sequence generated by the BRF model is determined by the voting rate of the trees in the BRF model for each ship in descending order. The voting rate for a ship is the number of trees predicting the ship to be detained divided by the total number of trees in the BRF model. The results are shown in Figure 2.

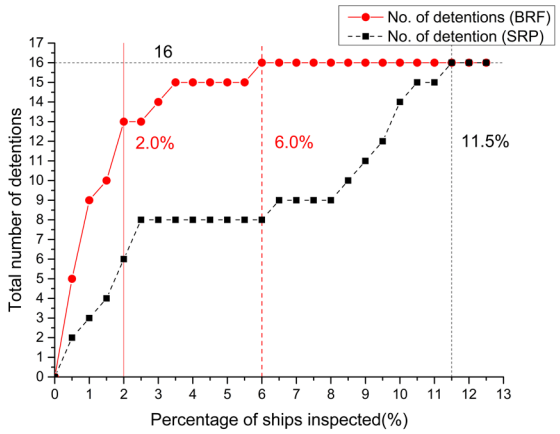

Figure 2. Comparison between BRF and SRP regarding ship detention

In Figure 2, the Y axis shows the total number of identified ships with detention, and the X axis shows the percentage of ships that are inspected among all the foreign visiting ships. We analyze Figure 2 from three perspectives:

(a) Two red vertical lines in Figure 2 show the differences between the BRF model and the SRP considering the ability to identify detained ships after inspecting the same number of ships (using the same inspection resources). For example, when the inspection rate is 2.0% (i.e. selecting 64 ships for inspection out of the total 3,200 coming

505 ships), the proposed BRF model can identify 13 ships with detention whereas the total number of detained ships
506 identified by SRP is only six. Therefore, the BRF model is 2.17 times more efficient than the SRP considering
507 the ability to identify detained ships. Meanwhile, the BRF model could pick out the total 16 detained ships after
508 inspecting 6% of all the visiting ships as indicated by the red vertical dash line. If 6% of ships are inspected by
509 the SRP, only eight of the detained ships can be identified.
510 (b) The intersections of the black dot horizontal line with the red dash vertical line and black dot vertical line
511 show the resources needed to identify all the 16 detained ships in the test set. The BRF model needs to inspect
512 6% of all ships to find out the total 16 detained ships whereas the SRP needs to inspect 11.5% of all ships to
513 identify all the detained ships. Therefore, the overall efficiency of BRF model is 1.92 times higher than that of
514 SRP ship selection scheme.
515 (c) Overall, after inspecting the total 400 ships, the average improvement ((No. of detentions identified by BRF
516 – No. of detentions identified by SRP)/No. of detentions identified by SRP) when the inspection rate is 0.5%,
517 1.0%, …, 12.0% of the BRF model over the SRP is 73.72%.

518 **5.3 Decision tree performance in the BRF model**

519 We further denote the ratio of ships selected for inspection as r, $r = 0.5\%, 1.0\%, \ldots, 12.5\%$, and the total
520 number of visiting ships as $N = 400$. We calculate the total number of detentions identified by an average tree
521 (the average detentions identified by one tree in BRF model, denoted by avg_tree for short), the whole BRF
522 model, and SRP selection scheme given the inspection ratio r to compare their performance. Denote the total
523 number of trees contained in the BRF model as M, and one tree is denoted by m. For tree m, denote the
524 number of ships predicted to be detained as α_m, and thus the number of ships predicted not to be detained is
525 $400 - \alpha_m$. β_m ships will actually be detained among the α_m ships predicted to be detained, and $\hat{\beta}_m$ ships will
526 actually be detained among the $400 - \alpha_m$ ships predicted not to be detained. The detention rate of the ships
527 predicted to be detained is $\eta_m = \beta_m / \alpha_m$ and the detention rate of the ships predicted not to be detained is
528 $\hat{\eta}_m = \hat{\beta}_m / (400 - \alpha_m)$. Given the inspection ratio r, the procedure to construct an avg_tree Θ^r is presented in
529 Procedure 2:

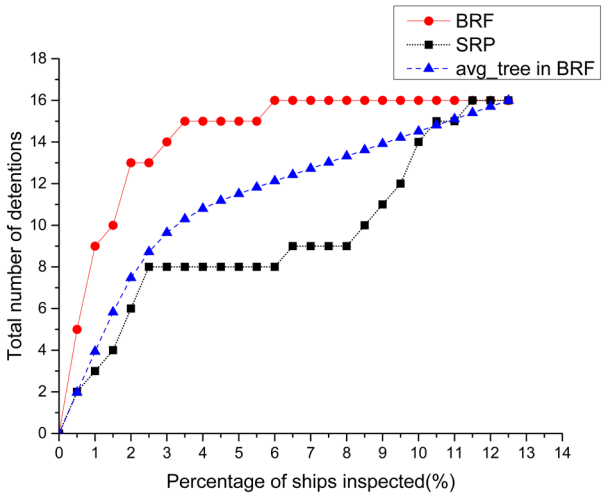

Figure 3. Comparison results of the BRF, SRP, and the avg_tree in BRF

Procedure 2: Construction of avg_tree given inspection ratio

Input inspection ratio r; total number of ships N; total number of trees M; the number of ships predicted to be detained α_m, $m = 1,...,M$; the detention rate η_m among ships predicted to be detained and $\hat{\eta}_m$ among ships predicted not to be detained, $m = 1,...,M$.

Output avg_tree Θ^r given r

Step 1: for tree $m = 1,...,M$

if $\alpha_m \geq N \times r$

Randomly pick out $N \times r$ ships from all the ships predicted to be detained for inspection. The estimated number of identified ships with detention is $\theta_m^r = N \times r \times \eta_m$.

else

Select all the α_m ships predicted to be detained and $N \times r - \alpha_m$ ships randomly from the ships predicted not to be detained. The estimated number of identified ships with detention is $\theta_m^r = \alpha_m \times \eta_m + (N \times r - \alpha_m) \times \hat{\eta}_m$.

end if

end for

Step 2: Calculate avg_tree Θ^r given r as $\Theta^r = (\sum_{m=1}^{M} \theta_m^r) / M$ and return Θ^r.

The comparison results of the performance of the avg_tree in BRF, the whole BRF model and the SRP ship selection scheme given different inspection ratio r are shown in Figure 3.

Figure 3 shows that the performance of the avg_tree in the BRF model is better than the currently used SRP ship selection scheme while is worse than the whole BRF model. Particularly, the avg_tree performs better than SRP when the inspection rate is between 1.0% and 10.0% and at 11.0%. When the inspection rate is 0.5%,

10.5%, 11.5% or 12.0%, the avg_tree performs a little worse than SRP. In addition, the average improvement of the avg_tree over SRP is 26.13%.

To better illustrate the structure of a decision tree in the BRF model, we randomly visualize a decision tree with average performance in the BRF model as shown in Figure 4.

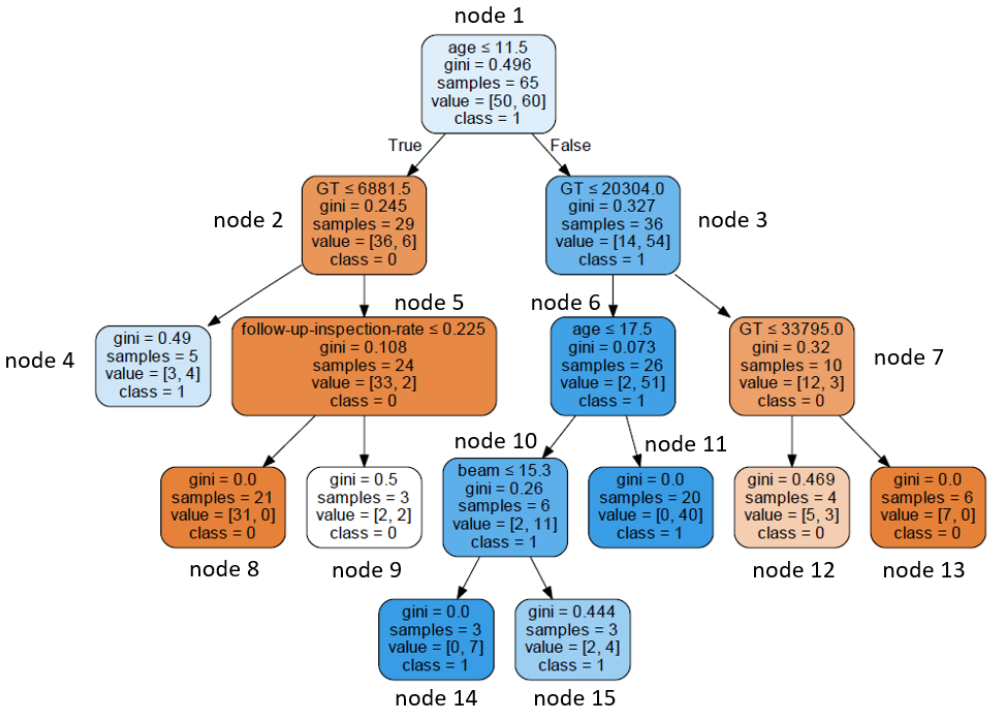

Figure 4. Visualization of a decision tree in BRF

The decision tree contains three types of nodes: root node, internal node, and leaf node. As shown in Figure 4, node 1 is the root node containing all the training examples. As the training set contains 55 examples in class "1", the total number of examples in the new dataset generated by sampling all the examples in class "1" and the same number of examples in class "0" without replacement is 110. Then a bootstrap sample is drawn from the new dataset, with 50 examples in class "0" and 60 examples in class "1". As bootstrapping relies on random sampling with replacement, there are 65 distinct examples among the total 110 selected examples. The leaf nodes (i.e. node 4, node 8, node 9, nodes 11 to 15) are the nodes that are not further split and give the final prediction results, while the internal nodes (i.e. node 2, node 3, nodes 5 to node 7, and node 10) are further split to other internal nodes or leaf nodes. The selected splitting feature and value are shown in the first line of the root node and the internal nodes. For example, the splitting feature selected by node 3 is "GT" and the

corresponding splitting value is “20304.0”. This means that among all the examples contained in node 3, examples with feature “GT” no more than “20304.0” are split to the left branch (i.e. node 6) while examples with “GT” more than “20304.0” are split to the right branch (i.e. node 7). In all the nodes shown in Figure 4, “gini” is the Gini index of the examples contained in this node. “Samples” is the number of distinct examples contained in this node, while “value” is a list with the first term representing the total number of samples in class “0” and the second term representing the total number of examples in class “1”. The confusion matrix generated by this decision tree on the test set is shown in Table 11, and its performance compared to the whole BRF model and SRP is shown in Figure 5.

Table 11. Confusion matrix of test set generated by a decision tree in BRF

	No. of predicted samples with detention	No. of predicted samples without detention	Total
No. of actual samples with detention	13	3	16
No. of actual samples without detention	101	283	384
Total	114	286	400

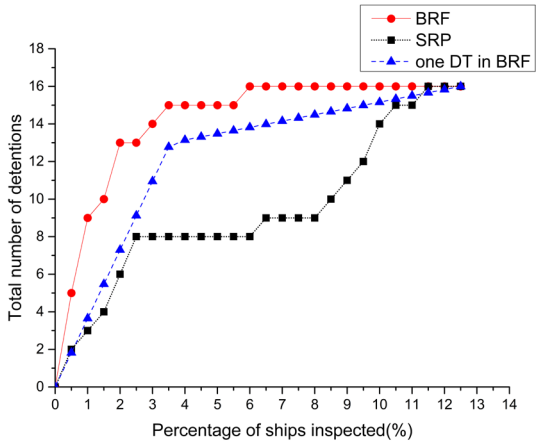

Figure 5. The performance of BRF, SRP and one decision tree in the BRF model

5.4 Model performance on balanced dataset

To access the overall accuracy of the proposed BRF model, we apply the BRF model to a new balanced dataset formulated by applying the SMOTE algorithm to the original imbalanced dataset. Recall that we totally have 2,000 records in the whole dataset, while 71 of them are with detention and 1,929 of them are without detention. We first apply the SMOTE algorithm to form a new balanced dataset which contains 3,858 sample, with 1,929 samples with detention (71 are from the original dataset while 1,858 are synthetic) and 1,929 samples

without detention (all of them are from original dataset). Then, we randomly divide the new balanced dataset into training set (80% samples) and test set (20% samples). To find the optimal values for hyperparameters *$max depth$*, *$max features$*, and *$min samples leaf$*, we apply grid search method for hyperparameter tuning by 5-fold cross-validation on the training set with metric as 'accuracy'. The optimal values for *$max depth$*, *$max features$*, *$min samples leaf$* are 9, 6, and 1, respectively. The BRF model is constructed on the whole training set with the optimal hyperparameter values. The confusion matrix of the prediction results is shown in Table 12 and the performance of the model is summarized in Table 13.

Table 12. Confusion matrix of the test set by using BRF (new balanced dataset)

	No. of predicted samples with detention	No. of predicted samples without detention	Total
No. of actual samples with detention	369	4	373
No. of actual samples without detention	21	378	399
Total	390	382	772

Table 13. Model performance on test set by using BRF (new balanced dataset)

Metric	Average* *accuracy*	Average *precision*	Average *recall*	Average *F-measure*	*ROC AUC*
Score	0.97	0.97	0.97	0.97	0.97

Note*: Average here means the arithmetic mean of the metric for class "1" and class "0"

It can be seen from Table 13 that when applying the BRF model on balanced dataset, it can achieve satisfactory performance with average *accuracy* as 0.97 and *ROC AUC* as 0.97. Actually, the working process of the BRF model on balanced dataset is quite similar to the traditional RF model as the bootstrap sample is also generated on the whole dataset (with equal number of samples in two classes) for each tree.

6. Conclusions and future work

PSC inspection is the guard of marine safety, the marine environment, and the decent working and living conditions of seafarers. To help the port state authorities to identify ships that are highly likely to be detained in PSC inspections, a BRF model which is able to address the imbalanced distribution of the dataset in the classification problem is developed. The BRF model is constructed by using 1,600 inspection records at the

591 Hong Kong port from Jan 2016 to Dec 2018 and its performance is validated by another 400 inspection records
592 conducted at the same port in the same time period. The average *F-measure* of the BRF model is 0.64 and the
593 *ROC AUC* is 0.85 on the test set. Besides, 81.25% of the ships that are actually detained can be identified by
594 the BRF model. If the same inspection resources (i.e. the resources used to inspect the same number of ships)
595 are used to inspect the ships selected by the SRP, only six of the detained ships can be identified. Meanwhile,
596 when the inspection resources allowing all the 16 ships with detention in the test set identified by the BRF model
597 are allocated to inspect the ships selected by the SRP, only eight of the detained ship can be identified. To allow
598 the SRP finding out all the ships with detention, 91.67% more ships need to be inspected compared to the BRF
599 model. Overall, after inspecting the total 400 ships, the average improvement of the BRF model over the SRP
600 is 73.72%. To better illustrate the working process of the BRF model, the performance of an average decision
601 tree in the BRF model is analyzed. The average decision tree performs better than SRP ship selection scheme
602 with the average improvement as 26.13%. A decision tree in the BRF model is also visualized and discussed.

603 The BRF model which aims to address the problem of classification on imbalanced dataset can be further
604 applied to address the practical problems in road transport and air transport where imbalance exists in datasets.
605 For road transport, the BRF model can help to solve the problems of vehicle classification on imbalanced
606 datasets, road traffic crashes detection and prediction, imbalanced traffic flow and traffic congestion prediction,
607 and severe traffic accident prediction, etc. For air transport, the BRF model can be applied to air crash prediction,
608 unqualified craft detection, airplane failure detection, and flight delay prediction, etc.

609 The proposed BRF model is the very first few models that take the imbalanced distribution of ships with
610 and without detention into account when developing prediction models on ship detention. It can help the port
611 state authorities to target high-risk ships more accurately and efficiently, and thus enhance the role of PSC
612 inspection for guaranteeing “safer shipping, cleaner oceans”. For future research, we can further combine
613 different databases to incorporate more ship and port state features, such as the database of ship accidents and
614 incidents, Lloyd’s Register database which provides information on ship assurance, certification, inspection,
615 and training, and the background of port state control officers. Moreover, once new ship selection schemes based
616 on machine learning models are implemented, we can continuously collect data and apply reinforced learning

to improve the outcome (Zhou et al., 2019; Qu et al., 2020). Besides, the proposed BRF model can be applied to other port states for ship detention prediction, and the predicted results can be analyzed and compared to generate insights on differences among ports as well as practical management strategies.

Acknowledgement

The authors thank two anonymous referees for their constructive comments and suggestions. This study is supported by the Policy Innovation and Co-ordination Office (PICO) of the Government of the HKSAR (Project number: 2020.A6.148.20A).

Reference

Abbassi, A., El hilali Alaoui, A., & Boukachour, J. (2019). Robust optimisation of the intermodal freight transport problem: Modeling and solving with an efficient hybrid approach. *Journal of Computational Science*, *30*, 127-142.

Babamail (2020). 10 of the oldest ships sailing the seas today. Accessed 14 July 2020. https://www.ba-bamail.com/content.aspx?emailid=26875.

Ballard, D. H. (1987). Modular Learning in Neural Networks. *AAAI*, 279-284.

Biau, G., & Scornet, E. (2016). A random forest guided tour. *Test*, *25*(2), 197–227.

Breiman, L. (1997). Arcing the edge. *Technical Report 486, Statistics Department, University of California at Berkeley.*

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32.

Breiman, L., Friedman, J., Stone, C. J., & Olshen, R. A. (1984). *Classification and Regression Trees*. Taylor & Francis, Abingdon, the United Kingdom.

Cariou, P., Mejia Jr, M. Q., & Wolff, F. C. (2007). An econometric analysis of deficiencies noted in port state control inspections. *Maritime Policy & Management*, *34*(3), 243–258.

Cariou, P., Mejia Jr, M. Q., & Wolff, F. C. (2008). On the effectiveness of port state control inspections. *Transportation Research Part E*, *44*(3), 491–503.

643 Cariou, P., Mejia, M. Q., & Wolff, F. C. (2009). Evidence on target factors used for port state control
644 inspections. *Marine Policy*, *33*(5), 847–859.
645 Cariou, P., & Wolff, F. C. (2011). Do port state control inspections influence flag-and class-hopping phenomena
646 in shipping? *Journal of Transport Economics and Policy*, *45*(2), 155–177.
647 Cariou, P., & Wolff, F. C. (2015). Identifying substandard vessels through port state control inspections: A new
648 methodology for concentrated inspection campaigns. *Marine Policy*, *60*, 27–39.
649 Chandola, V., Banerjee, A., & Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys*
650 *(CSUR)*, *41*(3), 1-58.
651 Chawla, N. V., Bowyer, K. W., Hall, L. O., & Kegelmeyer, W. P. (2002). SMOTE: synthetic minority over-
652 sampling technique. *Journal of artificial intelligence research*, *16*, 321-357.
653 Chen, C., Liaw, A., & Breiman, L. (2004). Using random forest to learn imbalanced data. *Tech-report of*
654 *University of California, Berkeley*, *110,* 1–12.
655 Chen, J., Zhang, S., Xu, L., Wan, Z., Fei, Y., & Zheng, T. (2019). Identification of key factors of ship detention
656 under Port State Control. *Marine Policy*, *102*, 21–27.
657 Degré, T. (2007). The use of risk concept to characterize and select high risk vessels for ship inspections. *WMU*
658 *Journal of Maritime Affairs*, *6*(1), 37–49.
659 Degre, T. (2008). From black-grey-white detention-based lists of flags to black-grey-white casualty-based lists
660 of categories of vessels? *The Journal of Navigation*, *61*(3), 485–497.
661 EMSA (2019). European Maritime Safety Agency. Annual overview of marine
662 casualties and incidents 2019. Accessed 18 Dec 2019. http://www.emsa.europa.eu/news-a-press-
663 centre/external-news/item/3734-annual-overview-of-marine-casualties-and-incidents-2019.html.
664 Fan, L., Luo, M., & Yin, J. (2014). Flag choice and Port State Control inspections—Empirical evidence using a
665 simultaneous model. *Transport Policy*, *35*, 350–357.
666 Fan, L., Wang, M., & Yin, J. (2020). The impacts of risk level based on PSC inspection deficiencies on ship
667 accident consequences. *Research in Transportation Business & Management, 33,* 1–9.

668 Friedman, J., Hastie, T., & Tibshirani, R. (2001). The Elements of Statistical Learning. Springer Publisher,
669 Berlin, Germany.
670 Galar, M., Fernandez, A., Barrenechea, E., Bustince, H., & Herrera, F. (2011). A review on ensembles for the
671 class imbalance problem: bagging-, boosting-, and hybrid-based approaches. *IEEE Transactions on*
672 *Systems, Man, and Cybernetics, Part C*, *42*(4), 463–484.
673 Gao, Z., Lu, G., Liu, M., & Cui, M. (2008). A novel risk assessment system for port state control inspection. In
674 Proceedings of 2008 *IEEE International Conference on Intelligence and Security Informatics*, 242–244.
675 Graziano, A., Schröder-Hinrichs, J. U., & Ölcer, A. I. (2017). After 40 years of regional and coordinated ship
676 safety inspections: Destination reached or new point of departure? *Ocean Engineering*, *143*, 217–226.
677 Graziano, A., Mejia Jr, M. Q., & Schröder-Hinrichs, J. U. (2018). Achievements and challenges on the
678 implementation of the European Directive on Port State Control. *Transport Policy*, *72*, 97–108.
679 Hänninen, M., & Kujala, P. (2014). Bayesian network modeling of port state control inspection findings and
680 ship accident involvement. *Expert Systems with Applications*, *41*(4), 1632–1646.
681 He, H., & Garcia, E. A. (2009). Learning from imbalanced data. *IEEE Transactions on Knowledge and Data*
682 *Engineering*, *21*(9), 1263–1284.
683 Heij, C., Bijwaard, G. E., & Knapp, S. (2011). Ship inspection strategies: Effects on maritime safety and
684 environmental protection. *Transportation Research Part D*, *16*(1), 42–48.
685 Heij, C., & Knapp, S. (2019). Shipping inspections, detentions, and incidents: an empirical analysis of risk
686 dimensions. *Maritime Policy & Management*, *46*(7), 866–883.
687 Imbalanced-learn API (2020). Imbalanced-learn API. Accessed 3 April 2020. https://imbalanced-
688 learn.readthedocs.io/en/stable/api.html.
689 IMO (2017). International Maritime Organization. Resolution A.1119(30): Procedure for port state cont
690 rol, 2017. Accessed 17 May 2019. http://www.imo.org/en/KnowledgeCentre/IndexofIMOResolutions/
691 Assembly/Documents/A.1119%2830%29.pdf.
692 Knapp, S., & Franses, P. H. (2007). A global view on port state control: econometric analysis of the differences
693 across port state control regimes. *Maritime Policy & Management*, *34*(5), 453–482.

694 Knapp, S., & Heij, C. (2020). Improved strategies for the maritime industry to target vessels for inspection and
695 to select inspection priority areas. *Safety*, *6*(2), 1–21.
696 Li, K. X., Yin, J., Bang, H. S., Yang, Z., & Wang, J. (2014). Bayesian network with quantitative input for
697 maritime risk analysis. *Transportmetrica A: Transport Science*, *10*(2), 89–118.
698 Liaw, A., & Wiener, M. (2002). Classification and regression by random forest. *R news*, *2*(3), 18–22.
699 Liou, S. T., Liu, C. P., Chang, C. C., & Yen, D. C. (2011). Restructuring Taiwan's port state control inspection
700 authority. *Government Information Quarterly*, *28*(1), 36–46.
701 Liu, F. T., Ting, K. M., & Zhou, Z. H. (2008). Isolation forest. In Proceedings of *2008 Eighth IEEE International*
702 *Conference on Data Mining*, 413-422.
703 Luo, M., Fan, L., & Li, K. X. (2013). Flag choice behaviour in the world merchant fleet. *Transportmetrica A:*
704 *Transport Science, 9(5), 429–450.*
705 Mansell, J. (2009). Port state control in the Asia-Pacific region: Issues and challenges. *Australian Journal of*
706 *Maritime & Ocean Affairs*, *1*(3), 73–87.
707 Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P.,
708 Weiss, R., Dubourg, V., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E. (2011). Scikit-
709 learn: machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825-2830.
710 Qu, X., Yu, Y., Zhou, M., Lin, C. T., & Wang, X. (2020). Jointly dampening traffic oscillations and improving
711 energy consumption with electric, connected and automated vehicles: A reinforcement learning based
712 approach. *Applied Energy*, 257, 114030.
713 Ramdin, M., Chen, Q., Balaji, S. P., Vicent-Luna, J. M., Torres-Knoop, A., Dubbeldam, D., ... & Vlugt, T. J.
714 (2016). Solubilities of CO2, CH4, C2H6, and SO2 in ionic liquids and Selexol from Monte Carlo
715 simulations. *Journal of Computational Science*, *15*, 74-80.
716 Paris MoU (2014). Ship risk profile. Accessed 19 April 2020. https://www.parismou.org/system/files/A
717 nnex%207%20ship%20risk%20profile.pdf.
718 Ravira, F. J., & Piniella, F. (2016). Evaluating the impact of PSC inspectors' professional profile: a case study
719 of the Spanish Maritime Administration. *WMU Journal of Maritime Affairs*, *15*(2), 221–236.

720 Rekik, I., & Elkosantini, S. (2019). A multi agent system for the online container stacking in seaport
721 terminals. *Journal of Computational Science*, *35*, 12-24.
722 Şanlıer, Ş. (2020). Analysis of port state control inspection data: The Black Sea Region. *Marine Policy*, *112*, 1–
723 11.
724 Sun, Y., Wong, A. K., & Kamel, M. S. (2009). Classification of imbalanced data: A review. *International Journal*
725 *of Pattern Recognition and Artificial Intelligence*, *23*(04), 687–719.
726 Tokyo MoU (2014). Information sheet of the New Inspection Regime (NIR). Accessed 20 November 2018.
727 http://www.tokyo-mou.org/doc/NIR-information%20sheet-r.pdf.
728 Tokyo MoU (2017a). Annual Report on Port State Control in the Asia-Pacific Region 2016. Accessed 27
729 December 2018. http://www.tokyo-mou.org/doc/ANN16.pdf.
730 Tokyo MoU (2017b). List of Tokyo MOU Deficiency Codes. Accessed 25 November 2018. http://www.tokyo-
731 mou.org/doc/NIR-information%20sheet-r.pdf.
732 Tokyo MoU (2018a). Annual Report on Port State Control in the Asia-Pacific Region 2017. Accessed 28
733 October 2018. http://www.tokyo-mou.org/doc/ANN17.pdf.
734 Tokyo MoU (2018b). Memorandum of Understanding on Port State Control in the Asia-Pacific Region.
735 Accessed 19 October 2019. http://www.tokyo-mou.org/.
736 Tokyo MoU (2019). Annual Report on Port State Control in the Asia-Pacific Region 2018. Accessed 21 July
737 2019. http://www.tokyo-mou.org/doc/ANN18.pdf.
738 Tokyo MoU (2020). Annual Report on Port State Control in the Asia-Pacific Region 2019. Accessed 11 July
739 2020. http://www.tokyo-mou.org/doc/ANN19-f.pdf.
740 Tsou, M. C. (2019). Big data analysis of port state control ship detention database. *Journal of Marine*
741 *Engineering & Technology*, *18*(3), 113–121.
742 Wang, S., Yan, R., & Qu, X. (2019). Development of a non-parametric classifier: Effective identification,
743 algorithm, and applications in port state control for maritime transportation. *Transportation Research Part*
744 *B*, *128*, 129–157.
745 Wu, J., Kulcsár, B., Ahn, S., & Qu, X. (2020). Emergency vehicle lane pre-clearing: From microscopic

cooperation to routing decision making. *Transportation Research Part B*, 141, 223–239.

Xia, D. F., Xu, S. L., & Qi, F. (1999). A proof of the arithmetic mean-geometric mean-harmonic mean inequalities. *RGMIA Research Report Collection*, *2*(1), 1–10.

Xiao, Y., Wang, G., Lin, K. C., Qi, G., & Li, K. X. (2020). The effectiveness of the New Inspection Regime for Port State Control: Application of the Tokyo MoU. *Marine Policy*, 1–8.

Xu, R. F., Lu, Q., Li, W. J., Li, K. X., & Zheng, H. S. (2007). 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. (2007). 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., Zhuge D., & Wang S. (2020). Development of two highly-efficient and innovative inspection schemes for PSC inspection. *Asia Pacific Journal of Operations Research*, in press.

Yang, Z., Yang, Z., & Yin, J. (2018). Realising advanced risk-based port state control inspection using data-driven Bayesian networks. *Transportation Research Part A*, *110*, 38–56.

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

Zhao, Y., Nasrullah, Z., & Li, Z. (2019). Pyod: A python toolbox for scalable outlier detection. *arXiv:1901.01588*.

Zhou, M., Yu, Y., & Qu, X. (2019). Development of an efficient driving strategy for connected and automated vehicles at signalized intersections: A reinforcement learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 21(1), 433–443.