

Determining the Most Influential Human Factors in Maritime Accidents: a Data-Driven Approach

Andrea Coraddu^{a,*}, Luca Oneto^b, Beatriz Navas de Maya^a, Rafet Kurt^a

^a*Department of Naval Architecture, Ocean & Marine Engineering
Strathclyde University, Glasgow G1 1XW, UK*

^b*DIBRIS Department, University of Genova, Via Opera Pia 11a, 16145 Genova, Italy*

Abstract

Marine accidents are complex processes in which many factors are involved and contribute to accident development. For this reason, effectively analyse what combination of factors lead an accident event is a complex problem, especially when human factors are involved. State-of-the-art methods such as Human Factor Analysis and Classification System, Human reliability Assessments, and simple Statistical Analysis are not effective in many situations since they require the intervention of human experts with their limitations, biases, and high costs. The authors propose to use a data-driven approach able to utilise the information present in historical databases of marine accident for the purposes of establishing the most influential human factors. For this purpose a two-stage approach is presented: first, a data-driven predictive model is built able to predict the type of accident based on the contributing factors, and then the different contributing factors are ranked based on their ability to influence the prediction. Results on a real historical database of accidents provided by the Marine Accident Investigation Branch will support the proposed novel approach.

Keywords: Human Factors, Shipping Accidents, Accident Investigation, Data Analytics, Random Forests, Kernel methods, Boolean Kernels, Feature Ranking, Gini Impurity, Backward Elimination

*Corresponding Author

Email addresses: andrea.coraddu@strath.ac.uk (Andrea Coraddu),
luca.oneto@unige.it (Luca Oneto), beatriz.navas-de-maya@strath.ac.uk (Beatriz Navas de Maya), rafet.kurt@strath.ac.uk (Rafet Kurt)

1. Introduction

Historic accident analysis reveals that shipping accidents have always been a concern for the maritime sector, incurring in significant economic consequences and social impacts (Eliopoulou et al., 2016; Wang, 2002). In order to reduce accidents, the shipping industry is continuously developing and implementing safety measures, which overall have a significant contribution to maritime safety (Hesse, 2003).

Marine accidents are complex processes, in which usually there is not just a single factor involved (Hetherington et al., 2006; Jin et al., 2002; Kristiansen, 2013). Therefore, it becomes difficult to effectively analyse what combination of factors lead to a specific accident event. For this reason, the increase in maritime safety studies and consequent breakthroughs in the literature is relatively slow compared to the rate of accidents (Soner et al., 2015). Considering that around 90% of world trading is still carried out by shipping companies (Chauvin et al., 2013), and that the world fleet is constantly enlarged, this can be translated into an increase in the maritime accidents rate. Hence, even a small improvement in maritime accidents prevention can make a tremendous positive impact on safety. The complexity of identifying all the variables involved in accidents and non scientific methods followed during accident investigations, make extremely challenging to integrate lessons learnt from past accidents into safety assessments. This situation creates a barrier to enhancing safety.

When analysing the available literature, it becomes evident that human factors had a high contribution to past accidents. According to Azadeh & Zarrin (2016), human factors are the cause of at least 66% of the accidents and more than 90% of the incidents not only in the maritime sector but also in different industries (i.e. aerospace or nuclear). For instance, the explosion of the *Space Shuttle Challenger* or the *Three Mile Island* accident were attributed to human errors (Azadeh & Zarrin, 2016). Furthermore, O'Hare et al. (1994) identified that human factors are responsible for up to 80% of the accidents reported in both civil and military aviation. Regarding the maritime industry, different authors have also highlighted that some form of human error triggers more than 80% of maritime casualties (Graziano et al., 2016; Navas de Maya et al., 2018; Rothblum, 2000; Turan et al., 2016). However, despite the high contribution of human factors into past accidents, it was not until the sink of the vessel *Herald of Free Enterprise* in 1987 that the International Maritime Organization (IMO) started considering human fac-

tors from a different perspective (Schröder-Hinrichs et al., 2013). In the last decades, various authors have conducted numerous studies to identify and successfully evaluate the contribution of human factors into past accidents. By analysing the relevant literature, it is possible to identify three different approaches regarding the analysis of human factors.

The first and most applied technique to analyse and classify human factors is the well-known HFACS method (Shappell & Wiegmann, 2000), which initially was designed for military aviation. Thus, HFACS was successfully expanded to other strategic industries as civil aviation, railway, mining and maritime (Celik & Cebi, 2009; Chauvin et al., 2013; Chen et al., 2013; Luo & Shin, 2016; Schröder-Hinrichs et al., 2011; Yıldırım et al., 2017). However, although HFACS has a strong theoretical foundation (i.e. it is based on Reason's ideas and theory), it presents some limitation. HFACS have industry restrictions, as it was designed to analyse the causes of aviation accidents, then some categories within this model are not applicable to other sectors. Hence, HFACS presents a remarkable lack of versatility. Moreover, although psychological factors can be identified by interviewing relevant personnel when applying HFACS, these psychological factors will be limited due to the subjectivity of using expert judgement (Fu et al., 2017). Finally, by applying HFACS, the scope of the investigation is limited to the predefined taxonomy and organisation level, hence; it will be altered in each accident outcome or scenario being modelled.

The second technique, HRAs, has been extensively applied to diverse sectors. It is possible to establish two different generations regarding HRAs. Within the first generation, the concept of human error is associated with the weaknesses of people. In this first category, some representatives HRA methods are the Technique for Human Error Rate Prediction (THERP), the Human Error Assessment and Reduction Technique (HEART) or the Justification of Human Error Data Information (JHEDI). On the other hand, the second generation of HRAs are characterised by approaching human behaviour to risk analysis. In this category, some representatives HRA methods are the Technique for Human Event Analysis (ATHEANA) or the Cognitive Reliability and Error Analysis Method (CREAM) (Yang et al., 2013). The above-mentioned methods have been applied in numerous case studies (Castiglia et al., 2015; Kirwan, 1997; Kumar et al., 2017; Ung, 2015, 2019; Yang et al., 2014; Zhang et al., 2019; Zhou et al., 2017).

However, the majority of the above HRAs apply some sort of expert judgement, as the human element contribution into an accident is difficult

to quantify numerically. Nevertheless, by incorporating expert opinion, the results are highly influenced by the own expert knowledge, hence, the results become subjective. Besides, HRAs are designed for a specific sector or scenario. This limitation does not allow creating a generic model, which could be easily applied to any maritime accident. Hence, the necessity to model a new scenario for each case study is time and cost consuming. Thus, there is a lack of a proper technique, which could be applied to any accident scenario, to identify the main human-contributing factors into maritime accident.

Finally, additional techniques have been applied in the literature to identify accident contributing-factors and prevent accidents. First, statistical analysis has been extensively performed to identify accident contributing-factors (Bye & Aalberg, 2018; Eliopoulou et al., 2016; Papanikolaou et al., 2014; Schlögl et al., 2019; Ventikos et al., 2018; Yıldırım et al., 2017).

However, the aforementioned efforts of statistically modelling the relationship between contributors and accidental outcomes have been difficult due to the type of data and inconsistency in data collection. In addition, alternative techniques that have been applied in the literature for risk assessment and accident prevention are Bayesian Networks (BN), key risk indicators or fuzzy logic (Abou, 2012; Hänninen, 2014; Shi et al., 2018).

Nevertheless, despite all the research carried out in the past decades on the analysis of human factors in accident analysis, there are still some open questions and limitations (Jeong et al., 2016). In particular, the main limitations of the above describe techniques lie in the fact that the human element contribution into an accident is difficult to quantify, therefore there is the necessity of applying expert judgement. Consequently, the results are biased by the experts' past experience and subjective to the single expert interpretation of the events, and the quality of the results may not be instrumental in understanding the actual most important maritime accidents contributing human factors. Moreover, human expert judgement is time consuming and expensive. Finally, the solutions available in the literature are frequently tailored to a specific case, therefore the transferability of approaches is difficult to achieve due to demanding adaptation procedures related to the time necessary to engage and involve the adequate expertise. Therefore the application on different accidental scenario and sectors may be either not doable or effective (Schröder-Hinrichs et al., 2011).

In this paper, the authors propose to utilise Data-Driven Models (DDMs) (Oneto, 2019; Shalev-Shwartz & Ben-David, 2014; Vapnik, 1998) to create models exploiting robust statistical inference procedures, and historical data

to perform unbiased predictions on previously unseen cases. Recently, DDMs have proved to be valid instruments for transport related problems (Cipollini et al., 2018; Coraddu et al., 2017; Iranitalab & Khattak, 2017; Osman et al., 2019; Petersen et al., 2012; Smith et al., 2013) and, in some preliminary works (Akyuz et al., 2017; Grech et al., 2002; Hänninen, 2014; Lützhöft, 2004; Praetorius et al., 2015; Tirunagari et al., 2012), it has been shown that DDMs applied to the study of the human factor in accident analysis may produce interesting results.

In this paper, the authors propose a novel approach that can recognise the most important human factors involved into a specific accident outcome, eliminating the necessity of expert judgement and therefore the human bias, by simply exploiting and historical database of marine accidents. Moreover, another advantage of this human unbiased proposed approach lies in the ability to adapt the methodology on a different accidental scenario and sectors (i.e. aviation and railway) and taxonomy. To this aim, the authors propose a two-stage approach:

- (I) first a data-driven predictive model is built, leveraging on an historical accident database provided by the Marine Accident Investigation Branch (MAIB), able to predict the type of accident bases on the contributing factors;
- (II) then the different contributing factors are ranked based on their ability to influence the prediction.

Ensemble methods (Breiman, 2001) and Kernel methods (Hsu & Lin, 2002; Polato et al., 2018; Shawe-Taylor & Cristianini, 2004) represent the state-of-the-art for dealing with Phase (I). During Phase (II), state-of-the-art feature ranking methods for Ensemble and Kernel methods (Guyon & Elisseeff, 2003; Guyon et al., 2008; Louppe, 2014; Louppe et al., 2013) will be exploited to rank the human factor and then to identify the most important maritime accidents contributing factors.

The paper is structured as follows. Section 2 provides a detailed description of the data-set, including the data source, the data structure and the data description. In Section 3 the authors' proposal is described. Section 4, includes the experimental results obtained. Finally, Section 5 concludes the paper and discuss future perspectives related to this work.

2. Problem Description and Available Data

As described in the previous section, marine accidents are complex processes in which, usually, there is not a single factor solely responsible for the accident outcome. Hence, efforts must be focused on developing a suitable approach that allows identifying those factors with a higher contribution to accidents. Thanks to this information, specific strategies (i.e. adequate training programs or amend of existing specification) can be developed and applied in order to mitigate them and enhance the overall maritime safety.

The method proposed in this paper for the identification of main human factors responsible for past maritime accidents is applied to a historical accident database, which was provided by the Marine Accident Investigation Branch (MAIB)¹. The database comprises marine accidents involving UK vessels worldwide and all vessels operating in UK territorial waters. Thus, it includes detailed information about 91 accidents, in which human factors were recorded, for the period 2011-2016. Moreover, the aforementioned period was selected as MAIB started recording human factors accidental data in 2011.

Each accident in the database consists of 94 Boolean features (see Table 1), corresponding to the presence or not of a particular accident contributing factor, and an additional feature which identifies the accident category (see Table 2). The full list of accident contributing factor are provided by the European Maritime Safety Agency (EMSA) and reported in (EMSA, 2016). In the available database nine accident categories, or accident outcomes, are present and the following description are taken and adapted from EMSA (EMSA, 2018):

- Grounding/Stranding (C1): event during which a moving navigating ship, either under command or not (i.e. power or drift conditions), strikes the sea bottom, shore or underwater wrecks;
- Capsizing/Listing (C2): event during which the ship no longer floats in the right-side-up mode due to external factors (i.e. negative initial stability, transversal shift of the centre of gravity, or the impact of external forces);
- Contact (C3): casualty caused by a ship striking or being struck by an external object. Thus, the sea bottom is excluded (i.e. a contact with

¹<https://www.gov.uk/government/organisations/marine-accident-investigation-branch>

the sea bottom is considered as a grounding event);

- Collision (C4): casualty caused by ships striking or being struck by another ship. This event might involve more than two vessels;
- Fire/Explosion (C5): an uncontrolled ignition on board of a ship;
- Flooding/Foundering (C6): event during which the ship is taking water on board; It can be progressive or massive;
- Hull Failure (C7): failure event affecting the general structural strength of the ship;
- Loss of Control (C8): event during which a total or temporary loss of the ability to operate or manoeuvre the ship, failure of electric power occurs;
- Damage to Ship or Equipment (C9): damage to equipment, system or the ship not covered by any of the previous casualty types.

In addition to what has been provided by MAIB in the database, the method proposed in this paper is applied to different relabelling of above described database, aiming to demonstrate that the results remain significantly constant, independently of the accident outcome analysed. Hence, the method is applied to four different aggregations of the outcome of the accident (see Table 3):

- Aggregation 1 (A1): all the nine accident categories outcomes have been considered as independent categories. However, by considering each accident category separately, the number of data points in each category is low, which results in an unbalanced category distribution;
- Aggregation 2 (A2): in order to mitigate the unbalancing effect among categories and to increase the number of data points for each category, the previous accident outcomes have been aggregated into two main categories: Navigational Accidents (NA), which includes Grounding/Stranding (C1), Contact (C3), and Collision (C4), and Non- Navigational Accidents (NNA), which includes all the remaining categories (C2 + C5 + C6 + C7 + C8 + C9) considered as a unique group;
- Aggregation 3 (A3): the accident outcomes have been aggregated into seven groups, the first group incorporate all NA, while the NNA are considered as an individual category;
- Aggregation 4 (A4): the nine accident outcomes have been aggregated into four similar accident categories. NAs have been still considered as an independent group as the first aggregation. The Hull Failure (C7), Loss of Control (C8) and Damage to Ship or Equipment (C9) con-

Abb.	Feature	[%]	Abb.	Feature	[%]
F01	Anthropometric factors	2.2	F48	Lack of motivation/morale	13.2
F02	Audit	2.2	F49	Lack of priority to IMR	2.2
F03	Checks	2.2	F50	Lack of resources	2.2
F04	Conflicting orders, cross-pressure	1.1	F51	Lack of responsibility for own job	2.2
F05	Contingency plans not updated	2.2	F52	Lack of skill	30.8
F06	Cowboy attitudes, horseplay	1.1	F53	Lack of warning systems	1.1
F07	Crisis handling	1.1	F54	Lacks initiative to deal with emergencies	8.8
F08	Cross-pressure from schedule	0.0	F55	Language problem	1.1
F09	Design	0.0	F56	Life saving equipment	0.0
F10	Design error	3.3	F57	Long working periods, much overtime	5.5
F11	Deviation from standards/specifications	1.1	F58	Low job satisfaction, monotony	1.1
F12	Display design, controls	2.2	F59	LTA assessment of needs and risks	2.2
F13	Emergency plans	1.1	F60	LTA communication	8.8
F14	Emergency procedures	2.2	F61	LTA design verification	1.1
F15	Emergency training program	7.7	F62	LTA Formal safety assessment, risk analysis	12.1
F16	Expectations of supervisor is unclear	4.4	F63	LTA medical services provided	0.0
F17	Failure not detected during IMR	1.1	F64	LTA mental and psychological state	17.6
F18	Follow-up of non-conformities	0.0	F65	LTA physical/physiological capability	3.3
F19	Frequent change of watch schedule	1.1	F66	LTA planning	1.1
F20	Hazardous/ messy workplace	0.0	F67	LTA Safety plan and program	2.2
F21	Health control of personnel	1.1	F68	LTA System review and evaluation	2.2
F22	Hiring and selection policy	1.1	F69	Management training	0.0
F23	Idleness, waiting	1.1	F70	No review or critical tasks	2.2
F24	Improper performance of maintenance/repair	2.2	F71	Person-to-person conflict	0.0
F25	Improper supervisory example	12.1	F72	Pressure to keep schedule and costs	1.1
F26	Inadequate briefing, instruction	12.1	F73	Regulation	2.2
F27	Inadequate control of life saving equipment	1.1	F74	Regulatory procedures	1.1
F28	Inadequate fighting equipment	1.1	F75	Regulatory standards	4.4
F29	Inadequate illumination	0.0	F76	Resistance to change	2.2
F30	Inadequate maintenance	1.1	F77	Restricted fairway	1.1
F31	Inadequate manning	5.5	F78	Right tools and equipment unavailable	1.1
F32	Inadequate procedures and check lists	14.3	F79	Safety awareness, cutting corners	22.0
F33	Inadequate promotion of Safety	4.4	F80	Sea motion	0.0
F34	Inadequate standards or specifications	5.5	F81	Selection/training of officers	1.1
F35	Inadequate testing	1.1	F82	Social and cultural barriers and conflicts	6.6
F36	Inadequate training program	12.1	F83	Supervision	0.0
F37	Inadequate work methods	15.4	F84	Supervisors not in touch	1.1
F38	Inadequate work preparation	18.7	F85	Surveillance	2.2
F39	Inappropriate peer pressure	1.1	F86	Too high work load/low work load	3.3
F40	Inappropriate regulations	1.1	F87	Too low visibility for observation	2.2
F41	Inspection	6.6	F88	Traffic density hinders vessel control	1.1
F42	Lack of communication and coordination	4.4	F89	Training ignored	8.8
F43	Lack of co-ordination of tasks	4.4	F90	Unclear roles and responsibility	0.0
F44	Lack of information	5.5	F91	Use of wrong equipment	1.1
F45	Lack of knowledge	34.1	F92	Work instruction	1.1
F46	Lack of leadership	4.4	F93	Work place inspections	0.0
F47	Lack of maintenance	1.1	F94	Wrong person assigned	0.0

Table 1: Dataset input variables. [%] means % of elements in that feature that are equal to 1 (actually contributing factor).

Abb.	Accident Type	Size
C1	Grounding/Stranding	23
C2	Capsizing/Listing	5
C3	Contact	12
C4	Collision	19
C5	Fire/Explosion	9
C6	Flooding/Foundering	12
C7	Hull Failure	1
C8	Loss of Control	7
C9	Damage to Ship or Equipment	3

Table 2: Accident categories and sizes.

formed a second cluster that represents Equipment, Operational and Structural Accidents (EOSA). This aggregation was created as EOSA might be considered as prior conditions that can derive into additional NA accidents. Fire/Explosion (C5) was considered as an independent group (F\EA) due to its own accident nature. As a final group, related to Safety and Stability Accidents (SSA), Capsizing/Listing (C2) and Flooding/Foundering (C6) were considered as single grouped.

Abb.	Group	Size
A1	C1, C2, C3, C4, C5, C6, C7, C8, C9	-
A2	NA (C1 + C3 + C4),	54
	NNA (C2 + C5 + C6 + C7 + C8 + C9)	37
A3	NA (C1 + C3 + C4),	54
	C2, C5, C6, C7, C8, C9	-
A4	NA (C1 + C3 + C4),	54
	EOSA (C7 + C8 + C9),	11
	F\EA (C5),	9
	SSA (C2 + C6)	17

Table 3: Proposed aggregation and sizes. - means that the size is reported already in Table 2

3. Data-Driven Methods

In order to address the specific problem described in the previous section, the authors will adopt state-of-the-art data-driven approaches. In particular, given the available data, the problem of determining the most influential human factors in maritime accidents can be mapped into a standard feature ranking problem (Shalev-Shwartz & Ben-David, 2014). The specific problem here is to rank the human factors based on their ability to be influential in predicting the type of accident in the set of accidents available. This will allow stating that, based on the available data, the top human factors in the ranking are the most influential in understanding a particular incident and the most influential in understanding the maritime accidents. To perform an accurate, effective, and reliable feature ranking a two-steps approach is implemented. The first step is to build a predictive model (Shalev-Shwartz & Ben-David, 2014) of the accident type based on the presence or not of the different human factors. This step can be performed using different approaches, in this paper the authors will exploit two different state-of-the-art approaches: Random Forests (RF) (Breiman, 2001; Harb et al., 2009) and Multiclass (Hsu & Lin, 2002) Support Vector Machines (Shawe-Taylor &

Cristianini, 2004) with Boolean Kernels (Polato et al., 2018) - MSVM-BK, since the presence or not of the different human factors can be represented with a Boolean vector. The second step is to rank (Guyon & Elisseeff, 2003; Guyon et al., 2008) the different human factors based on their ability to influence the model outputs. For what concerns this second steps RF naturally provide, as a by-product of the model creation, the ranking of the different input features (Louppe, 2014) while for MSVM-BK the authors will exploit the Backward Elimination algorithm to extract the same information (Larose, 2015).

3.1. Building a Predictive Model

Let us consider the now-classical multiclass classification problem (Shalev-Shwartz & Ben-David, 2014) with a Boolean feature space. In particular, let us consider an input space $\mathcal{X} = 0, 1^d$, consisting of d features (the presence or not of a particular human factor), and an output space $\mathcal{Y} = \{1, 2, \dots, c\}$ consisting of c possible classes (the accident type, c depends on the type of grouping). The authors' proposed approach is to estimate the unknown rule that maps an item $X \in \mathcal{X}$ to an item $Y \in \mathcal{Y}$. It is worth noting that the rule can be non-deterministic (Shalev-Shwartz & Ben-David, 2014) and some values of X may be missing (Donders et al., 2006). For this reason, an additional category is introduced for that particular feature. A set of labelled samples $\mathcal{D}_n = \{(X_1, Y_1), \dots, (X_n, Y_n)\}$ is available for the learning phase. A learning algorithm \mathcal{A}_h , characterised by its hyperparameters h , maps \mathcal{D} to a function $f = \mathcal{A}_h(\mathcal{D}_n)$. The error that f commits in approximating the unknown mapping rule is measured with reference to a loss function ℓ . In this particular classification problem framework, the authors decided to use the loss function that counts the number of misclassified samples have been chosen: $\ell(f(X), Y) = [f(X) \neq Y]$, in which the Iverson bracket notation is exploited. The expected error of f is generally referred as to generalisation error (Shalev-Shwartz & Ben-David, 2014) and defined as

$$L(f) = \mathbb{E} \{ \ell(f(X), Y) \}. \quad (1)$$

It is evident that $L(f)$ cannot be computed, as the distribution of the samples is unknown, nevertheless, it is still possible to compute its empirical estimator (empirical error) which is defined as

$$\widehat{L}(f, \mathcal{T}_t) = \frac{1}{t} \sum_{(X,Y) \in \mathcal{T}_t} \ell(f(X), Y), \quad (2)$$

where $\mathcal{T}_t = \{(X_1, Y_1), \dots, (X_t, Y_t)\}$ must be a different set from \mathcal{D}_n , which has been used to build f to ensure that the estimator of the model quality is unbiased Oneto (2019).

When it comes to binary classification problems, namely $\mathcal{Y} \in \{\pm 1\}$, RF (Breiman, 2001) can be considered as a state-of-the-art powerful learning algorithm. Nevertheless, to describe the RFs in detail, the authors would like to recall the definition and construction of a binary Decision Tree (DT) (Rokach & Maimon, 2008). A binary DT represents a flowchart-like structure in which each internal node represents a test of a feature, each branch is associated with the outcome of the test, and each leaf node represents a class label. In this framework, a path from the root to a leaf is representing a classification rule. A recursive schema is utilised to build the DT until it reaches its desired depth n_d . Each node of the DT, starting from the root node, can be built by choosing both the attribute and the cut that most effectively splits the set of samples into two subsets based on the information gain.

Formally, for Boolean features, the node test of a feature j (briefly called cut or split) can be defined as

$$X_j = 0. \quad (3)$$

Based on this cut, it is possible to split the original data \mathcal{D}_n into two subsets \mathcal{D}_n^0 and \mathcal{D}_n^1 such that

$$\mathcal{D}_n^0 = \{(X, Y) \in \mathcal{D}_n | X_j = 0\}, \quad \mathcal{D}_n^1 = \mathcal{D}_n \setminus \mathcal{D}_n^0. \quad (4)$$

Then, it is possible to define the information gain of Eq. (3), namely the reduction in entropy (or confusion) induced by the cut

$$G(j) = H(\mathcal{D}) - \frac{|\mathcal{D}_n^0|}{n} H(\mathcal{D}_n^0) - \frac{|\mathcal{D}_n^1|}{n} H(\mathcal{D}_n^1), \quad (5)$$

where H is the entropy of the set of samples

$$H(\mathcal{D}) = \sum_{i \in \{\text{Distinct values of } Y \in \mathcal{D}\}} \frac{|\mathcal{D}(Y = i)|}{|\mathcal{D}|} \log_c \left(\frac{|\mathcal{D}(Y = i)|}{|\mathcal{D}|} \right) \quad (6)$$

$$\mathcal{D}(Y = i) = \{(X, Y) \in \mathcal{D} | Y = i\}. \quad (7)$$

Consequently, the best cut is defined as

$$j^* = \arg \max_{j \in \{1, \dots, d\}} G(j). \quad (8)$$

At this point, it is possible to briefly describe the learning phase of each of the n_t trees that are composing the RF. In particular, from \mathcal{D}_n , n_b samples are taken with replacement, and \mathcal{D}' is built. A tree is constructed with \mathcal{D}' , but the best split is chosen among a subset of n_v predictors over the possible n_f predictors randomly chosen at each node. The tree is grown until its depth reaches the maximum value of n_d or until all the samples in \mathcal{D}' are correctly classified. During the classification phase of a previously unseen X , each tree classifies X into a class Y ; the final classification is the mode of all the answers of each tree in the RF. Note that n_b , n_v , n_d , and n_t are the hyperparameters of the RF. If $n_b = n$, $n_v = \sqrt{n_f}$, and $n_d = \infty$ we obtain the original RF formulation (Breiman, 2001), where n_t is usually chosen as a tradeoff between accuracy and efficiency (Orlandi et al., 2016).

In the case of binary classification problems, namely $\mathcal{Y} \in \{\pm 1\}$, one of the most powerful learning algorithm, are the SVM. Then, the SVM classifier is defined as

$$f(X) = \text{sign}(W^T \phi(X) + b), \quad (9)$$

where ϕ is an unknown nonlinear function which maps $\{0, 1\}^r \rightarrow \mathbb{R}^D$ the weights $W \in \mathbb{R}^D$ and the bias $B \in \mathbb{R}$ are found by solving the following (primal) convex constrained quadratic programming problem:

$$\begin{aligned} \min_{W, B, \xi} \quad & \frac{1}{2} \|W\|^2 + C 1^T \xi \\ & Y_i (W^T \phi(X_i) + B) \geq 1 - \xi_i, \quad \xi_i \geq 0, \end{aligned} \quad (10)$$

where $1_i = 1 \forall i \in \{1, \dots, n\}$ (Shawe-Taylor & Cristianini, 2004). This problem is known as the Tikhonov formulation of the SVM, as it can be seen as a regularised ill-posed problem. The introduction of the n Lagrange multipliers $\alpha_1, \dots, \alpha_n$, allow writing the Problem (10) in its dual form, for which state-of-the-art solvers are available (Shawe-Taylor & Cristianini, 2004)

$$\begin{aligned} \min_{\alpha} \quad & \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j Y_i Y_j K(X_i, X_j) - \sum_{i=1}^n \alpha_i \\ & 0 \leq \alpha_i \leq C, \quad \sum_{i=1}^n y_i \alpha_i = 0, \end{aligned} \quad (11)$$

where $K(X_i, X_j) = \phi(X_i)^T \phi(X_j)$ is a suitable kernel function (Schölkopf et al., 2001) which allow solving the problem without knowing explicitly

ϕ (Shawe-Taylor & Cristianini, 2004). After solving Problem (11), it is possible to use the Lagrange multipliers in order to define the dual form of the SVM classifier

$$f(\mathbf{x}) = \sum_{i=1}^n Y_i \alpha_i K(X_i, X) + b. \quad (12)$$

There are many hyperparameters in SVM. C in Problems (10) and (11) is the first one. The second one is the kernel in Problem (11). The third one is the kernel parameters if present. For what concerns the type of kernels, some of them are specifically developed for the case when $\mathcal{X} = \{0, 1\}^d$ and are called Boolean Kernels, reviewed and evolved here (Polato et al., 2018). Without describing them in details, since the mathematical background needed is over-complicated for the purposed of this paper, the idea is to not represent all the possible analogical functions in \mathbb{R}^d but just all the possible Boolean functions in $\{0, 1\}^d$. This allows to remove much useless function and improve the quality of the learning procedure. All these Boolean Kernels are characterised by one or more hyperparameters and in this work we will explore all the kernels and all the ranges of hyperparameters adopted in (Polato et al., 2018). Even if SVM is a very powerful tool, real world problems are frequently characterised by $c > 2$ classes. SVM cannot naively tackle multi-class classification problems, and to overcome this limitation of SVM several techniques can be applied (Hsu & Lin, 2002). As the problem under investigation is a multiclass classification problem, the authors decided to exploit the All-Versus-All (AVA) approach. The main reasons behind this choice are: (a) AVA is the simplest method, (b) if the original problem is balanced, AVA does not create unbalanced classification problems (c) the proposed learning problem is characterised by a small size. In particular, the AVA method consists in building $c(c-1)/2$ training sets each one containing data from two different classes, u and v . These sets are used for training $c(c-1)/2$ different binary classifiers and the resulting models are saved for the forward phase. To classify a new pattern, the resulting model is applied in input to all the binary classifiers, and the majority voting is taken as a predictor.

As described for both M-SVM-BK, all the proposed algorithms have a series of hyperparameters that need to be tuned and once the model is built a statistical reliable estimation of its performance need to be constructed. The optimal hyperparameters selection and the unbiased estimation of the performance of the final model are fundamental problems when it comes to learning

from data (Oneto, 2019). The authors applied the resampling methods because of their simplicity and proved effectiveness in real world problems application (Oneto, 2020). The original dataset \mathcal{D}_n is resampled once (or many times (r)), with or without replacement, to build three independent datasets: training \mathcal{L}_l^i , validation \mathcal{V}_v^i , and test sets \mathcal{T}_t^i , with $i \in \mathcal{I}_r = \{1, \dots, r\}$. Note that $\forall i \in \mathcal{I}_r$

$$\mathcal{L}_l^i \cap \mathcal{V}_v^i = \emptyset, \quad \mathcal{L}_l^i \cap \mathcal{T}_t^i = \emptyset, \quad \mathcal{V}_v^i \cap \mathcal{T}_t^i = \emptyset, \quad \mathcal{L}_l^i \cup \mathcal{V}_v^i \cup \mathcal{T}_t^i = \mathcal{D}_n. \quad (13)$$

The best set of hyperparameters has been selected to create models based on \mathcal{L}_l^i with $i \in \mathcal{I}_r$ which is characterised by good, on average, performances with respect to \mathcal{V}_v^i with $i \in \mathcal{I}_r$. Since the data belonging to \mathcal{L}_l^i are independent from \mathcal{V}_v^i , the best set of hyperparameters should belong to the one which can achieve a small error on a data set that is independent from the training set. When the best set of hyperparameters is found, it is possible to select the best model training the algorithm using the whole dataset and test it in on \mathcal{T}_t^i . The model's error, achieved with the training on $\mathcal{L}_l^i \cup \mathcal{V}_v^i$ is independent from \mathcal{T}_t^i , and for this reason, the error committed on the test set is an unbiased estimator of the generalisation error. In this work the non-parametric Bootstrap has been utilised, as it is the most powerful resampling method. For this reason, $l = n$ and \mathcal{L}_l^i must be sampled with replacement from \mathcal{D}_n , while \mathcal{V}_v^i and \mathcal{T}_t^i must be sampled without replacement from the sample of \mathcal{D}_n that has not been sampled in \mathcal{L}_l^i (Oneto, 2019). It is worth noting that for the non-parametric Bootstrap procedure $r \leq \binom{2n-1}{n}$.

3.2. Feature Ranking

As described above from the models build with RF and M-SVM-BK, it is possible to rank the Boolean features (human factor present or not for a particular accident) in order to understand the most important ones for understanding the maritime accidents.

In this context, feature rankings methods based on RF is surely a state-of-the-art approach (Louppe, 2014; Louppe et al., 2013). Several measures are available for feature importance in Random Forests. One approach is the one based on the Gini Importance or Mean Decrease in Impurity (MDI) which evaluates each feature importance as the sum over the number of splits (across all the trees) including the feature, proportionally to the number of samples it splits. Another powerful approach is the one based on the Permutation Importance or Mean Decrease in Accuracy (MDA), where the importance is

assessed for each feature by removing the association between that feature and the target. This can be achieved with a random permutation (Good, 2013) of the values of the feature, and measuring the resulting increase in error. The influence of the correlated features is also removed. In particular, for every tree, two quantities are computed: the first quantity is the error on the out-of-bag samples utilised during the prediction, while the second quantity is the error on the out-of-bag samples after a random permutation of the values of a variable. Consequently, these two values are subtracted, and then the average of the result over all the trees in the ensemble represents the raw importance score for the variable under consideration. Both MDI and MDA can be adopted since they can be easily carried out during the main prediction process inexpensively.

Nevertheless, also M-SVM-BK is a very powerful technique for building a predictive model but unfortunately it does not provide, naturally, also a ranking of the features exploited for building the model. For this purpose a wrapper method (Guyon & Elisseeff, 2003; Guyon et al., 2008) has been exploited which allow deriving this information a posteriori. The method that will be used is the backward elimination (Guyon & Elisseeff, 2003) which eliminates one feature. In particular, for each of the features two quantities are computed: the error on a test set when all the features are exploited and the same quantity when the particular feature is removed. Then the raw importance is computed by subtracting the second from the first. Then the features are ranked based on this score.

Note that for each method, multiple ranking results are available. One for each of the type of labels described in Section 2 and, in case of RF, one for each of the feature ranking methods (for RF the authors considered MDI and MDA, while for M-SVM-BK just the backward elimination). Finally, the different rankings have been aggregated into a unified merged ranking by using Borda’s method (Sculley, 2007), where we summed the two position of each feature in the two rankings and sorted the ranking accordingly.

4. Experimental results

In this section, the authors report the results of applying the methodology proposed in Section 3 on the data described in Section 2 for solving the problem of understanding what the most important factors which characterise the maritime accidents are. For this purpose, during the model selection phase the authors search $n_b \in \{0.70, 0.75, \dots, 1.20\}n$, $n_v \in d^{0.00, 0.05, \dots, 1}d$

and $n_d \in \{1, 2, 4, 8, 16, 32, 64\}$ for RF where we set $n_t = 5000$ since increasing more n_t does not change the results. For what concern M-SVM-BK the authors search $C \in 10^{\{-6.0, -5.5, \dots, 6.0\}}$ and select the kernels and its hyperparameters between all the possible combination proposed in (Polato et al., 2018). Tables 4 and 5 report the confusion matrices (in %) for RF and M-SVM-BK on the different classification problems described in Section 2. Because of space constraints, the average % has been reported, in any case variance is always below the 5% of the results. Both RF and M-SVM-BK models are characterised by a satisfactory accuracy for the task under consideration.

From Tables 6 to 13, we reported the top 10 features ranked by RF and M-SVM-BK on the different classification problems described in Section 2, considering the different aggregations. From the results, it is possible to observe that the methodology is robust and not affected by the particular aggregation choice, providing a consistent ranking of the features. Table 14 reports the features ranking obtained by counting the frequency of occurrence on all the classification problems. While Tables 15 and 16 report the top 10 features ranked by RF and M-SVM-BK on all the classification problems described in Section 2. It is possible to note that the features reported in Table 14 are quite different from the ones obtained using the proposed approach and reported in Table 15 and 16. Moreover, the results obtained based on a simplistic evaluation of the frequency of occurrences does not support or show any predictive power ability of these features.

From the Top 10 features that were ranked from the RF and M-SVM-BK methods, it is possible to note that the majority of these factors are related with inadequate procedures or deviation from the Standard Operating Procedure (SOP) (i.e. F05 - Contingency plans not updated, F37 - Inadequate work methods and F38 - Inadequate work preparation), or with inadequate training (i.e. F15 - Emergency training program, F45 - Lack of knowledge, F52 - Lack of skill and F89 - Training ignored). Inadequate procedures have been often referred to in the literature as a potential cause for maritime accidents. For example, research conducted in EU funded SEAHORSE project concluded that up to a third of SOP are ineffective hence not being adequately followed during ship operations (Kurt et al., 2016, 2015). Thus, in their study Chang & Lin (2006) revealed that inadequate procedures and/or inadequate resources are the major contributors into fire and explosion accidents. Regarding inadequate training, it has also been reported in the literature as a highly contributing factor into maritime accidents. For instance, Puisa et al. (2018) conducted an analysis of various maritime accidents, which

revealed that inadequate training was often an observable feature, appearing frequently across all accident reports analysed. In addition, Graziano et al. (2016) applied the Technique for Retrospective and Predictive Analysis of Cognitive Errors (TRACEr) to reveal that fatigue or inadequate training and instruction led to most of the failures identified in their study. Moreover, also a study performed by Kum & Sahin (2015) concluded that maritime accidents in extreme environments are often associated with inadequate quality and extension of training. Furthermore, additional accident contributing factors, which are neither related to inadequate procedures nor insufficient training, have been also ranked in the Top 10 contributors (i.e. F01 - Anthropometric factors, F24 - Improper performance of maintenance/repair, F26 - Inadequate briefing, instruction, F31 - Inadequate manning and F79 - Lack of knowledge). Above-mentioned factors are not often referred to in the literature as a common accidental cause (e.g. anthropometric factors). Nevertheless, one of the main limitations in EMCIP taxonomy is the lack of a proper description for each contributing factor, which often lead to misinterpretation when accident investigators are identifying which accidental factors are involved in an specific accident scenario. Thus, this misinterpretation may often lead to the inadequate selection of certain contributing factors when analysing an accident, which causes that certain factors are accounted more times that they appear, justifying that those factors have also been ranked in the Top 10 features even if they are not known as common accidental causes.

5. Conclusions

In this work, a method to define numerical models able to establish the most influential human factors contributing to the maritime accident has been presented. The work aims to identify the main relevant human factors involved in accident development and ranking their contribution. With this purpose in mind, a two-stage approach is introduced to build a robust and reliable data-driven model capable of predicting the type of accident based on the contributing factors. The outcome of the model is utilised for the factors' ranking based on their influence on the predictions. Two different state-of-the-art methods, Ensemble and Kernel methods, have been presented, able to exploit the sources of information provided by a real database containing marine accidents involving UK vessels worldwide and all ships operating in UK territorial waters.

Table 4: Confusion matrices (in %) of RF on the different classification problems described in Section 2. Where the considered categories are: C1 - Grounding/stranding, C2 - Capsizing/Listing, C3 - Contact, C4 - Collision, C5 - Fire/Explosion, C6 - Flooding/Foundering, C7 - Hull failure, C8 - Loss of control, C9 - Damage to ship or equipment, NA - Navigational Accident, NNA - Non Navigational, EOSA - Operational and Structural Accidents, F\EA - Fire and Explosions, SSA - Safety and Stability Accidents.

		Truth								
		A1	C1	C2	C3	C4	C5	C6	C7	C8
Pred.	C1	12.13	0.76	2.00	3.47	1.44	2.00	0.15	1.10	0.45
	C2	0.97	2.64	0.43	0.75	0.31	0.43	0.03	0.24	0.10
	C3	2.32	0.40	6.33	1.81	0.75	1.04	0.08	0.57	0.23
	C4	3.67	0.63	1.65	10.02	1.19	1.65	0.12	0.90	0.37
	C5	1.74	0.30	0.78	1.36	4.75	0.78	0.06	0.43	0.18
	C6	2.32	0.40	1.04	1.81	0.75	6.33	0.08	0.57	0.23
	C7	0.19	0.03	0.09	0.15	0.06	0.09	0.53	0.05	0.02
	C8	1.35	0.23	0.61	1.06	0.44	0.61	0.04	3.69	0.14
	C9	0.58	0.10	0.26	0.45	0.19	0.26	0.02	0.14	1.58

		Truth	
		A2	NA
Pred.	NA	52.22	4.88
	NNA	7.12	35.78

		Truth						
		A3	C2	C5	C6	C7	C8	C9
Pred.	C2	3.24	0.25	0.34	0.03	0.19	0.08	3.29
	C5	0.24	5.84	0.62	0.05	0.34	0.14	5.92
	C6	0.31	0.59	7.78	0.06	0.45	0.18	7.89
	C7	0.03	0.05	0.07	0.65	0.04	0.02	0.66
	C8	0.18	0.35	0.48	0.04	4.54	0.11	4.60
	C9	0.08	0.15	0.21	0.02	0.11	1.95	1.97
	NA	1.41	2.67	3.70	0.27	2.03	0.83	35.01

		Truth			
		A4	EOSA	F\EA	NA
Pred.	EOSA	8.10	0.44	5.82	0.92
	F\EA	0.45	6.63	4.76	0.75
	NA	2.69	2.15	39.76	4.50
	SSA	0.85	0.68	9.00	12.52

Table 5: Confusion matrices (in %) of M-SVM-BK on the different classification problems described in Section 2. Where the considered categories are: C1 - Grounding/stranding, C2 - Capsizing/Listing, C3 - Contact, C4 - Collision, C5 - Fire/Explosion, C6 - Flooding/Foundering, C7 - Hull failure, C8 - Loss of control, C9 - Damage to ship or equipment, NA - Navigational Accident, NNA - Non Navigational, EOSA - Operational and Structural Accidents, F\EA - Fire and Explosions, SSA - Safety and Stability Accidents.

		Truth								
A1		C1	C2	C3	C4	C5	C6	C7	C8	C9
Pred.	C1	17.44	0.46	1.19	2.07	0.86	1.19	0.09	0.65	0.27
	C2	0.58	3.79	0.26	0.45	0.19	0.26	0.02	0.14	0.06
	C3	1.38	0.24	9.10	1.08	0.45	0.62	0.05	0.34	0.14
	C4	2.19	0.38	0.98	14.41	0.71	0.98	0.07	0.54	0.22
	C5	1.04	0.18	0.47	0.81	6.82	0.47	0.03	0.26	0.10
	C6	1.38	0.24	0.62	1.08	0.45	9.10	0.05	0.34	0.14
	C7	0.12	0.02	0.05	0.09	0.04	0.05	0.76	0.03	0.01
	C8	0.81	0.14	0.36	0.63	0.26	0.36	0.03	5.31	0.08
	C9	0.35	0.06	0.16	0.27	0.11	0.16	0.01	0.09	2.27

		Truth	
A2		NA	NNA
Pred.	NA	57.44	1.30
	NNA	1.90	39.36

		Truth						
A3		C2	C5	C6	C7	C8	C9	NA
Pred.	C2	4.23	0.14	0.19	0.01	0.11	0.04	1.84
	C5	0.13	7.62	0.35	0.03	0.19	0.08	3.32
	C6	0.18	0.33	10.15	0.03	0.25	0.10	4.43
	C7	0.01	0.03	0.04	0.85	0.02	0.01	0.37
	C8	0.10	0.19	0.27	0.02	5.92	0.06	2.58
	C9	0.04	0.08	0.12	0.01	0.06	2.54	1.11
	NA	0.79	1.50	2.07	0.15	1.14	0.47	45.69

		Truth			
A4		EOSA	F\EA	NA	SSA
Pred.	EOSA	10.76	0.15	1.94	0.31
	F\EA	0.15	8.80	1.59	0.25
	NA	0.90	0.72	52.81	1.50
	SSA	0.28	0.23	3.00	16.63

Table 6: Top 10 features ranked by RF on the Aggregation No. 1 classification problem described in Section 2.

Aggregation 1		
Pos.	Feat.	Description
1	F45	Lack of knowledge
2	F89	Training ignored
3	F15	Emergency training program
4	F05	Contingency plans not updated
5	F82	Social and cultural barriers and conflicts
6	F01	Anthropometric factors
7	F24	Improper performance of maintenance/repair
8	F26	Inadequate briefing, instruction
9	F25	Improper supervisory example
10	F54	Lacks initiative to deal with emergencies

Table 7: Top 10 features ranked by RF on the Aggregation No. 2 classification problem described in Section 2.

Aggregation 2		
Pos.	Feat.	Description
1	F38	Inadequate work preparation
2	F89	Training ignored
3	F37	Inadequate work method
4	F31	Inadequate manning
5	F15	Emergency training program
6	F52	Lack of skill
7	F01	Anthropometric factors
8	F16	Expectations of supervisor is unclear
9	F64	LTA medical services provided
10	F45	Lack of knowledge

Table 8: Top 10 features ranked by RF on the Aggregation No. 3 classification problem described in Section 2.

Aggregation 3		
Pos.	Feat.	Description
1	F45	Lack of knowledge
2	F89	Training ignored
3	F15	Emergency training program
4	F38	Inadequate work preparation
5	F05	Contingency plans not updated
6	F24	Improper performance of maintenance/repair
7	F79	Safety awareness, cutting corners
8	F37	Inadequate work method
9	F52	Lack of skill
10	F64	LTA medical services provided

Table 9: Top 10 features ranked by RF on the Aggregation No. 4 classification problem described in Section 2.

Aggregation 4		
Pos.	Feat.	Description
1	F89	Training ignored
2	F15	Emergency training program
3	F38	Inadequate work preparation
4	F05	Contingency plans not updated
5	F24	Improper performance of maintenance/repair
6	F37	Inadequate work method
7	F45	Lack of knowledge
8	F79	Safety awareness, cutting corners
9	F60	LTA communication
10	F52	Lack of skill

Table 10: Top 10 features ranked by M-SVM-BK on the Aggregation No. 1 classification problem described in Section 2.

Aggregation 1		
Pos.	Feat.	Description
1	F15	Emergency training program
2	F01	Anthropometric factors
3	F89	Training ignored
4	F05	Contingency plans not updated
5	F38	Inadequate work preparation
6	F26	Inadequate briefing, instruction
7	F31	Inadequate manning
8	F37	Inadequate work method
9	F24	Improper performance of maintenance/repair
10	F45	Lack of knowledge

Table 11: Top 10 features ranked by M-SVM-BK on the Aggregation No. 2 classification problem described in Section 2.

Aggregation 2		
Pos.	Feat.	Description
1	F89	Training ignored
2	F15	Emergency training program
3	F24	Improper performance of maintenance/repair
4	F05	Contingency plans not updated
5	F01	Anthropometric factors
6	F26	Inadequate briefing, instruction
7	F37	Inadequate work method
8	F31	Inadequate manning
9	F38	Inadequate work preparation
10	F45	Lack of knowledge

Table 12: Top 10 features ranked by M-SVM-BK on the Aggregation No. 3 classification problem described in Section 2.

Aggregation 3		
Pos.	Feat.	Description
1	F15	Emergency training program
2	F05	Contingency plans not updated
3	F01	Anthropometric factors
4	F89	Training ignored
5	F24	Improper performance of maintenance/repair
6	F31	Inadequate manning
7	F26	Inadequate briefing, instruction
8	F37	Inadequate work method
9	F45	Lack of knowledge
10	F38	Inadequate work preparation

Table 13: Top 10 features ranked by M-SVM-BK on the Aggregation No. 4 classification problem described in Section 2.

Aggregation 4		
Pos.	Feat.	Description
1	F15	Emergency training program
2	F24	Improper performance of maintenance/repair
3	F01	Anthropometric factors
4	F05	Contingency plans not updated
5	F89	Training ignored
6	F38	Inadequate work preparation
7	F31	Inadequate manning
8	F37	Inadequate work method
9	F26	Inadequate briefing, instruction
10	F52	Lack of skill

Table 14: Top 10 features ranked counting the frequency of occurrence on all the classification problems described in Section 2.

Frequency Count		
Pos.	Feat.	Description
1	F45	Lack of knowledge
2	F52	Lack of skill
3	F79	Safety awareness, cutting corners
4	F38	Inadequate work preparation
5	F64	LTA physical/physiological capability
6	F37	Inadequate work methods
7	F32	Inadequate procedures and check lists
8	F48	Lack of motivation/morale
9	F25	Improper supervisory example
10	F26	Inadequate briefing, instruction

Table 15: Top 10 features ranked by RF on all the classification problems described in Section 2.

RF		
Pos.	Feat.	Description
1	F38	Inadequate work preparation
2	F45	Lack of knowledge
3	F89	Training ignored
4	F15	Emergency training program
5	F05	Contingency plans not updated
6	F37	Inadequate work methods
7	F24	Improper performance of maintenance/repair
8	F01	Anthropometric factors
9	F79	Safety awareness, cutting corners
10	F52	Lack of skill

Table 16: Top 10 features ranked M-SVM-BK on all the classification problems described in Section 2.

M-SVM-BK		
Pos.	Feat.	Description
1	F15	Emergency training program
2	F89	Training ignored
3	F01	Anthropometric factors
4	F05	Contingency plans not updated
5	F24	Improper performance of maintenance/repair
6	F26	Inadequate briefing, instruction
7	F31	Inadequate manning
8	F37	Inadequate work methods
9	F38	Inadequate work preparation
10	F45	Lack of knowledge

Although the proposed data-driven approach has been proved accurate and robust, some limitations are worth to be highlighted. In particular, as human factor related studies and investigations are relatively new, a limited amount of data is available. Therefore the statistical relevance of the results is limited. Moreover, a lack of proper guidelines for accident investigators led to an inconsistent data collection procedure that needs further improvement. The data collection is a blind process where the accidents' authorities make available only anonymous data, without sufficient description of the events. Therefore the process of linking data with accident report is difficult and challenging.

The outcome of this research will allow helping the researchers and policy-makers to prioritise areas of investigation and development a set of measures to mitigate, reduce and eventually eliminate the root causes of maritime accidents. Furthermore, as the approach is human unbiased, the developed methodology can be applied to other sectors that have a high contribution of human factors in accidents. Aviation, railway and nuclear industries are potential candidates for the methodology application. Next steps of the research will consider the utilisation of a more extensive data set, including accidental data not only from the European Union (EMSA) but also global data such as the one collected by IMO.

Acknowledgements

The authors of this paper would like to acknowledge the support of the Marine Accident Investigation Branch (MAIB) who provided the required data for performing this study.

References

- Abou, S. C. (2012). Fuzzy-logic-based network for complex systems risk assessment: Application to ship performance analysis. *Accident Analysis & Prevention, 45*, 305–316.
- Akyuz, E., Ilbahar, E., Cebi, S., & Celik, M. (2017). Maritime environmental disaster management using intelligent techniques. In *Intelligence Systems in Environmental Management: Theory and Applications*.
- Azadeh, A., & Zarrin, M. (2016). An intelligent framework for productivity assessment and analysis of human resource from resilience engineering, motivational factors, hse and ergonomics perspectives. *Safety science, 89*, 55–71.
- Breiman, L. (2001). Random forests. *Machine learning, 45*, 5–32.
- Bye, R. J., & Aalberg, A. L. (2018). Maritime navigation accidents and risk indicators: An exploratory statistical analysis using ais data and accident reports. *Reliability Engineering & System Safety, 176*, 174–186.
- Castiglia, F., Giardina, M., & Tomarchio, E. (2015). Therp and heart integrated methodology for human error assessment. *Radiation Physics and Chemistry, 116*, 262–266.
- Celik, M., & Cebi, S. (2009). Analytical hfacs for investigating human errors in shipping accidents. *Accident Analysis & Prevention, 41*, 66–75.
- Chang, J. I., & Lin, C.-C. (2006). A study of storage tank accidents. *Journal of loss prevention in the process industries, 19*, 51–59.
- Chauvin, C., Lardjane, S., Morel, G., Clostermann, J.-P., & Langard, B. (2013). Human and organisational factors in maritime accidents: Analysis of collisions at sea using the hfacs. *Accident Analysis & Prevention, 59*, 26–37.

- Chen, S. T., Wall, A., Davies, P., Yang, Z., Wang, J., & Chou, Y. H. (2013). A human and organisational factors (hofs) analysis method for marine casualties using hfacs-maritime accidents (hfacs-ma). *Safety science*, *60*, 105–114.
- Cipollini, F., Oneto, L., Coraddu, A., Murphy, A. J., & Anguita, D. (2018). Condition-based maintenance of naval propulsion systems: Data analysis with minimal feedback. *Reliability Engineering & System Safety*, *177*, 12–23.
- Coraddu, A., Oneto, L., Baldi, F., & Anguita, D. (2017). Vessels fuel consumption forecast and trim optimisation: a data analytics perspective. *Ocean Engineering*, *130*, 351–370.
- Donders, A. R. T., van der Heijden, G. J. M. G., Stijnen, T., & Moons, K. G. M. (2006). Review: a gentle introduction to imputation of missing values. *Journal of clinical epidemiology*, *59*, 1087–1091.
- Eliopoulou, E., Papanikolaou, A., & Voulgarellis, M. (2016). Statistical analysis of ship accidents and review of safety level. *Safety science*, *85*, 282–292.
- EMSA (2016). *Glossary of reporting attributes*.
- EMSA (2018). *Safety Analysis of Data Reported In EMCIP*.
- Fu, G., Cao, J. L., Zhou, L., & Xiang, Y. C. (2017). Comparative study of hfacs and the 24model accident causation models. *Petroleum Science*, *14*, 570–578.
- Good, P. (2013). *Permutation tests: a practical guide to resampling methods for testing hypotheses*. Springer Science & Business Media.
- Graziano, A., Teixeira, A. P., & Guedes Soares, C. (2016). Classification of human errors in grounding and collision accidents using the tracer taxonomy. *Safety science*, *86*, 245–257.
- Grech, M. R., Horberry, T., & Smith, A. (2002). Human error in maritime operations: Analyses of accident reports using the leximancer tool. *Proceedings of the human factors and ergonomics society annual meeting*, *46*, 1718–1721.

- Guyon, I., & Elisseeff, A. (2003). An introduction to variable and feature selection. *Journal of machine learning research*, *3*, 1157–1182.
- Guyon, I., Gunn, S., Nikravesh, M., & Zadeh, L. A. (2008). *Feature extraction: foundations and applications*. Springer.
- Hänninen, M. (2014). Bayesian networks for maritime traffic accident prevention: benefits and challenges. *Accident Analysis & Prevention*, *73*, 305–312.
- Harb, R., Yan, X., Radwan, E., & Su, X. (2009). Exploring precrash maneuvers using classification trees and random forests. *Accident Analysis & Prevention*, *41*, 98–107.
- Hesse, H. (2003). Maritime security in a multilateral context: Imo activities to enhance maritime security. *The International Journal of Marine and Coastal Law*, *18*, 327–340.
- Hetherington, C., Flin, R., & Mearns, K. (2006). Safety in shipping: The human element. *Journal of safety research*, *37*, 401–411.
- Hsu, C. W., & Lin, C. J. (2002). A comparison of methods for multiclass support vector machines. *IEEE transactions on Neural Networks*, *13*, 415–425.
- Iranitalab, A., & Khattak, A. (2017). Comparison of four statistical and machine learning methods for crash severity prediction. *Accident Analysis & Prevention*, *108*, 27–36.
- Jeong, K. S., Choi, B. S., Moon, J. K., Hyun, D. J., Lee, J. H., Kim, I. J., Kim, G. H., & Kang, S. Y. (2016). Risk assessment on abnormal accidents from human errors during decommissioning of nuclear facilities. *Annals of Nuclear Energy*, *87*, 1–6.
- Jin, D., Kite-Powell, H. L., Thunberg, E., Solow, A. R., & Talley, W. K. (2002). A model of fishing vessel accident probability. *Journal of safety Research*, *33*, 497–510.
- Kirwan, B. (1997). The development of a nuclear chemical plant human reliability management approach: Hrms and jhedi. *Reliability Engineering & System Safety*, *56*, 107–133.

- Kristiansen, S. (2013). *Maritime transportation: safety management and risk analysis*. Routledge.
- Kum, S., & Sahin, B. (2015). A root cause analysis for arctic marine accidents from 1993 to 2011. *Safety science*, *74*, 206–220.
- Kumar, A. M., Rajakarunakaran, S., & Prabhu, V. A. (2017). Application of fuzzy heart and expert elicitation for quantifying human error probabilities in lpg refuelling station. *Journal of Loss Prevention in the Process Industries*, *48*, 186–198.
- Kurt, R., Arslan, V., Khalid, H., Comrie, E., Boulougouris, E., & Turan, O. (2016). Seahorse procedure improvement system: development of instrument. In *International SEAHORSE Conference on Maritime Safety and Human Factors* (pp. 1–8).
- Kurt, R., Arslan, V., Turan, O., de Wolff, L., Wood, B., Arslan, O., Kececi, T., Winkelman, J., van Wijngaarden, M., & Papadakis, G. (2015). Seahorse project: Dealing with maritime workarounds and developing smarter procedures, .
- Larose, D. T. (2015). *Data mining and predictive analytics*. John Wiley & Sons.
- Louppe, G. (2014). Understanding random forests: From theory to practice. *arXiv preprint arXiv:1407.7502*, .
- Louppe, G., Wehenkel, L., Sutera, A., & Geurts, P. (2013). Understanding variable importances in forests of randomized trees. In *Advances in neural information processing systems*.
- Luo, M., & Shin, S. H. (2016). Half-century research developments in maritime accidents: Future directions. *Accident Analysis & Prevention*, .
- Lützhöft, M. (2004). *“The technology is great when it works”: Maritime Technology and Human Integration on the Ship’s Bridge*. Ph.D. thesis Linköping University Electronic Press.
- Navas de Maya, B., Kurt, R. E., & Turan, O. (2018). Application of fuzzy cognitive maps to investigate the contributors of maritime collision accidents. In *Transport Research Arena (TRA)*.

- O'Hare, D., Wiggins, M., Batt, R., & Morrison, D. (1994). Cognitive failure analysis for aircraft accident investigation. *Ergonomics*, *37*, 1855–1869.
- Oneto, L. (2019). *Model Selection and Error Estimation in a Nutshell*. Springer.
- Oneto, L. (2020). *Model selection and error estimation in a nutshell*. Springer.
- Orlandi, I., Oneto, L., & Anguita, D. (2016). Random forests model selection. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*.
- Osman, O. A., Hajij, M., Karbalaieali, S., & Ishak, S. (2019). A hierarchical machine learning classification approach for secondary task identification from observed driving behavior data. *Accident Analysis & Prevention*, *123*, 274–281.
- Papanikolaou, A., Bitha, K., Eliopoulou, E., & Ventikos, N. P. (2014). Statistical analysis of ship accidents occurred in the period 1990-2012 and assessment of safety level of ship types. *Maritime Technology and Engineering*, .
- Petersen, J. P., Winther, O., & Jacobsen, D. J. (2012). A machine-learning approach to predict main energy consumption under realistic operational conditions. *Ship Technology Research*, *59*, 64–72.
- Polato, M., Lauriola, I., & Aiolli, F. (2018). A novel boolean kernels family for categorical data. *Entropy*, *20*, 444.
- Praetorius, G., Kataria, A., Petersen, E. S., Schröder-Hinrichs, J. U., Baldauf, M., & Kähler, N. (2015). Increased awareness for maritime human factors through e-learning in crew-centered design. *Procedia Manufacturing*, *3*, 2824–2831.
- Puisa, R., Lin, L., Bolbot, V., & Vassalos, D. (2018). Unravelling causal factors of maritime incidents and accidents. *Safety science*, *110*, 124–141.
- Rokach, L., & Maimon, O. (2008). *Data mining with decision trees: theory and applications*. World Scientific.

- Rothblum, A. M. (2000). Human error and marine safety. In *National Safety Council Congress and Expo*.
- Schlögl, M., Stütz, R., Laaha, G., & Melcher, M. (2019). A comparison of statistical learning methods for deriving determining factors of accident occurrence from an imbalanced high resolution dataset. *Accident Analysis & Prevention*, *127*, 134–149.
- Schölkopf, B., Herbrich, R., & Smola, A. J. (2001). A generalized representer theorem. In *International conference on computational learning theory*.
- Schröder-Hinrichs, J. U., Baldauf, M., & Ghirxi, K. T. (2011). Accident investigation reporting deficiencies related to organizational factors in machinery space fires and explosions. *Accident Analysis & Prevention*, *43*, 1187–1196.
- Schröder-Hinrichs, J. U., Hollnagel, E., Baldauf, M., Hofmann, S., & Kataria, A. (2013). Maritime human factors and imo policy. *Maritime Policy & Management*, *40*, 243–260.
- Sculley, D. (2007). Rank aggregation for similar items. In *SIAM International Conference on Data Mining*.
- Shalev-Shwartz, S., & Ben-David, S. (2014). *Understanding machine learning: From theory to algorithms*. Cambridge university press.
- Shappell, S. A., & Wiegmann, D. A. (2000). The human factors analysis and classification system—hfacs, .
- Shawe-Taylor, J., & Cristianini, N. (2004). *Kernel methods for pattern analysis*. Cambridge university press.
- Shi, X., Wong, Y., Li, M., & Chai, C. (2018). Key risk indicators for accident assessment conditioned on pre-crash vehicle trajectory. *Accident Analysis & Prevention*, *117*, 346–356.
- Smith, T. W. P., O’Keeffe, E., Aldous, L., & Agnolucci, P. (2013). Assessment of shipping’s efficiency using satellite ais data. In *International Council on Clean Transportation*.
- Soner, O., Asan, U., & Celik, M. (2015). Use of hfacs-fcm in fire prevention modelling on board ships. *Safety Science*, *77*, 25–41.

- Tirunagari, S., Hanninen, M., Stanhlberg, K., & Kujala, P. (2012). Mining causal relations and concepts in maritime accidents investigation reports. *International Journal of Innovative Research and Development*, 1, 548–566.
- Turan, O., Kurt, R. E., Arslan, V., Silvagni, S., Ducci, M., Liston, P., Schraagen, J. M., Fang, I., & Papadakis, G. (2016). Can we learn from aviation: safety enhancements in transport by achieving human orientated resilient shipping environment. *Transportation research procedia*, 14, 1669–1678.
- Ung, S. T. (2015). A weighted cream model for maritime human reliability analysis. *Safety science*, 72, 144–152.
- Ung, S. T. (2019). Evaluation of human error contribution to oil tanker collision using fault tree analysis and modified fuzzy bayesian network based cream. *Ocean Engineering*, 179, 159–172.
- Vapnik, V. N. (1998). *Statistical learning theory*. Wiley New York.
- Ventikos, N. P., Papanikolaou, A. D., Louzis, K., & Koimtzoglou, A. (2018). Statistical analysis and critical review of navigational accidents in adverse weather conditions. *Ocean Engineering*, 163, 502–517.
- Wang, J. (2002). Offshore safety case approach and formal safety assessment of ships. *Journal of safety research*, 33, 81–115.
- Yang, K., Tao, L., & Bai, J. (2014). Assessment of flight crew errors based on therp. *Procedia Engineering*, 80, 49–58.
- Yang, Z. L., Bonsall, S., Wall, A., Wang, J., & Usman, M. (2013). A modified cream to human reliability quantification in marine engineering. *Ocean Engineering*, 58, 293–303.
- Yıldırım, U., Başar, E., & Uğurlu, Ö. (2017). Assessment of collisions and grounding accidents with human factors analysis and classification system (hfacs) and statistical methods. *Safety Science*, .
- Zhang, S., He, W., Chen, D., Chu, J., & Fan, H. (2019). A dynamic human reliability assessment approach for manned submersibles using pmv-cream. *International Journal of Naval Architecture and Ocean Engineering*, .

Zhou, Q., Wong, Y. D., Xu, H., Van Thai, V., Loh, H. S., & Yuen, K. F. (2017). An enhanced cream with stakeholder-graded protocols for tanker shipping safety application. *Safety science*, *95*, 140–147.