Journal of
Marine Science and Technology

Volume 31 | Issue 4 Article 2

Machine Learning with Multi-Source Data to Predict and Explain Marine Pilot Occupational Accidents

Gokhan Camliyurt
Department of Navigation, Graduate School, Korea Maritime, and Ocean University

Youngsoo Park
Division of Navigation Convergence Studies, Korea Maritime, and Ocean University

Daewon Kim
Division of Navigation Convergence Studies, Korea Maritime, and Ocean University

Won Sik Kang
College of Ocean Sciences, Jeju National University

Sangwon Park
Logistics and Maritime Industry Research Department, Korea Maritime Institute, psw6745@kmi.re.kr

Recommended Citation
Camliyurt, Gokhan; Park, Youngsoo; Kim, Daewon; Kang, Won Sik; and Park, Sangwon (2023) "Machine Learning with Multi-Source Data to Predict and Explain Marine Pilot Occupational Accidents," *Journal of Marine Science and Technology*: Vol. 31: Iss. 4, Article 2.
DOI: 10.51400/2709-6998.2709
Available at: https://jmstt.ntou.edu.tw/journal/vol31/iss4/2

RESEARCH ARTICLE

Machine Learning With Multi-Source Data to Predict and Explain Marine Pilot Occupational Accidents

Gokhan Camliyurt [a], Youngsoo Park [b], Daewon Kim [b], Won-Sik Kang [c], Sangwon Park [d],*

[a] Department of Navigation, Graduate School, Korea Maritime, and Ocean University, South Korea
[b] Division of Navigation Convergence Studies, Korea Maritime, and Ocean University, South Korea
[c] College of Ocean Sciences, Jeju National University, South Korea
[d] Logistics and Maritime Industry Research Department, Korea Maritime Institute, South Korea

Abstract

Marine pilot occupational accidents during transfer to/from ships are the primary concern of the International Marine Pilots' Association (IMPA) and industry professionals. There are multiple transfer methods for marine pilots, with the most common being the pilot boat. To reach the mother ship bridge, the following stages must be safely completed: car transfer, walking on the pier, pier to pilot boat, pilot transfer by boat, cutter to pilot ladder, mother ship freeboard climbing, and ship deck to the bridge. Each stage has its own risk. Previous accident records and expert opinions are commonly used to conduct a risk analysis and take preventive actions. However, the reports vary in scope and are often complex, making qualitative analysis a timeintensive task. To overcome this challenge, this study aggregates 500 reports to create a multi-source dataset describing instances of undesired events. A ML (machine learning) approach is used to predict and explain marine pilot occupational accidents. Analyzing the importance of factors distinguishing between accidents, incidents, and non-compliance, we conclude that workplace factors are more dangerous than environmental factors. The findings of this study provide a foundation for developing a unified accident reporting system for predicting accidents on a wider scale.

Keywords: Marine pilot, Pilot ladder, Occupational accident, RF (random forest), Explainable ML

1. Introduction

The shipping industry has witnessed rapid growth over the last three decades due to the increasing demand for cargo delivery. The global fleet's size has increased linearly from 1.3 to 2.1 million within ten years [52]. A significant increase in traffic demand has increased traffic movements within port waters. In general, the number of traffic movements on a busy fairway in port waters can be as high as 2000 per day [56]; this number is expected to increase further with the continuing growth of traffic demand [53]. An increase in traffic density also increases the probability of ship accidents.

Traffic density and accident rates are increasing; moreover, the passage of vessels in ports or channels without a marine pilot is hazardous. Such hazards can be observed in the Istanbul Strait, as pilotage is not compulsory for transit vessels under some conditions. There is a drastic difference between accident rates of vessels under pilotage and those without pilotage. The unusual characteristics of the Bosporus, the Istanbul Strait, and its climate, coupled with the failure to request pilotage in this treacherous waterway, have resulted in more than 200 accidents in the past decade [51]. Therefore, the demand for marine pilots navigating ships through hazardous or busy waters, such as harbors, is increasing globally. Even though the overall number of vessels calling Busan port decreased between 2010 and 2020, the number of piloted vessels increased from 24,921 to 30,249 [41].

Received 3 March 2023; revised 26 September 2023; accepted 27 September 2023.
Available online 15 December 2023

* Corresponding author.
E-mail address: psw6745@kmi.re.kr (S. Park).

https://doi.org/10.51400/2709-6998.2709

Vessel pilotage is a centuries-old profession. The term "pilot" first appeared as early as the 6th century BC in Ezekiel's book; the book described the pilot as the "guide" of the ship [17]. Marine pilots take the vessel through a highly treacherous part of the ship's voyage and guide her to safely berth, unberth, and pass narrow channels. They use a boat and pilot ladder in conjunction to embark on the vessel. However, embarkation is only sometimes safe and secure owing to ship defects, mistakes, or environmental conditions. Pilots conduct the transfer operation multiple times daily, increasing their risk probability. Hence, early detection of risk factors and applying preventive actions are vital.

Using a prediction model for maritime pilot occupational accidents is essential because it allows for early detection of risk factors and the implementation of preventive actions. By analyzing previous accidents and identifying the associated circumstances and safety factors, the model can provide valuable insights and recommendations to prevent future accidents and ensure the safety of marine pilots.

Therefore, this study aims to develop a marine pilot occupational accident prediction model using a methodology including a decision tree, RF (Random Forest), and explainable ML. The purpose of the model is to determine the circumstances and safety factors of previous accidents to provide a basis for making recommendations to prevent further marine pilot casualties and accidents from occurring in the future.

2. Literature Review

2.1. Marine pilot occupational accidents and ML application in the marine accident domain

Occupational health and safety are vital to practical work [25]. The workplace's physical conditions and mental demands majorly influence workers' performance. Occupational accidents have significant human, social, and economic costs (Leigh et al., 2004). Therefore, workplace safety is paramount to eliminating occupational accidents.

Merchant shipping is recognized as an occupation with a high rate of fatalities [12] caused by maritime disasters and occupational accidents. Detailed information exists for the Danish merchant fleet, showing a fatality rate over ten times that of shore-based industries.

Marine pilots have amphibious duties that expose them to shore, sea, and transition hazards. They must judge the correct timing and wave crest to jump at the correct moment successfully. This role is not described in their training literature or contract but must be undertaken as there is no alternative transfer method [35].

Previous research [47] shows that marine pilots are more prone to accidents while using a pilot ladder, with a risk of life-threatening injuries [11]. Hazards are due to transfer arrangement (equipment) failure [26] or both lateral and vertical movement of the two vessels, where the swell causes the marine pilot to misjudge their steps and fall from a potentially dangerous height.

Less severe and rare pilot transfer accidents also occur during the following stages: car transfer from pilot office to port, pier walking, pier to cutter transfer, cutter navigation, cutter to pilot/accommodation ladder climbing, traversing deck to the bridge, pier to gangway climbing (at port), during helicopter flight, or winching and landing [13].

Traditionally, marine pilots use a pilot boat and ladder in the open sea and the gangway alongside the ship. H helicopter landing and winching [24] are rarely used in specific ports in Australia, the United States, France, Germany, and South Africa. However, they are less common globally than the pilot boat and ladder method. Even though marine pilots have a vital role in the safety of vessels, ports, and channels, their safety has received little attention. The maritime industry has been modernized with many new inventions and automated systems; however, the method of marine pilot transfer remains unchanged and relatively archaic. Thus, marine pilot occupational accidents still occur. Therefore, this study aims to develop a new model to predict marine pilot occupational accidents during their transfer to or from ships. It uses an ML algorithm to determine the most influential risk factors.

ML is a significant technological development, influencing the adoption of new habits and behaviors in society. Industrial development and cost-saving demands require repetitive tasks to be carried out by machines with less effort than human operation. Robotics, intelligent cities, smart homes, and autonomously driven vessels are already integrated into society. ML has several subdomains, like Supervised Learning, Unsupervised learning, and Deep Learning, and is used to identify patterns and classify massive datasets [36]. It refers to techniques aimed at programming computers to learn from experience.

The RF [9] algorithm is based on the idea of an ensemble of various decision trees [14,32], aggregating their predictions. This achieves better performance and provides more robust predictions, but at the cost of interpretability, as humans cannot easily comprehend multiple decision trees. RFs are widely used in fields such as finance [58], healthcare [37], and e-commerce. For example, banks use RFs for client credibility evaluation, credit card fraud detection, and options pricing [33].

Considering the ability of RF to predict events, the benefits of prediction are beginning to be explored in other industries. Logistically, if a factor is predictable, then it is also preventable. Therefore, RFs can be used to predict accidents and mitigate contributing factors. Road traffic accident studies show that driver experience, light conditions, age, car type, and annual car service contribute to accident severity [55].

In the maritime industry also, RF usage is expected. Marine applications of RF include environmental and ecological research, meteorological data modeling [28], ship collision prediction and prevention, and identifying human error in accident occurrence [48].

2.2. Explainable ML (machine learning)

In modern ML, owing to increased volumes of data and computing power, practitioners use complex models with a wide variety of parameters to improve predictive performance. These are often called “black-box” approaches, such as RFs and deep neural networks, where predictions cannot be explicitly interpreted. However, it is usually possible to obtain model-specific variable importance measurements that indicate the overall most influential variables in a dataset, such as split gain in the case of an RF model [38]. Black-box approaches differ from interpretable by-design algorithms, such as decision trees and linear regression, which provide explicit reasoning [45].

Various explainability methods have been developed to overcome the interpretability performance tradeoff in ML applications [22], allowing new insights into the models' reasoning [4,44]. Model-agnostic explanations that work with any black-box approach [39], such as permutational variable Importance [18] and partial dependence profiles [19], show a global-level overview of the model. Breakdown profiles [6] and Shapley values [34] provide a crucial interpretation of the model's predictions, which is particularly useful in responsible decision-making [21]. It is good practice to compare the model-agnostic variable Importance with the native model-specific measure to check for a consistent result. A detailed description of explainable ML methods and tools used in this study is available in previous publications [7,29,38].

Explainable ML was recently used in marine research to support monitoring of the marine engine's condition [30] and classifying species [23,36]. However, to the authors' knowledge, explainable ML has yet to previously be used to explain models for predicting marine pilot occupational accidents. Data scarcity [1] was one of the main struggles for marine domain [3] research. However, with new technologies on board, more data becomes available. The new system will produce a massive amount of data. For this reason, ML will be better than other methodologies in dealing with big data.

3. Methodology

This study aims to predict marine pilot occupational accidents during their transfer to/from the ship, using decision tree-based ML algorithms, and describe the ranking of contributing factors with explainable ML methods. First, a multi-source dataset with metadata of 406 accidents was created from reports, as explained later in Section 3.1. Then, the decision tree and RF models were trained on the data to predict the target outcome. Both models were tuned using a parameter search to achieve the best possible performance. The research flow is shown In Fig. 1. The data and code used in this study are available at https://github.com/hbaniecki/marine-pilot-occupational-accidents.

After training the data, bootstrapping was applied to the split algorithm. Bootstrapping [8] is a statistical sampling method based on reusing samples, where each observation can be repeated more than once. This method allows several equally sized sampling groups to achieve more effective results [46]. In mathematical notation, bootstrapping can be expressed as follows:

$$\hat{f}_{bag}(x)=\frac{1}{B}\sum_{b=1}^{B}f^{*}b(x) \tag{1}$$

A set of observations, X = …, is defined with responses Y = ,… with B bagging repeats, in which a random set is sampled with replacement.

For $b = 1, \ldots,B$:

1. n training examples from X and Y are sampled, with replacement, defined as defined as X_b, Y_b.
2. A classification tree is trained using f_b on X_b, Y_b.,.

The probability of not selecting a row in a random sample is as follows:

$$\frac{\mathrm{N}-1}{\mathrm{N}} \tag{2}$$

Where N is the number of rows in the training dataset. Using sampling with replacement, the

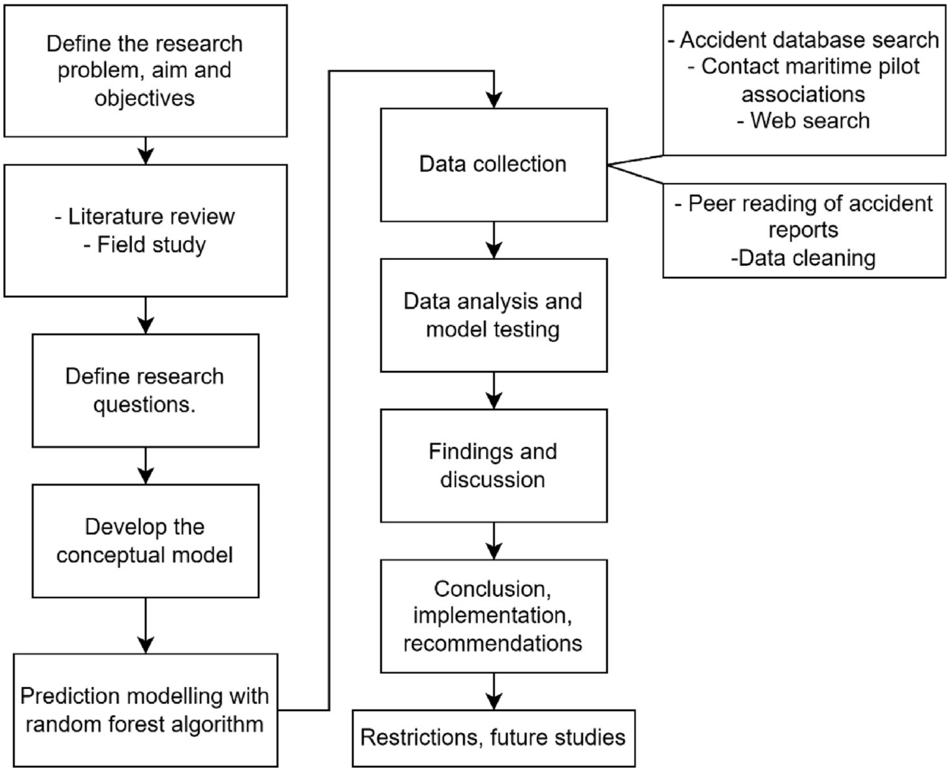
Define the research problem, aim and objectives
- Literature review
- Field study
Define research questions.
Develop the conceptual model
Prediction modelling with random forest algorithm
Data collection
- Accident database search
- Contact maritime pilot associations
- Web search
- Peer reading of accident reports
-Data cleaning
Data analysis and model testing
Findings and discussion
Conclusion, implementation, recommendations
Restrictions, future studies

Fig. 1. Research flow diagram.

probability of not selecting N rows from randomly sampling rows is as follows:

$$\left(\frac{N-1}{N}\right)^{N} \tag{3}$$

Which is the limit of N becoming equal to:

$$\lim_{N\to\infty}\left(1-\frac{1}{N}\right)^{N}=0.368 \tag{4}$$

In ML, the decision-making process is called a decision tree. As mentioned earlier, the process starts with a node and continues splitting until it reaches a leaf. The leaf is the endpoint of the process where no further splitting occurs. The first node is called the root node and remains at the top of the diagram. Branch trees are generated during processing, consisting of internal and leaf nodes.

RF creates multiple decision trees from the dataset in a pool and acquires decisions from trees. Although a more significant number of trees tends to provide more stable results after a certain level, there will be no significant effect on the result. There are limitations to increased computation time and costs when using big data [40]. In this study, an optimal number of trees was selected by conducting multiple tests with different sizes of tree numbers.

Finally, an interpretation of the tree decisions and an explanation of the RF predictions were visualized, discussed, and contrasted with maritime domain knowledge.

3.1. Dataset

This study is based on reports of perceived accidents that occurred during marine pilot transfers at various stages. The accident reports were acquired from global sources accessed between October and December 2020. After data processing, 406 reports of accidents, incidents, near misses, and non-compliance occurred from 1993 to 2019. These reports were acquired from maritime institutions including MAIB(Marine Accident Investigation Branch) IMPA (International Marine Pilot Organization), UKMPA (United Kingdom Marine Pilot Association), EMPA (European Marine Pilot Association), KMPA (Korean Marine Pilot Association), MARS (Marine Accident Reporting Scheme), IMCA (International Marine Contractors Association), Marshall Island Flag, Malta Flag, Bahamas Flag, TSCB(Transportation Safety Board of Canada), Hong Kong, ATBS(Australian

Transport Safety Bureau), AMSA (Australian Maritime Safety Authority), DMAIB(Danish Maritime Accident Investigation Board), TAIC(Transport Accident Investigation Commission). Table 1 presents example observations from the dataset. The target variable was constructed from the categorization of UDE (undesired events). The event outcome resulting in an accident was encoded as “1”, whereas incidents, non-compliance, nonconformance, and near-miss events were encoded as “0”.

Factors affecting accidents were categorized into environmental factors (e.g., sea state, wind force, visibility, month, and time of day) and workplace factors (e.g., accident location, ship type, age, dynamic status, geographical position, gross ton, and length).

Sea and wind states have a significant effect on ship-boat motion. Increased rolling of both vehicles creates an unstable platform, an unsafe condition during transfer. Swells also cause a potentially dangerous sea state. However, the swell variable was not included in this study owing to a lack of available data.

In impaired visibility [20], there is no direct effect during marine pilot transfer; however, the pilot cutter might be involved in a collision while marine pilots are on board under low visibility navigation.

Time of day is also crucial because marine pilots work irregular shifts. Evening and night shift workers are exposed to more occupational accidents than day shift workers. Under impaired weather conditions [15], darkness risk factors are increased.

Ship types are varied and have unique construction features, which might restrict the design and ergonomics of the marine pilot access point and passageway. For example, tanker decks are obstructed due to cargo tank frames and fish plates. In small vessels, cargo hold frames are too close to the ship side; this does not allow the ship to deploy the pilot ladder properly. For example, this is when an overboard pipe outlet is located on the same frame as the pilot access point.

Ship dynamic status describes the mother ship's voyage status. Marine pilots may join the vessel while the mother ship is underway, at anchor, or at the pier, which is a crucial factor to the mother ship-pilot cutter relative motion. The ship's geographical position affects the mother and pilot cutter's relative motion. Out of port limits, both vessels are exposed to environmental conditions, which increase the possibility of an accident. Finally, marine pilots pass through various stages to embark on the vessel. Each section has a specific accident risk with a different injury severity level.

3.2. Data analysis

The acquired dataset included 500 unique accident reports. Figs. 2 and 3 show the distribution of variables concerning the target outcome. Variables “age of the vessel,” “gross ton,” “length overall,” “sea state,” and “wind force” had several missing values, which are further analyzed in Fig. 4. The target outcome was divided into 315 observations of class 1 (77.5 %) and 91 observations of class 0 (22.5 %).

3.3. Model analysis

The two models compared in this study were a decision tree and an RF, which is an ensemble of multiple decision trees. Decision trees are a straightforward approach to predicting a target outcome and rely on sequentially creating decision rules that best split the observations into, in this case, two classes. The main strength of this approach is its built-in interpretability and simplicity for stakeholders working with automated predictions. Moreover, decision trees can operate with a dataset containing missing values, as in this study. RF is a more complex algorithm that creates many decision trees fitted to the same predictive tasks. These trees then predict each observation.

Table 1. Example observations from the dataset

VAR ID	Target	Accident location	Age of vessel [years]	Dynamic status	Gross ton	Length overall [meters]
99	Accident	Pilot/ Combination ladder	2	Underway	(1k,3k]	(90,108]
109	Accident		34	Moored	(7k,10k]	(140,160]
235	Non-accident		8	Underway	(7k,10k]	(70,90]
Month	**Part of day**	**Sea state [wave height]**	**Ship position**	**Ship type**	**Visibility**	**Wind force [Beaufort]**
3	Sunset/ twilight	4	Port area	Tanker	Good	5
3	Daytime	4	Pier	Miscellaneous	Poor	2
1	Daytime	3	Port area	Tanker	Moderate	3

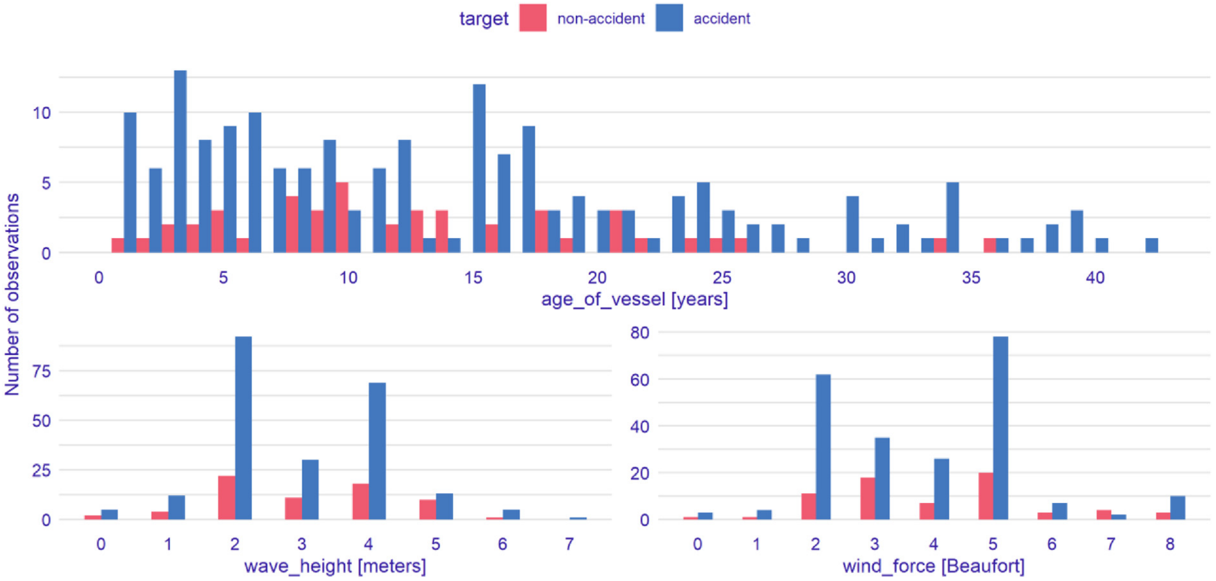

Fig. 2. Distribution of four numerical variables in the dataset.

The ensemble approach has proven robust and increases ML performance in practice. Although RFs operate with missing values in the dataset,[1] Both their main strength and weaknesses are the model complexity. A decision from an RF cannot be directly interpreted, as explainable ML approaches are required to analyze such models.

Experiments were conducted on the whole data set described in the previous section. The data dimensions were small; hence, five-fold cross-validation was used to assess the models at each step unless otherwise specified. The performance measures of interest were conventional classification metrics: ACC (Accuracy), AUC (the area under the receiver operating characteristic curve), and AUPRC (the area under the precision–recall curve). Observations were weighted based on the fraction of a given class in the target variable, which is crucial to interpret the results effectively. Multiple parameters of interest can describe decision trees. This study focused on maximum tree depth, the minimum number of observations needed for the algorithm to attempt another split, and the complexity of the tree that restricts weak splits. RFs were additionally described by three more parameters in this study: the number of trees, the fraction of randomly sampled rows for each tree, and the fraction of randomly sampled columns for each tree. Decision tree and RF models with default algorithm parameters were fitted to the data as a baseline, and then a grid search algorithm was used to tune the models to improve their performance. Details of specific parameter values are included in Appendix A.

The Importance of factors associated with accidents was assessed using two measures:

1. A tree-specific measure that evaluated the gain of model performance for all the splits (decision rules) in the tree using a given variable. For RFs, the measure was aggregated for all the trees.
2. A model-agnostic measure that evaluated the loss of model performance when a contribution of a given variable from the data was not active. Permutational variable Importance was used to simulate such an effect.

Additional explanations, in the form of partial dependence plots, breakdowns, and Shapley values, were used to interpret the predictions of the RF model. All the mentioned methods are described in detail [7,38].

3.4. Software

The analysis was performed using the R language for statistical computing v4.1 [43] with additional packages: ggplot2 [54] and visdat [50] for data analysis, rpart [49] and caret [31] for ML, and DALEX [5] and modelStudio [2] for model analysis. The data and code used in this study are openly available on GitHub at [https://github.com/hbaniecki/marine-pilot-occupational-accidents].

[1] Decision trees can be used as weak learners, which operate with missing values in the dataset.

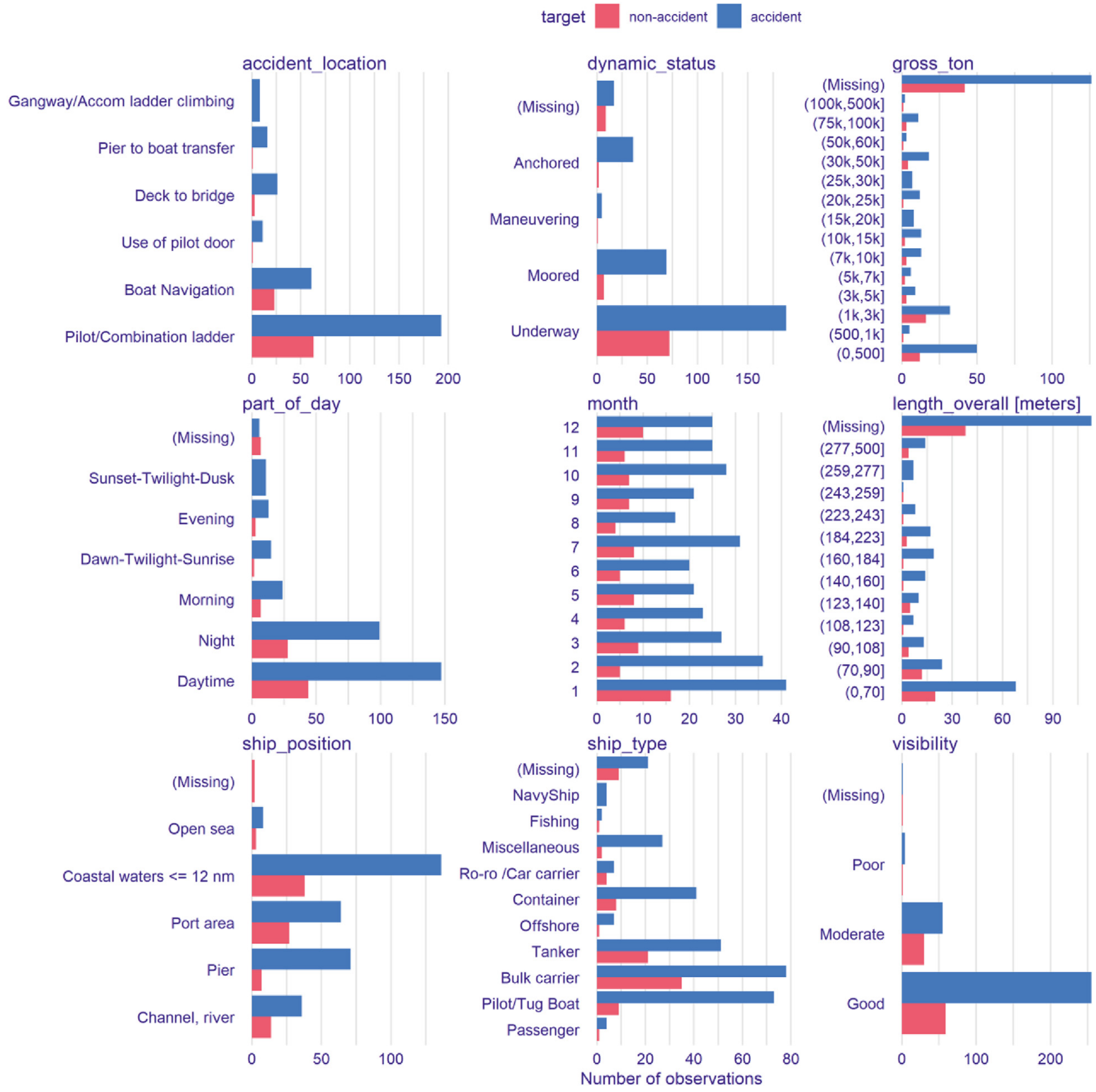

Fig. 3. Distribution of eight categorical variables in the dataset.

4. Results

Table 2 shows the performance of the baseline and tuned models, where the mean and standard deviation were estimated from the five cross-validation runs. The tuned RF model achieved the best scores, although the grid parameter search process did not meaningfully improve the baseline. It was crucial to check for better model parameters as an ablation study. In this case, both AUC and AUPRC were used to compare the models, while ACC determined the context; hence, the tuned decision tree was slightly better. Fig. 5 shows the performance curves. A clear improvement was seen for using an RF over a decision tree, as an ensemble of trees consistently offered more accurate predictions. Each of the five cross-validation runs is a distinct line.

The study's main result explains the complex RF model that distinguished between accidents and other UDE (Un Desired Events) outcomes. Fig. 6 shows both variable importance measures for the tuned models. The Importance of factors associated with the accidents measured with a performance drop (left) and split gain (right) are displayed. The most critical variables yielded higher measured values (top).

The focus is on the split gain importance of the tuned RF (right, red), whereas the decision tree and

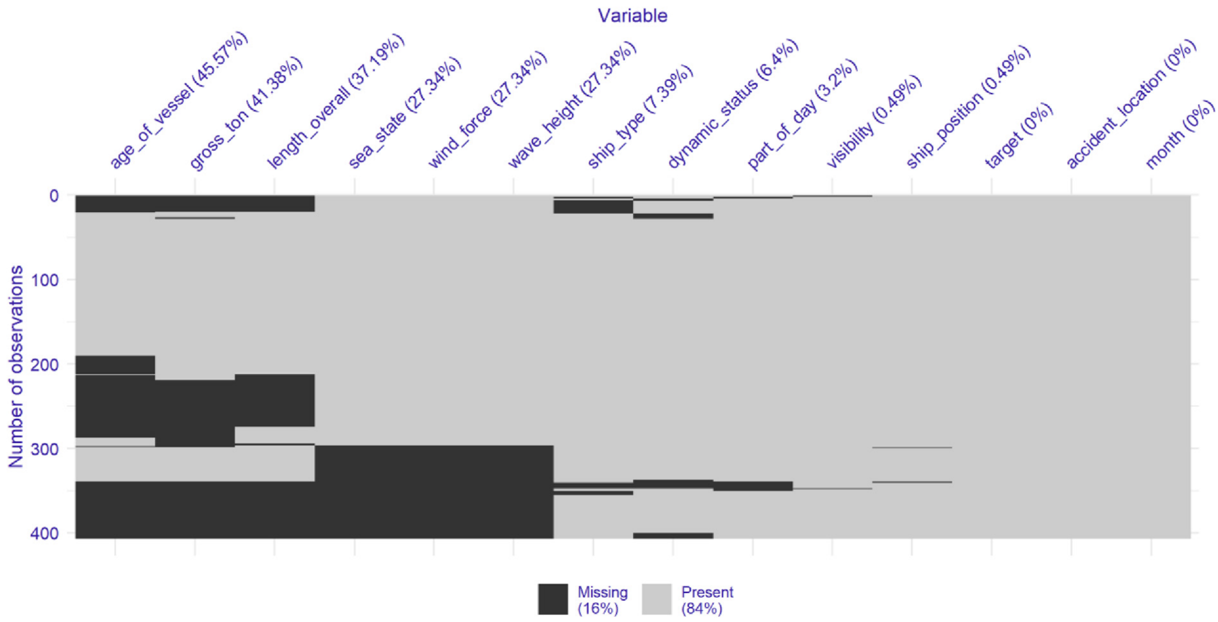

Fig. 4. Distribution of missing values while black ones are missing and grey ones are available data indication.

Table 2. Performance measures for the baseline and tuned models

Model	ACC	AUC	AUPRC
Decision tree (baseline)	0.65 ± 0.05	0.60 ± 0.06	0.85 ± 0.03
Decision tree (tuned)	0.62 ± 0.05	0.63 ± 0.09	0.86 ± 0.04
Random forest (baseline)	0.75 ± 0.06	0.71 ± 0.05	0.91 ± 0.03
Random forest (tuned)	**0.76 ± 0.05**	**0.72 ± 0.05**	**0.91 ± 0.02**

permutational Importance are provided for a broader context. The essential variables are seasonality (month), ship type (length overall, gross ton), and dynamic status. Consistently, workplace factors were important in the permutational measure, apart from the variable "month," as its value did not lower the model's performance. Variables not crucial for the model fitted to this specific dataset and task were mainly the environmental factors (e.g., sea state, wind force, and time of day).

For interpretation purposes, a single decision tree and RF model have been trained on the whole dataset best to capture the variable dependence on

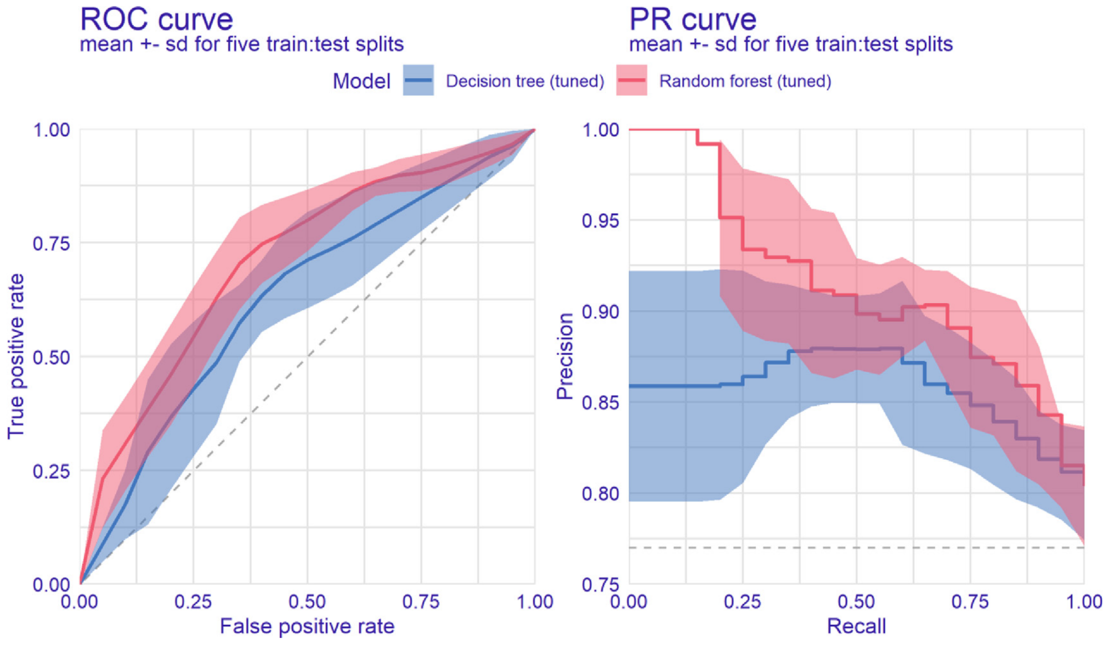

Fig. 5. ROC and PR curves for the final models.

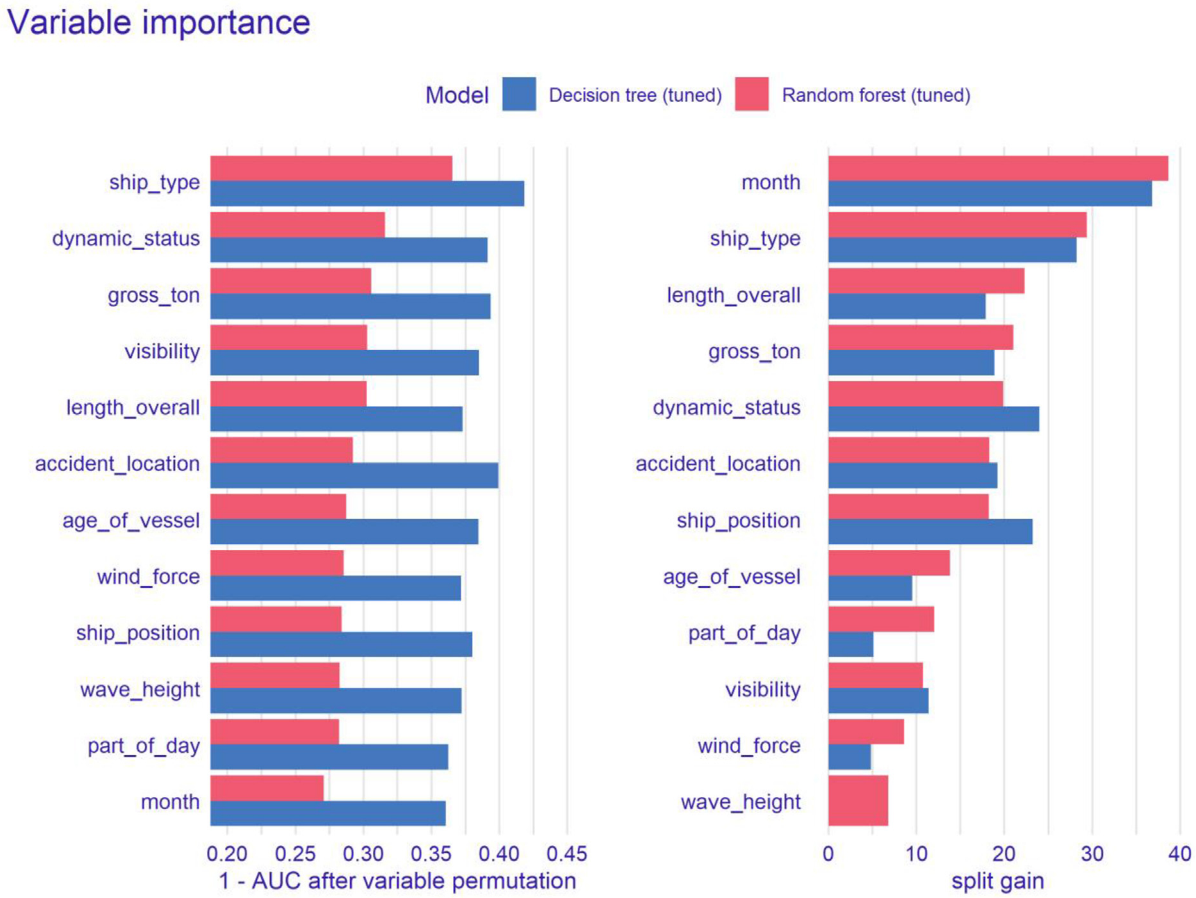

Fig. 6. Importance of factors associated with accidents.

the target outcome. This was to allow for knowledge discovery and not to evaluate the models' performance. Fig. 7 visualizes an example decision tree, which supports the finding that workplace factors are more important than environmental factors in predicting accidents. Values in the leaves (at the bottom) are observation counts weighted by the class fraction in the dataset. Labels of the gross ton variable were truncated for clarity but imitate the monotonicity of values.

Fig. 8 shows a partial dependence plot of the essential variables from an example RF model, which provides insight into how the variables affect average prediction in a global sense.

Additionally, the RF prediction for a given observation can be explained using the breakdown method, which shows the contribution of variables. Fig. 9 presents explanations for three different observations from the dataset (from Table 2), where four of the essential variables are highlighted and other factors are truncated. Complementary explanations using the Shapley value method are reported in Appendix B. The intercept is a mean prediction that serves as a baseline.

5. Application

This chapter applied the developed model to the sea area near Busan Port in South Korea. Busan Port is a complex port with the most extensive seaborne trade in Korea, so there is much pilotage work [42]. According to [27]; at least three days of maritime traffic data is required to confirm the maritime traffic characteristics of the target sea area for one year. In addition [57], said that at least seven days are required for maritime traffic surveys, considering the day-of-week index. Therefore, to apply the developed model, September 01, 2019, ~7th Sept. maritime traffic survey. To minimize the impact of pilot work due to COVID-19, we used data from 2019. The sea area near Busan Port is divided into inbound and outbound; accordingly, the pilot's embarkation and disembarkation areas are divided[10]. In bad weather, it is

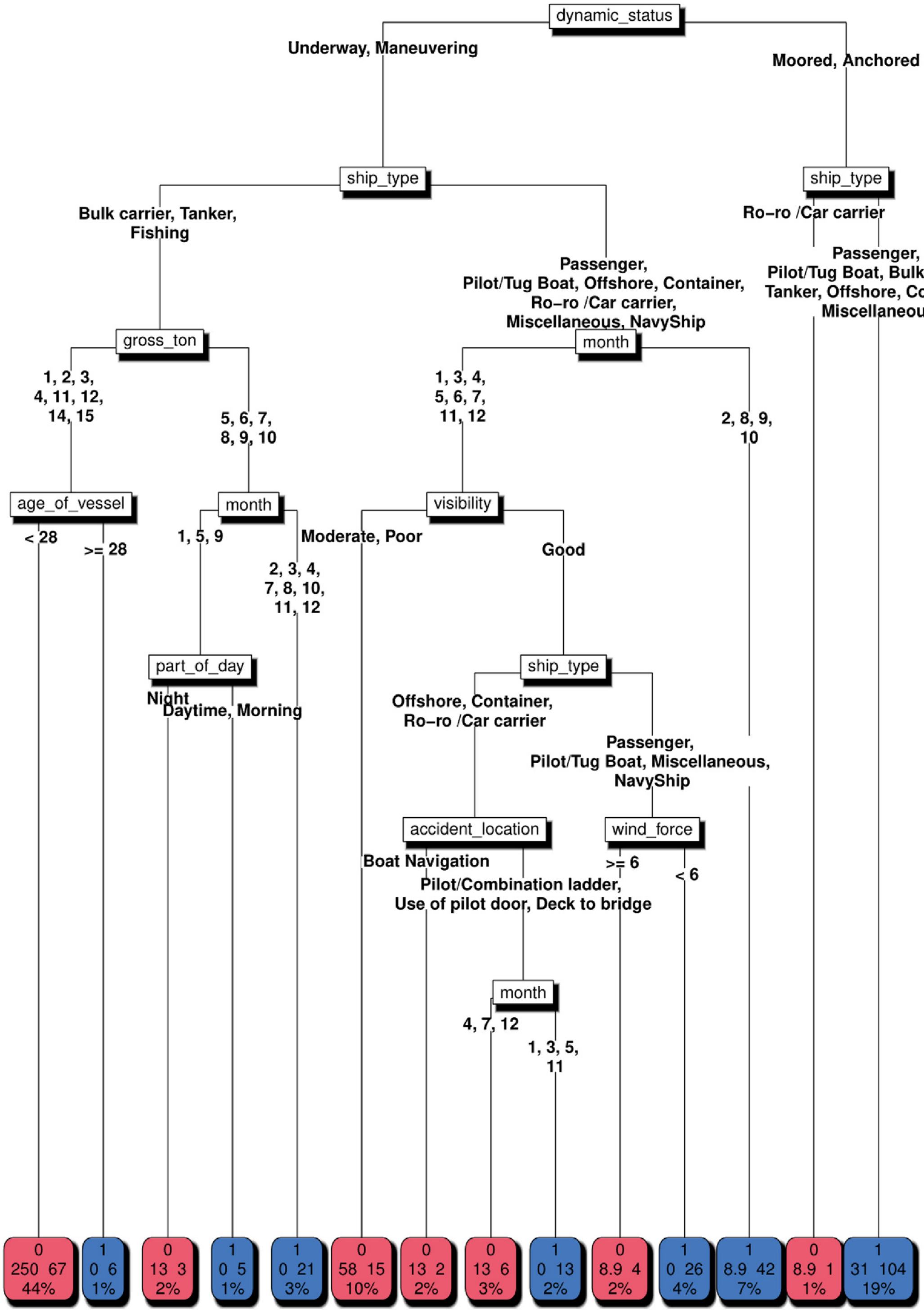

Fig. 7. Visualization of the decision tree trained on a whole dataset.

designated to use the pilot boarding and leaving points on the inner route of the breakwater (BHPA, 2023).

Fig. 10 shows the location of the ship where the pilot boarded and disembarked for seven days and the probability of a pilot accident using the model at that time. The pilot accident probability means the probability of being classified as a pilot accident when using the model. There were 237 cases in which the pilot boarded and disembarked for seven days; among them, there were 31 cases with a probability of 0.5 or higher.

According to the Importance of factors associated with an accident in Fig. 5, it was confirmed that

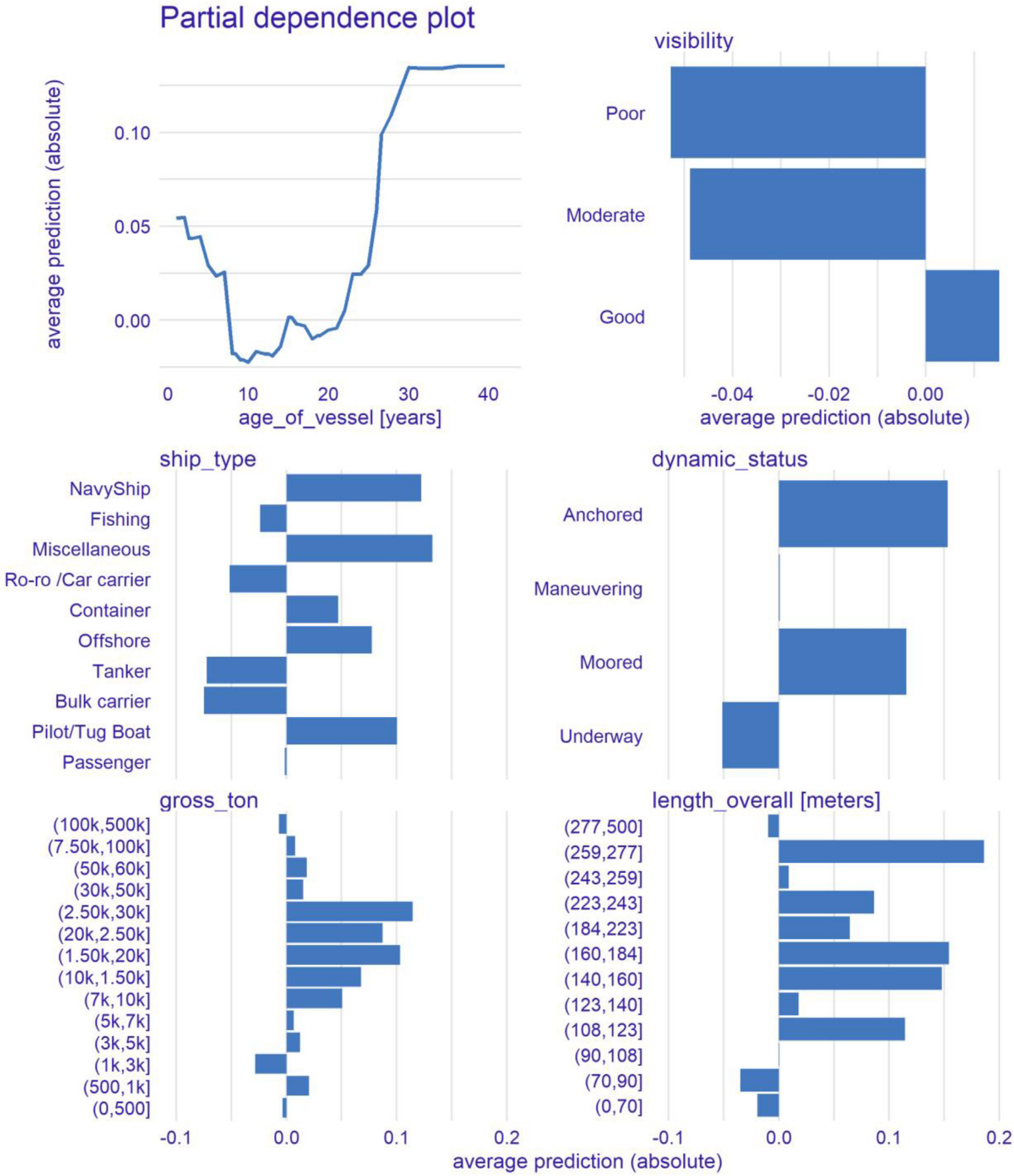

Fig. 8. Partial dependence of the selected variables in the RF model.

workplace factors had a more significant impact on accidents than environmental factors such as wind force and wave height. Among the 7-day data, the average gross tonnage of cases with a high probability of accident occurrence was 11,903.2 Tons, and the average ship length was 144.8 m.

As a result of applying the random forest model to the Busan Port 7-day traffic survey, it was confirmed that the probability distribution follows a normal distribution (R-square: 0.96). Fig. 11 is the distribution of the probability of a pilot accident using the 7-day data of Busan Port. The average accident

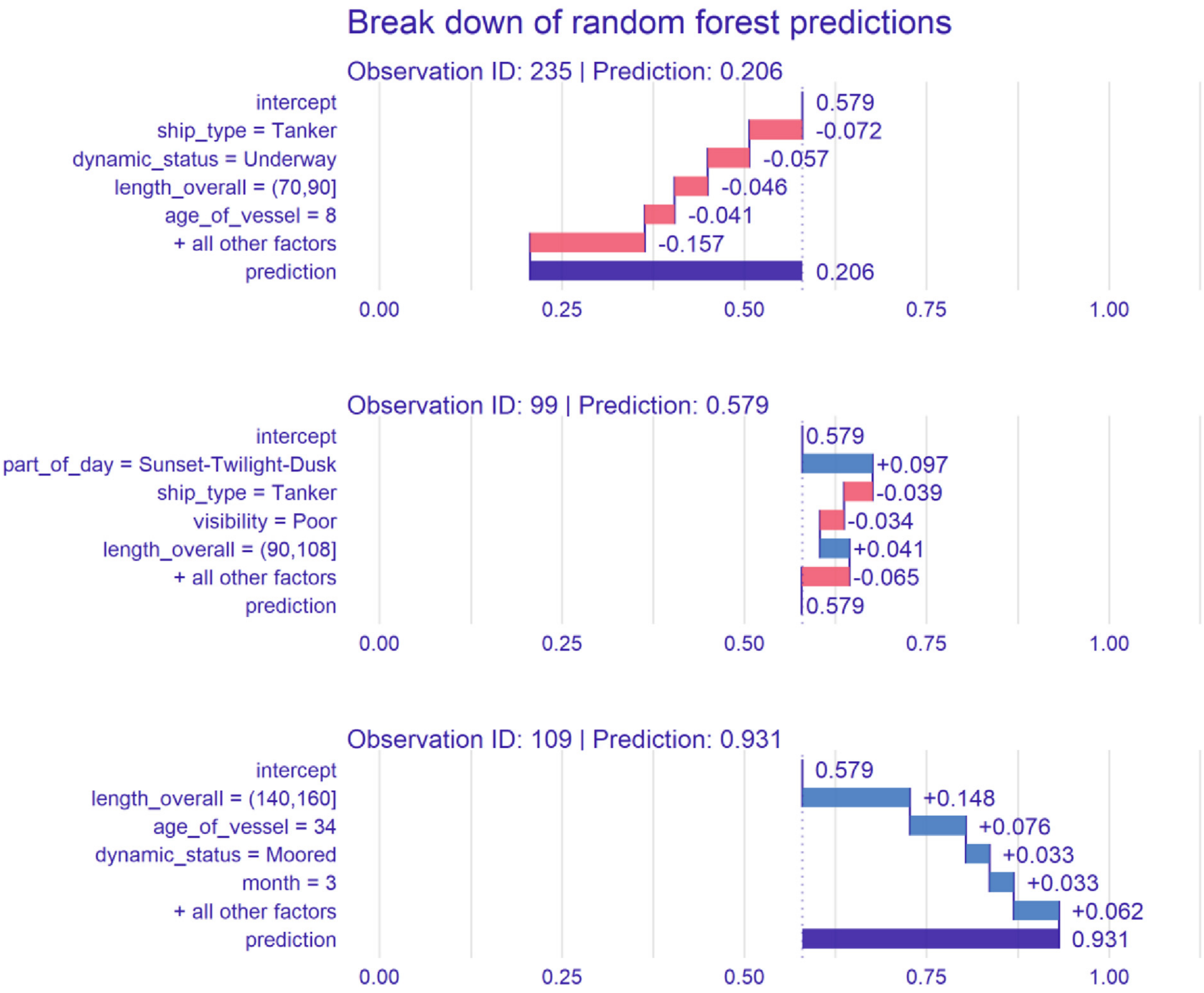

Fig. 9. Breakdown of the RF model's predictions for three distinct observations from the dataset.

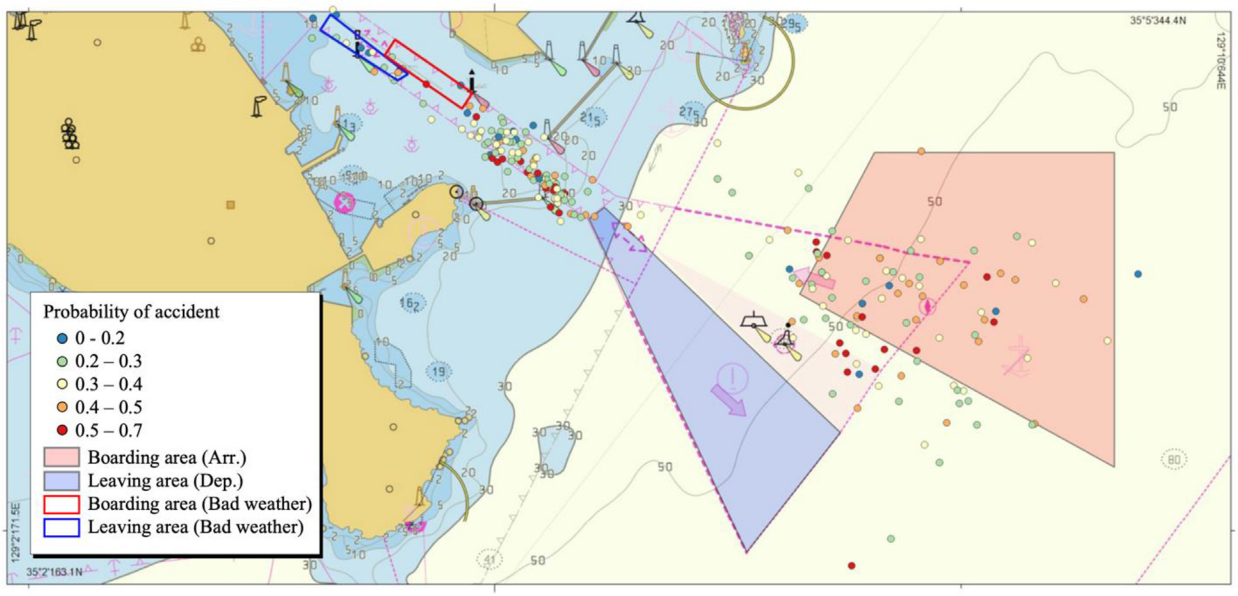

Fig. 10. Location of the ship where the pilot boarded and disembarked for 7 days and the probability of pilot accident using the model at that time.

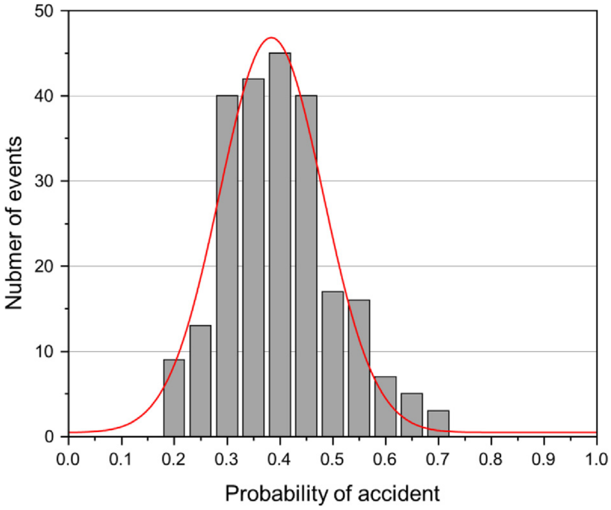

Fig. 11. The distribution of the probability of a pilot accident using the 7-day data of Busan Port.

probability, according to the model, was 38.345 %, and it could be expressed as the following equation.

$$y = y_0 + \frac{A}{w\sqrt{\frac{\pi}{2}}}\exp\left(-2\left(\frac{x - x_c}{w}\right)^2\right) \quad (5)$$

where:

$y_0 = 0.4621$

$x_c = 0.38345$

$w = 0.19551$

$A = 11.36493$

According to previous studies, seven days of maritime traffic data can replace one year of maritime traffic. Hence, on average, the risk of pilot accidents in Busan Port for one year is about 38 %. Therefore, if the model is used, the risk level can be measured using the data of piloting ships when entering and leaving the port. The port authorities can use this result to make policy decisions, such as selecting areas that require priority response. However, since this application only used the traffic data of Busan Port for seven days, it is necessary to improve the Accuracy by using data for a more extended period. In addition, since the application of the maritime traffic-based model only considers environmental factors, it is necessary to analyze the human factors in the future. In other words, the current application is the probability of leading to an accident in the environment if there is a human factor failure.

6. Conclusion

This study proposed a prediction model for advanced accident prediction during marine pilot transfer. The study's novelty is the application of artificial intelligence for accident predictions. 13 pre-accident factors were determined, with decision trees and RFs trained to distinguish marine pilot accidents and non-accident Variable Importance. Using only decision trees and RFs is insufficient for interpreting the decisions of RF models, thus necessitating explainable ML approaches to analyze the model. Significantly, workplace factors were more important than environmental factors. Unimportant variables were the sea state, wind force, and time of day outside the operators' control.

The model developed through this study was applied using the maritime traffic data of Busan Port, which is the most complex in South Korea. As a result of the application, it was found that the average probability of a pilot accident in Busan Port was about 38.3 %, and it was confirmed that the distribution of accident probability followed a normal distribution. In other words, by using the model of this study, it is possible to derive the probability of a pilot accident in a specific port, and it is judged that port authorities will be able to make policy judgments using the data.

Overall, human factors have a more significant influence on undesired event occurrence. Thus, the adequacy of pilot boarding appliances and organization plays a vital role in preventing accidents. Constant monitoring of equipment on both the mother ship and the pilot cutter would be an efficient method of preventing accidents. Furthermore, organizations should ensure marine pilots are always alerted to prevent complacency.

Moreover, the model can be rerun after collecting an increased dataset with further variables, such as Ship Flag Authority, Classification Societies, ship owner and manager, P&I club, which strong inspection regimes affect the safe managing of ships, and additionally minimum safety manning and the actual number of crew, relative height from boat to ship deck, age of the marine pilot working hours, drug and alcohol use, body mass index, and years on duty.

6.1. Implementation

The findings of this study present essential theoretical and managerial contributions for preventing marine pilot occupational accidents and mitigating the impact of accidents:

1) With the proposed model, pre-accident factors can be determined, and their relative Importance provides a supportive database for organizations to determine and implement working conditions for marine pilots. Factors such as maximum weather permits, wave height, and precautions for ship types should be considered.
2) A reporting system for all pilot organizations globally should be implemented). Respective pilot associations or responsible authorities can conduct report gathering and analysis. This data can then be shared with the International Marine Pilots' Association as a general dataset for future rule and regulation implementation, efficacy control, and developments for arrangement

6.2. Limitations and future research

In this study, reports were collected from maritime institutions globally. However, many organizations need a robust database for accident records. On the other hand, some organizations are unwilling to share information; there needs to be a reporting system, or the existing system needs to record all facts in a minimum criterion, which is required to predict future events using statistical methods.

This study's acquired dataset included a unique gathering of 500 unique accident reports, which could be extended in future research to perform a larger-scale analysis. Furthermore, the low number of reports accounting for other undesired events (incidents, non-compliance, nonconformance, and near-misses) is the main limitation of this study. Nevertheless, maritime stakeholders should consider the outcomes a crucial reference point. Enlarging the dataset with a higher volume and quality of reports is required to achieve a more reliable ML analysis in the future.

In this study, there is a limit to the application to Busan Port, the representative port of South Korea. In the future, it will be applied to other ports to derive the accident probability and use it to suggest improvement measures to reduce the risk related to pilotage accidents in each port.

Other factors that may cause accidents are age, body mass index, alcohol/drug use, medication, fatigue, historical illness, and a person's injury. Due to the scarcity of data, these variables cannot be evaluated, which is a limitation of this study. In future studies, marine pilot accident root causes and injury statistics should be evaluated using fault and event tree methodology. In addition, a vessel traffic analysis based on real-time AIS (Automatic Identification System) data should be conducted to determine the safety of pilots embarking/disembarking ships through a minimum safe distance calculation.

Author contributions

Gokhan Camliyurt: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Validation; Visualization; Roles/Writing - original draft. **Youngsoo Park:** Conceptualization; Investigation; Methodology; Resources; Supervision; Validation; Roles/Writing - original draft; Writing - review & editing. **Daewon Kim:** Writing – review & editing.

Won-Sik Kang: *Writing - review & editing.* **Sangwon Park:** *Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Resources; Software; Validation; Visualization; Roles/Writing - original draft.*

Conflicts of interest

None.

Appendix A

The best parameters for both decision tree and RF algorithms were selected using a grid search. The parameter search was conducted with a five-fold cross-validation on the whole dataset. This was performed owing to low data availability, and for the best model fit, where the explanations and interpretations were the most accurate, rather than relying on prediction accuracy.

The parameter grid for the decision tree involved the following: the maximum tree depth from the set {3, 4, 5, 6, 7, 8}, the minimum number of observations that must exist in a node for a split to be attempted from the set {8, 10, 13, 15, 17, 20, 24}, and the complexity parameter from the set {0.1, 0.05, 0.03, 0.01, 0.005, 0.001}.

The parameter grid for the RF involved the following: the number of trees from the set {100, 200, 300}, the fraction of observations from the set {0.7, 0.8}, the fraction of variables from the set {0.6, 0.7, 0.8}, the maximum tree depth from the set {3, 5, 7}, the minimum number of observations that must exist in a node for a split to be attempted from the set {10, 15, 20}, and the complexity parameter from the set {0.05, 0.01, 0.005}.

The best parameter sets determined based on the AUC measure were {4, 13, 0.001} and {200, 0.7, 0.6, 7, 10, 0.01} for the decision tree and RF, respectively.

Appendix B

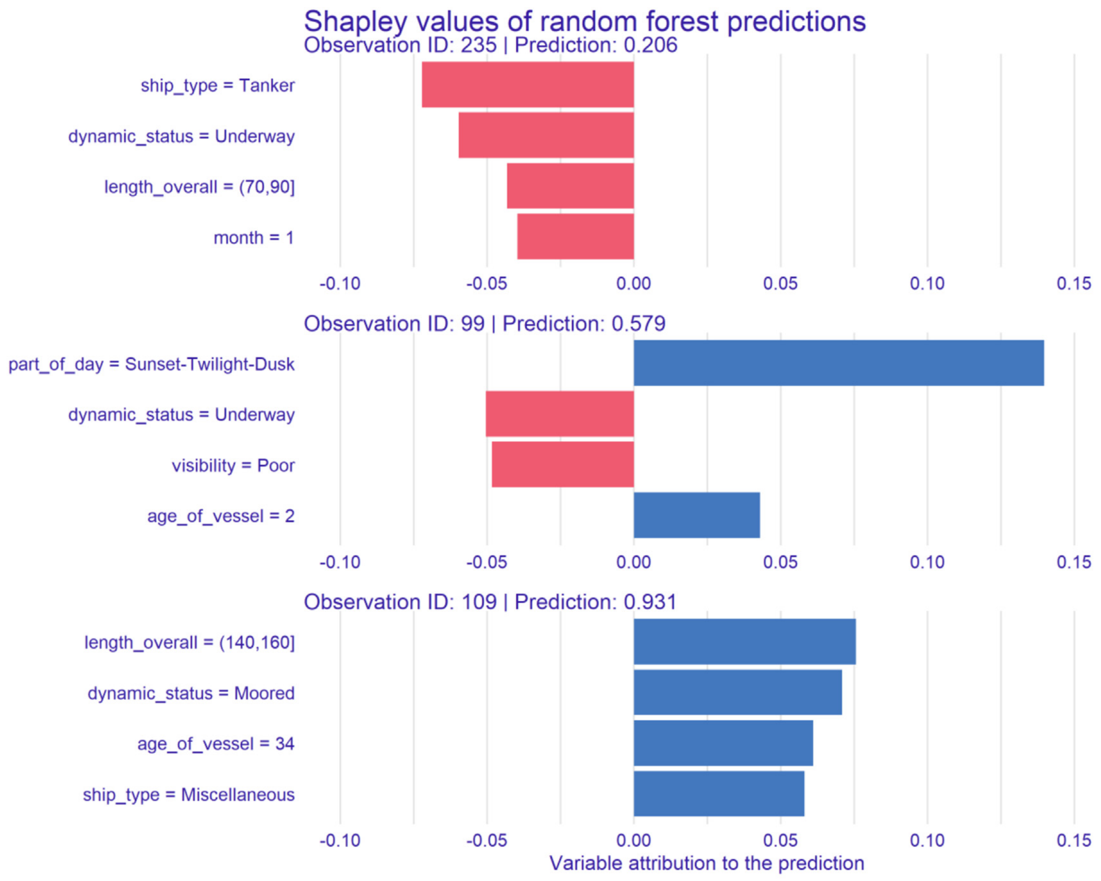

Fig. B1. Shapley values the RF model's predictions for three distinct observations from the dataset. These are absolute values (not relative); hence, they start at zero.

References

[1] Alzubaidi L, Bai J, Al-Sabaawi A, Santamaría J, Albahri AS, Al-dabbagh BSN, Gu Y. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. J Big Data 2023;10(1):46. https://doi.org/10.1186/s40537-023-00727-2.

[2] Baniecki H, Biecek P. modelStudio: Interactive Studio with Explanations for ML Predictive Models. J Open Source Softw 2019;4(43):1798. https://doi.org/10.21105/joss.01798.

[3] Becktor J, Schöller FET, Boukas E, Blanke M, Nalpantidis L. Bolstering Maritime Object Detection with Synthetic Data. IFAC-PapersOnLine 2022;55(31):64–9. https://doi.org/10.1016/J.IFACOL.2022.10.410.

[4] Bhatt U, Xiang A, Sharma S, Weller A, Taly A, Jia Y, Ghosh J, Puri R, Moura JMF, Eckersley P. Explainable machine learning in deployment. In: Proceedings of the 2020 conference on fairness, accountability, and transparency; 2020. p. 648–57. https://doi.org/10.1145/3351095.3375624.

[5] Biecek P. DALEX: Explainers for Complex Predictive Models in R. J Mach Learn Res 2018;19(84):1–5. https://jmlr.org/papers/v19/18-416.html.

[6] Biecek P. Explanations of Model Predictions with live and breakdown Packages. R J 2018;10(2):395–409. https://journal.r-project.org/archive/2018/RJ-2018-072/index.html.

[7] Biecek P, Burzykowski T. Explanatory model analysis – explore, explain, and examine predictive models [data science series]. New York: Chapman & Hall/CRC Press; 2021. https://pbiecek.github.io/ema.

[8] Breiman L. Bagging predictors. Mach Learn 1996;24(2): 123–40. https://doi.org/10.1007/BF00058655.

[9] Breiman L. Random forests, vol. 45. Machine Learning; 2001. p. 5–32. https://doi.org/10.1023/A:1010933404324.

[10] Busan Harbour Pilot's Association. Operational rules for embarkation and disembarkation areas of busan port pilots. 2023. http://www.busanpilot.co.kr/rule/control_new. [Accessed 28 February 2023].

[11] Camliyurt G, Kim SCS, Turgut A, Park GY. Risk Assessment for Marine Pilot Occupational Accidents using Fault Tree and Event. Tree Analysis 2022;46(5):400–8.

[12] Carter T, Williams JG, Roberts SE. Crew and passenger deaths from vessel accidents in United Kingdom passenger ships since 1900. Int Marit Health 2019;70(1):1–10. https://doi.org/10.5603/IMH.2019.0001.

[13] Čokorilo O, Mirosavljević P, Vasov L, Stojiljković B. Managing safety risks in helicopter maritime operations. J Risk Res 2013; 16(5):613–24. https://doi.org/10.1080/13669877.2012.737828.

[14] Chen Chien-Ta, Hung Chia-Tse, Lin Jyh-Dong, Sung Po-Hsun. Application of a decision tree method with a spatiotemporal object database for pavement maintenance and management.

J Mar Sci Technol 2015;23(3). https://doi.org/10.6119/JMST-014-0327-5. Article 5, https://jmstt.ntou.edu.tw/journal/vol23/iss3/5.

[15] Chou Chien-Chang, Su Yuh-Ling, Li Ren-Fu, Tsai Chaur-Luh, Ding Ji-Feng. Key Navigation Safety Factors In Taiwanese Harbors And Surrounding Waters. J Mar Sci Technol 2015;23(5). Article 12.

[17] FFPM (n.d.). History of maritime pilots. FFPM − Fédération Française des Pilotes Maritimes. https://pilotes-maritimes.com/en/history-of-maritime-pilots/.

[18] Fisher A, Rudin C, Dominici F. All Models Are Wrong, but Many Are Useful: Learning a Variable's Importance by Studying an Entire Class of Prediction Models Simultaneously. J Mach Learn Res 2019;20(177):1−81. https://jmlr.org/papers/v20/18-760.

[19] Friedman JH. Greedy function approximation: A gradient boosting machine. Ann Stat 2001;29(5):1189−232. https://doi.org/10.1214/aos/1013203451.

[20] Gao K, Tu H, Sun L, Sze NN, Song Z, Shi H. Impacts of reduced visibility under hazy weather conditions on collision risk and car-following behavior: Implications for traffic control and management. Int J Sustain Transport 2020;14(8):635−42. https://doi.org/10.1080/15568318.2019.1597226.

[21] Gill N, Hall P, Montgomery K, Schmidt N. A Responsible Machine Learning Workflow focusing on Interpretable Models, Post-hoc Explanation, and Discrimination Testing. Information 2020;11(3):137. https://doi.org/10.3390/info11030137.

[22] Gosiewska A, Kozak A, Biecek P. Simpler is better: Lifting interpretability-performance tradeoff via automated feature engineering. Decis Support Syst 2021;150:113556. https://doi.org/10.1016/j.dss.2021.113556.

[23] Harding J, Bridge T, Demartini G. In: Hiemstra D, Moens MF, Mothe J, Perego R, Potthast M, Sebastiani F, editors. CoralExp: an explainable system to support coral taxonomy research. Advances in information retrieval. ECIR 2021. Lecture notes in computer science; 2021. p. 12657. https://doi.org/10.1007/978-3-030-72240-1_55.

[24] ICS. Guide to Helicopter/Ship Operations (4 ed.). In: International chamber of shipping; 2008. Retrieved 03/03/2021, https://publications.ics-shipping.org/single-product.php?id=5.

[25] ILO. Global trends on occupational accidents and diseases. World Day for Safety and Health At Work; 2015. *April* 01, http://www.ilo.org/legacy/english/osh/en/story_content/external_files/fs_st_1-ILO_5_en.pdf.

[26] IMPA. International maritime pilots' association. Safety Campaign; 2020. 2020, chromeextension://efaidnbmnnnibpcajpcglclefindmkaj/viewer.html?pdfurl=https%3A%2F%2Fwww.impahq.org%2Fsites%2Fdefault%2Ffiles%2Fcontent-files%2Fimpa-safety-campaign-results-2020high.pdf&clen=8076692&chunk=true.

[27] Inoue K, Hara K. Relations between the number of observational days and the accuracy on the estimation of average annual daily traffic, vol. 50. Japan Institute of Navigation; 1974. p. 1−8.

[28] Jellen C, Oakley M, Nelson C, Burkhardt J, Brownell C. Machine-learning informed macro-meteorological models for the near-maritime environment. Appl Opt 2021;60(11):2938. https://doi.org/10.1364/ao.416680.

[29] Jeong Sang-ki, Choi Hyeung-Sik, Ji Dea-Hyung; Vu, Mai The, Kim Joon-Young, Hong Sung Min, Cho Hyun Joon. A Study on an Accurate Underwater Location of Hybrid Underwater Gliders Using Machine Learning. J Mar Sci Technol 2020;28(6). https://doi.org/10.6119/JMST.202012_28(6).0007. Article 7, https://jmstt.ntou.edu.tw/journal/vol28/iss6/7.

[30] Kim D, Antariksa G, Handayani M, Lee S, Lee J. Explainable Anomaly Detection Framework for Maritime Main Engine Sensor Data. Sensors 2021;21(15):5200. https://doi.org/10.3390/s21155200.

[31] Kuhn M. Building Predictive Models in R Using the Caret Package. J Stat Software 2008;28(5):1−26. https://doi.org/10.18637/jss.v028.i05.

[32] Lin Le-Hui, Chen Kee Kuo, Chiu Rong-Her. Predicting Customer Retention Likelihood in The Container Shipping Industry Through The Decision Tree Approach. J Mar Sci Technol 2017;25(1). https://doi.org/10.6119/JMST-016-0714-1. Article 3, https://jmstt.ntou.edu.tw/journal/vol25/iss1/3.

[33] Liu C, Chan Y, Kazmi SHA, Fu H. Financial Fraud Detection Model: Based on Random Forest. Int J Econ Finance 2015;7(7):178−88. https://doi.org/10.5539/ijef.v7n7p178.

[34] Lundberg SM, Lee S-I. A Unified Approach to Interpreting Model Predictions. Adv Neural Inf Process Syst 2017;30(NeurIPS). https://doi.org/10.48550/arXiv.1705.07874.

[35] Meere K, Van Damme J, Van Sprundel M. Occupational injuries in Flemish pilots in Belgium. A questionnaire survey. Int Marit Health 2005;56(1−4):67−77. https://pubmed.ncbi.nlm.nih.gov/16532586/.

[36] Miao Z, Gaynor KM, Wang J, Liu Z, Muellerklein O, Norouzzadeh MS, McInturff A, Bowie RCK, Nathan R, Yu SX, Getz WM. Insights and approaches using deep learning to classify wildlife. Sci Rep 2019;9(1):8137. https://doi.org/10.1038/s41598-019-44565-w.

[37] Mohanty SD, Lekan D, McCoy TP, Jenkins M, Manda P. Machine learning for predicting readmission risk among the frail: Explainable AI for healthcare. Patterns 2022;3(1):100395. https://doi.org/10.1016/j.patter.2021.100395.

[38] Molnar C. Interpretable machine learning − a guide for making black box models explainable. 2021. https://christophm.github.io/interpretable-ml-book.

[39] Narwaria M. Does explainable machine learning uncover the black box in vision applications? Image Vis Comput 2022;118:104353. https://doi.org/10.1016/j.imavis.2021.104353.

[40] Oshiro TM, Perez PS, Baranauskas JA. How Many Trees in a Random Forest? In: Perner P, editor. Machine learning and data mining in pattern recognition. MLDM 2012. Lecture notes in computer science; 2012. p. 7376. https://doi.org/10.1007/978-3-642-31537-4_13.

[41] Port-MIS (Management Information System). Ship Entry Data of Busan Port in Korea. https://new.portmis.go.kr/. [Accessed 20 December 2022].

[42] Port-MIS (Management Information System). Ship Entry Data of Busan Port in Korea. https://new.portmis.go.kr/. [Accessed 10 February 2023].

[43] R Core Team. R: a language and environment for statistical computing. 2021. https://www.R-project.org. [Accessed 6 June 2021].

[44] Roscher R, Bohn B, Duarte M, Garcke J. Explainable Machine Learning for Scientific Insights and Discoveries. IEEE Access 2020;8:42200−16. https://doi.org/10.1109/ACCESS.2020.2976199.

[45] Rudin C. Stop explaining black-box machine learning models for high-stakes decisions and use interpretable models instead. Nat Mach Intell 2019;1:206−15. https://doi.org/10.1038/s42256-019-0048-x.

[46] Singh K, Xie M. Bootstrap method. third ed. International Encyclopedia of Education; 2010. p. 46−51. https://doi.org/10.1016/B978-0-08-044894-7.01309-9.

[47] Sugihara S, Hayashi Y, Murai K, Ishikura A. Investigation of pilots' transfer between a vessel and a pilot boat at the mouths of Ise-Mikawa Bay and Osaka Bay. J Jpn Inst Navig 2013;128:235−42. https://doi.org/10.9749/jin.128.235.

[48] Tang L, Tang Y, Zhang K, Du L, Wang M. Prediction of grades of ship collision accidents based on random forests and Bayesian networks. In: 5th International Conference on Transportation Information and Safety (ICTIS); 2019. p. 1377−81. https://doi.org/10.1109/ICTIS.2019.8883590.

[49] Therneau T, Atkinson B. Rpart: recursive partitioning and regression trees. 2019. https://CRAN.R-project.org/package=rpart. [Accessed 6 June 2021].

[50] Tierney N. visdat: Visualising Whole Data Frames. J Open Source Softw 2017;2:355. https://doi.org/10.21105/joss.00355.

[51] Rulusçu ÖS, Özbaş B, Altiok T, Or I. Risk analysis of the vessel traffic in the strait of Istanbul. Risk Anal 2009;29(10):1454−72. https://doi.org/10.1111/j.1539-6924.2009.01287.x.

[52] UNCTAD. UNCTAD. 2021 (2021), https://unctad.org/statistics.

[53] Weng J, Xue S. Ship collision frequency estimation in port fairways: A case study. J Navig 2015;68(3):602−18. https://doi.org/10.1017/S0373463314000885.

[54] Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag; 2016. https://doi.org/10.1007/978-0-387-98141-3.

[55] Yassin SS, Pooja. Road accident prediction and model interpretation using a hybrid K-means and random forest algorithm approach. SN Appl Sci 2020;2(9):1−13. https://doi.org/10.1007/s42452-020-3125-1.

[56] Yip TL. Port traffic risks - A study of accidents in Hong Kong waters. Transport Res Part E: Logist Transport Rev 2008;44(5):921−31. https://doi.org/10.1016/j.tre.2006.09.002.

[57] Yoo, et al. A Study on the Observation Days of Maritime Traffic Investigation. J Korean Soc Marine Environ Safety 2015;21(4):397−402.

[58] Zou Z Bin, Peng H, Luo LK. The Application of Random Forest in Finance. Appl Mech Mater 2015;740:947−51. https://doi.org/10.4028/www.scientific.net/amm.740.947.