Natural Language Processing and Data-Driven Methods for Aviation Safety and Resilience: from Extant Knowledge to Potential Precursors

Irene Buselli, Luca Oneto, Carlo Dambra
ZenaByte s.r.l.
Genova, Italy

Christian Verdonk Gallego, Miguel García Martínez
CRIDA
Madrid, Spain

Anthony Smoker, Nnenna Ike
School of Aviation
Lund University
Ljungbyhed, Sweden

Patricia Ruiz Martino
Safety, Quality and Environment Division
ENAIRE
Madrid, Spain

Tamara Pejovic
Aviation Intelligence Unit
EUROCONTROL
Brussels, Belgium

***Abstract*—Demand upon the future Air Traffic Management (ATM) system will possibly grow to exceed available system capacity, pushing forward the need for automation and digitisation to maintain safety while increasing efficiency. This work focuses on a manifestation of ATM safety, the loss of separation (LoS), and its analysis via Natural Language Processing (NLP) and Data-Driven Methods (DDMs), able to extract meaningful and actionable information from the LoS-related data. These data are, primarily, safety reports and ATM-system data (e.g, flights information, radar tracks, and Air Traffic Control events).**

Current research in this field mainly exploits NLP to categorise the reports and DDMs to predict safety events. The limitation of current NLP-based approaches is that the considered categories need to be manually annotated by experts and general taxonomies are seldom exploited. At the same time, current DDMs are rarely able to support safety practitioners in the process of investigation of an incident after it happened.

To fill these gaps, the authors propose to (i) perform Exploratory Data Analysis on safety reports combining state-of-the-art techniques like topic modelling and clustering, then to (ii) develop an algorithm able to extract the recent Toolkit for ATM Occurrence Investigation (TOKAI) taxonomy factors from the free-text safety reports based on Syntactic Analysis, and finally to (iii) develop a DDM able to automatically assess if the Pilots or the Air Traffic Controller (ATCo) or both contributed to the incident, almost immediately after the LoS.

The results on LoSs reported in the public database of the Comisión de Estudio y Análisis de Notificaciones de Incidentes de Tránsito Aéreo (CEANITA) support the authors' proposal.

***Keywords*—ATM; Safety; Digitisation; Resilience; Resilient Performance; Data-Driven Models; Natural Language Processing; Safety Reports; Text Mining; TOKAI; Taxonomy.**

I. Introduction

Demand upon the future Air Traffic Management (ATM) system will possibly grow to exceed available system capacity at the same time as the economic challenges for both service providers — i.e., Air Navigation Service Providers (ANSPs) and aircraft operators — will manifestly grow in intensity [1]. The Single European Sky ATM Masterplan [2] defines the philosophy and concept of operations that will lead to the modernisation of the European ATM system. The goal is to provide the system with a sustainable capacity that is able to absorb this growth through an efficient and effective management of the ATM system, maintaining safety while increasing efficiency. A cornerstone of the ATM Masterplan is to further deploy automation and digitisation tools, leading to a significant integration of human and technical systems [3]. Therefore, organisations need to adopt new approaches to understand system safety performance of an increasingly complex operational environment with new actors and stakeholders as well as maintain extant approaches [4].

How organisations are able to quantify and understand the impact of changes in the ATM system is the focal research object of the FARO project - saFety And Resilience guidelines for aviatiOn. In this work, which is framed within the context of FARO, the authors focus on a manifestation of ATM safety, the loss of separation (LoS). There are two main sources of data which can inform about what happened during a LoS:

- the Safety Reports produced by states' Civil Aviation Authorities and ANSPs after investigating the safety-related events;
- the Automatic Safety Monitoring Tools (ASMTs), which allow the monitoring and recording of safety-related events. These tools are usually augmented with ATM system data, which gather surveillance (e.g., flight tracks) and operational data (e.g., ATC events) [5].

The LoS events considered in this study are the ones reported in the public database of the Comisión de Estudio y Análisis de Notificaciones de Incidentes de Tránsito Aéreo (CEANITA).

This work exploits Natural Language Processing (NLP) and Data-Driven Methods (DDMs) towards the automation of both the analysis of the above-mentioned reports and, partly, the early estimation of the investigation results through ASMTs data, for safety and resilience purposes.

For what concerns the application on NLP to safety reports, research has largely focused on developing models and algorithms for categorising incident reports [6]–[9]. All of these works rely on an initial set of labels and training data that include incident reports previously labelled by domain experts. The biggest limitation of this approach is the lack of generality: it would take substantial effort to generate a new set of labels and training data. In this framework, the importance of referring to a common taxonomy became evident. On one hand, tools like the Toolkit for ATM Occurrence Investigation (TOKAI) have been developed to generate structured safety data [10], and their outcomes have been deeply analysed [11]; on the other hand, NLP techniques have been applied to categorise the safety reports according to taxonomy factors [12]. Another limitation of the categorisation approach is that it just aims at automating a task performed by domain experts,

without trying to add further knowledge or to discover unknown patterns. To overcome this limitation, in recent years researchers have focused on topic modelling [13]–[16] and similarity clustering [13], [17], which have been shown to be extremely valuable tools. As regards the application of DDMs in the ATM scope, research has focused on a number of different fields, such as taxi-out time prediction [18], [19], trajectory prediction [20], [21], air traffic flow extraction [22], [23], and flight delay prediction [24], [25]. In the safety scope, some relevant applications of DDMs are proposed in literature to predict safety events or performance [26], [27], or to provide safety metrics [28] or accident precursors [29]. However, there are very few references aiming at supporting safety practitioners in facilitating the investigation of an incident after it happened but before it is reported [30].

To fill the current gaps in the literature, in this work the authors propose a threefold approach:

- First, an Exploratory Data Analysis (EDA) was performed on safety reports combining state-of-the-art techniques like topic modelling and clustering.
- Then, for the first time, an algorithm able to extract TOKAI taxonomy factors from the free-text safety reports was developed, based on Syntactic Analysis.
- Finally, for the first time, an Automatic Contribution Assessment model was developed, able to leverage data to assess if the Pilots or the Air Traffic Controller (ATCo) or both contributed directly to the incident, almost immediately after the LoS and before investigation.

The first two steps of the proposed approach focus on the mining of free-text safety reports through NLP techniques, with the purpose of identifying hidden patterns (e.g., recurrent behaviours during LoS events) via topic modelling and clustering and of associating patterns of behaviour to TOKAI taxonomy factors (e.g., perception, conformance to procedures, or memory). The choice of the TOKAI taxonomy is due to many reasons. The first one is that it is particularly suited to allow aggregation at different levels. The second one is that it makes a significant shift from traditional causal taxonomies based on negative perspectives (i.e., describing errors or failures) thanks to its neutralised language: TOKAI factors are neither negatively nor positively oriented, so they can be ideally used to explain both ordinary work situations and safety occurrences [11]. This is aligned with a basic principle of Resilience Engineering: successes and failures do not emerge from different pathways through the work system, but the same set of conditions can evolve to either [31], [32]. The last reason is that the TOKAI, as developed by EUROCONTROL, is intended to harmonise future investigations and to allow ANSPs to share lessons from ATM occurrences: the automatic identification of TOKAI factors could help harmonising not only future analyses, but also the existing ones.

The last step of the proposed approach aims at partially automating the process of contribution assessment (which may take weeks to be completed by human practitioners) based on the ATC events registered in temporal proximity to the LoS (which are readily available). Indeed, there are a number of safety-related occurrences which went unnoticed by the old systems, which can now be identified thanks to the new ASMTs. As a consequence, probably many more LoSs will need to be investigated and studied in the future. Since human review of incidents is an extensive process, providing the ability to partially estimate the results of these investigations timely (a few minutes after the LoS) would facilitate the safety practitioners in prioritising the investigations and in understanding potential precursors of these LoS events.

II. Scope of the Work

The scope of this work is to automatically extract meaningful and actionable information from CEANITA LoS reports and related contextual data (i.e., radar tracks of the aircraft involved, flights information, and related ATC events) with a particular focus on human contributing factors. To achieve this scope, a threefold approach was applied.

- First, an EDA was performed in order to get insights into the incidents phenomena. Initially, the most recurrent topics in the corpus of CEANITA LoS reports were automatically identified using unsupervised NLP techniques, in particular Topic Modelling [33]. The prevalence of the different topics in each report was then computed, obtaining numerical features able to describe at high level the content of the reports and compare them without the need to actually read and understand them. Furthermore, the combination of the above-mentioned features with other structured information extracted from the very same reports led to the development of a Cluster Analysis [34], which automatically grouped incidents that appeared to be similar.
- Then, an algorithm able to extract TOKAI taxonomy factors from the free text of CEANITA reports was developed, based on Syntactic Analysis. Every CEANITA report contains in its conclusions a free-text description of the main actions performed by ATCo and Pilots before and after the incident. Many of these actions are crucial factors in the dynamic of the LoS. Thus, it can be fundamental to be able to automatically extract these actions from the free-text conclusions and then to classify them according to a standard taxonomy (in our case the TOKAI one). Indeed, this would enable the application of quantitative-analysis techniques (e.g., to structure a proactive risk-assessment strategy) on these actions [11]. The authors used state-of-the-art tools for Syntactic Analysis [35] to estimate the occurrences of actions associated to each TOKAI taxonomy factor, together with their subject (i.e., usually Pilots or ATCo).
- Finally, an Automatic Contribution Assessment model able to leverage the ATC events to assess whether the Pilots or the ATCos or both contributed to the incident was developed. The model was able to assess contribution before (i.e., 10 minutes after the incident) human evaluation (which is usually concluded even weeks after the incident). This data-driven model [36] leverages recorded ATC events and other contextual data (i.e., radar tracks of the aircraft and flight information) to make its prediction.

III. Data Description

For the purpose of this study there were two main data sources available: CEANITA reports (see Sections III-A) and structured data from ENAIRE-CRIDA data warehouse, containing contextual information about the LoSs together with ATC events (see Section III-B).

A. CEANITA LoS Reports

The considered CEANITA LoS reports consist of 89 safety reports, written in Spanish and published by Spanish Safety Aviation Agency (AESA), covering safety-related occurrences that happened in the Spanish airspace between January 2018 and July 2019. These incidents reported by CEANITA are just a subset of the total amount of losses of separation, where high-severity incidents are over-represented.

The initial sections of these reports are written in fixed formulas or tabular format, enabling the direct extraction of some categorical or numerical variables, such as:

- the *main causes*: the most frequent ones in the corpus are wrong clearance (52%), deviation from procedures (22%), wrong or no resolution (17%-15%), coordination problems (17%), and late or no detection (15%-16%) — note that multiple causes are possible;
- the *airspace class*: most of the reported incidents happened in class C, D (40% each), and A (11%), while only 6% in G and 3% in E (classes assigned according to the ICAO classification [37]);
- the *Pilots and ATCo contribution*: Pilots contribution is classified as direct in 36% of the cases, as indirect in 15%, and as none in 49%. ATCo contribution is, instead, direct in the majority of cases (72%), indirect in 9% of the incidents, and none in 19%.

The remaining part of each report is written as a free text.

B. ENAIRE-CRIDA Contextual Information

The contextual information, arranged in structured form, was provided by ENAIRE-CRIDA. In particular, they provide high-granularity ATM data such as flight plans, flight tracks, and ATM-processed information about the Spanish airspace. More precisely, two main sources were exploited:

- flight tracks and related contextual flight information (e.g., type, speed, and heading);
- ATC events of the interactions between ATCos and the Controller Working Position (CWP).

The integration of these sources (only needed for the development of the DDM) led to the reduction of the sample from the initial 89 incidents to 70, since not all the LoSs could be linked to structured data with sufficient certainty as flights involved in the incidents are anonymised in CEANITA reports.

IV. Methods

This section presents the methods and tools exploited to achieve the scope of the work (see Section II) leveraging the data described in Section III. Four main technologies are exploited: Topic Modelling (Section IV-A), Clustering Analysis (Section IV-B), Syntactic Analysis (Section IV-C), and Data-Driven Predictive Models (Section IV-D).

A. Topic Modelling

Topic Modelling is an NLP method initially designed by David Blei and John Lafferty [33]. The aim is to represent a collection of documents in terms of a certain number or topics (i.e., latent dimensions), calculated in a completely unsupervised fashion, based only on the distribution of words in the documents. Topic Modelling has already been widely exploited in the transportation domain, since it is particularly suited to summarise the main themes in a corpus of documents [38], [39]. The statistical intuition behind Topic Modelling can be summarised in three points:

- A document can be defined as a set of words.
- A document contains different topics according to a certain distribution.
- A topic can be defined through words according to a certain distribution.

As a consequence, by observing a collection of documents, one can empirically estimate the two distributions that fit the observed frequencies of words in documents. In the most widely used technique for Topic Modelling, the Latent Dirichlet Allocation (LDA), the estimation of these distributions is based on the Dirichlet probability distribution [13], [14]. This intuitive framework also originates a topic-word matrix in which each topic is represented through weights associated to each word. This information can be used to interpret the (otherwise unlabelled) topics.

TABLE I. Example of Syntactic Analysis with UDPipe for the sentence "Sector SAU instruye a la aeronave 2 a proceder directo a el punto LOTEE" ("SECTOR SAU clears Aircraft 2 to proceed direct to the point LOTEE"). The meaning of "Part of Speech" and "Dependency" elements is standard[A].

Sentence	Lemma	Part of Speech	Dependency
Sector	sector	noun	nsubj
SAU	SAU	propn	appos
instruye	instruir	verb	root
a	a	adp	case
la	el	det	det
aeronave	aeronave	noun	obj
2	2	num	nummod
a	a	adp	mark
proceder	proceder	verb	advcl
directo	directo	adj	advmod:lmod
a	a	adp	case
el	el	det	det
punto	punto	noun	obl
LOTEE	LOTEE	propn	appos

[a]https://universaldependencies.org/u/dep/all.html

B. Clustering Analysis

Clustering Analysis [34] allows the grouping of data in a database according to a definition of similarity. In this context, Hierarchical Clustering is one of the most widely exploited methods [40]. In particular, the agglomerative Hierarchical Clustering, as opposed to the divisive one, was used in this work, since it has been shown to be the most effective [40]. The idea behind the agglomerative Hierarchical Clustering is simple: at the beginning, each point in the database is considered as an individual cluster. Then, each cluster is merged with other clusters until the data converge to a single cluster. Finally, the practitioner has to select the best number of clusters based on the knowledge of the subject, or the intra-cluster variability, or exploiting particular statistical metrics [41]. A crucial issue is how to map the data in the database into a space where a definition of distance well describes the notion of data similarity. In this case, data were merged according to Ward's minimum variance criterion.

C. Syntactic Analysis

Syntactic Analysis is the process of analysing a string in natural language to identify the syntactic relations between words. In this work, Syntactic Analysis is performed through the UDPipe [35], a state-of-the-art open-source library which automatically generates sentence segmentation, tokenisation, part-of-speech tagging, lemmatisation, and dependency parsing. Models are provided for 50 languages. An example of the output of the UDPipe library can be found in Table I. A detailed explanation of the UDPipe library can be found in [35].

D. Data-Driven Predictive Models

Data-driven predictive models are able to learn relations between inputs (e.g., ATC events) and outputs (e.g., incident direct contribution) based on a series of examples (i.e., historical data).

In this context two (Shallow) Machine Learning algorithms, Support Vector Machines (SVMs) [42] and Random Forests [43], represent state-of-the-art solutions for many real-world applications [44], [45] — at least when Deep Learning algorithms cannot be applied due to limited data availability.

SVMs are the most effective algorithms in the family of Kernel Methods [42] (i.e., methods exploiting the "kernel trick" to extend linear techniques to the solution of nonlinear problems). SVMs have a series of hyperparameters which deeply influence their performance and need to be tuned during the model selection phase [46]: the kernel, the kernel hyperparameter, and the complexity hyperparameter.

Random Forests, instead, are one of the most effective approaches in the family of the ensemble methods [43]. It is a tree-based ensemble algorithm, combining bagging to random-subset feature selection. In bagging, each tree is independently constructed using a bootstrap sample of the dataset. Random Forests add a further layer of randomness to bagging, also changing how trees are constructed (the best split at each node of the tree is chosen among a subset of predictors randomly sampled at that node). Eventually, a simple majority vote is taken for prediction. Random Forests are less influenced by their hyperparameters [47], even if the number of trees and features to be sampled still need to be tuned.

As just described, the data-driven predictive models need to be tuned, but, at the same time, their performance needs to be estimated in a rigorous statistical way, in order to estimate their behaviour in production environment. Model Selection and Error Estimation deal exactly with this problem [46]. Resampling techniques like k-fold cross validation and non-parametric bootstrap are commonly exploited solutions, which work well in many situations [46]. The idea is that the original dataset is re-sampled once or more, without replacement, to build three independent datasets called learning, validation, and test set. The learning set is exploited to train the model, the validation set to find the optimal hyperparameters (namely the ones that lead to the optimal performance), and the test set to estimate the performance of the final model: in this way, the test is independent from both the learning and the validation, so results are statically sound [48]. Performance measures strongly depend on the task to be solved. In this case, dealing with classification problems, Accuracy, Confusion Matrix, Area Under the Receiving Operating Characteristics (AUC), F1 score, Sensitivity, and Specificity are the most commonly used metrics [36].

Once the model is built and has been confirmed to be sufficiently effective, it can be of interest to investigate how this model is affected by the different input features [49], [50]. This procedure is called Feature Ranking and allows the user to detect if the features are appropriately taken into account by the learned models, from the perspective of the domain experts. In particular, Feature Ranking based on Random Forest via Mean Decrease in Accuracy (i.e., the importance of each feature is assessed by randomly permuting the values of the feature and measuring the resulting increase in error) is one of the most effective techniques [51], [52].

V. Experimental Results

This section shows how the methods presented in Section IV were exploited to achieve the scope of the work (see Section II) demonstrating the effectiveness of the proposed approach on the data described in Section III. Specifically, Section V-A presents the results of EDA, obtained first by exploiting Topic Modelling to extract the main topics from the CEANITA reports and then by clustering the different incidents. Subsequently, Section V-B presents the results of Syntactic Analysis applied to the same reports to connect them with the TOKAI taxonomy, validating also the quality of the methodology. Finally, Section V-C reports the performance of the data-driven model in estimating who directly contributed to the incident before the actual human evaluation.

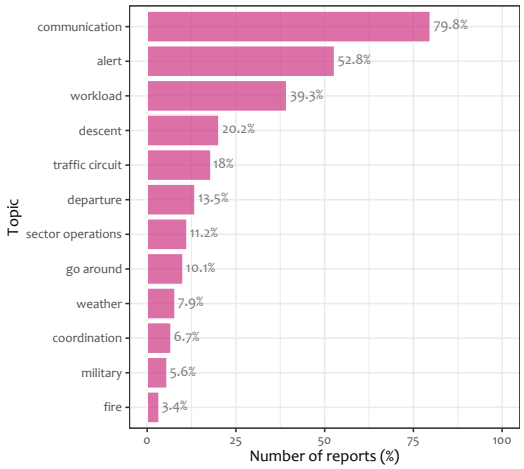

Figure 1. Prevalence of each of the 12 topics of Table II over CEANITA reports.

A. Exploratory Data Analysis

As a first step in EDA, this section shows how Topic Modelling (see Section IV-A) can extract the main topics from the 89 CEANITA reports.

With this technique, the reports can be organised according to the discovered topics. Indeed, the probability of finding each topic, namely the prevalence, can be associated to every report, generating a set of numerical features describing the document. The use of LDA for Topic Modelling led to the identification of 12 main topics. This result is obtained combining both automated procedures (optimising the coherence metrics) and more handcrafted fine-tuning (feedback from FARO operational experts, which allowed the selection of the most significant and coherent topics according to their domain knowledge). These 12 topics can be described by words and bigrams, to which experts have associated representative labels (see Table II). Topic Modelling results allow a finer granularity than simple descriptive analysis: while the main causes of the incident are identified using variables described in Section III-A (e.g., a wrong ATCo clearance was responsible), topic modelling also provides additional information (e.g., if the ATCo's wrong clearance was due to excessive workload or an emergency situation).

Figure 1 shows the prevalence of each topic over the different reports. Observing Figure 1 one can easily observe that exogenous factors like fire or weather problems are quite rare (only about 10% of the incidents contain one of these topics), while workload is present in around 40% of the reports.

At this point, as a second step of the EDA, a further analysis was conducted to find the relation of the 12 topics with the main causes of the incidents applying Clustering Analysis (see Section IV-B).

For this purpose, for each CEANITA report, a feature set was created, composed of the prevalence of the topics, the main causes, and the level of Pilots' and ATCo's contribution to the incident. Subsequently, Hierarchical Clustering with Ward distance was applied on the resulting dataset. After looking at dendrograms and screeplots (i.e., the two most common methods for cluster selection, which are not reported

TABLE II. WORDS AND BIGRAMS OF THE 12 TOPICS EXTRACTED WITH LDA FROM CEANITA REPORTS, TOGETHER WITH THE REPRESENTATIVE LABEL ASSOCIATED TO EACH TOPIC BY FARO'S EXPERTS (ENGLISH TRANSLATION FROM SPANISH).

Words/Bigrams							Topic
helicopter	drop	water	fires	extinguishing	coordination	drop area	fire
load	work	high	alone	workload	instructions	previous	workload
departure	to take off	aircraft climb	runway	to take off aircraft	rate	they are	departure
wind	tail	down-wind	leg	wind leg	right tail	runway	traffic circuit
weather	adverse	adverse weather	detours	meteorologic conditions	due to weather	thunderstorm	weather
runway	go around	go	around	to take off	to land	aircraft established	go around
sectors	sector aircraft	frequency sector	high	coordination	transfer	limit	sector operations
answer	received	finally	decided	they saw	communication	visual contact	communication
clearance	course descent	aircraft to descend	descent rate	sector to descend	aircraft to maintain	rate	descent
received	coordinating	confirming	to confirm receipt	maintaining formation	sector informs	receipt	coordination
alert	early	early alert	activation function	activation	function	alert function	alert
military	military formation	formation	military aircraft	defence	air defence	main centre	military

for space constraints) together with the FARO experts, 8 different clusters were identified:

- Two very small subgroups are identified as particularly different from the others: one is composed of three incidents where the main topic is "fire" (indeed, they are the reports referred to Llutxent fire in summer 2018), while the other contains the three incidents caused by level bust.
- The largest cluster is mainly composed of wrong-clearance and late-detection incidents, with clearly the highest frequency of ATCo contribution and an interesting high prevalence of "descent" topic.
- The fourth cluster contains incidents mainly caused by "wrong resolution" of the ATCo, with high prevalence of topics related to go-around, departure, and weather.
- The fifth cluster is composed of incidents caused mainly by transfer or coordination problems. The most frequent topics here are "sector operations" and "military".
- Incidents in the sixth cluster are essentially due to Pilots' errors, in particular to airspace infringement and unfulfillment of the Visual Flight Rules (VFR).
- The seventh cluster is characterised by incidents due to Pilots' deviations from procedures, especially in the landing phase (topic "traffic circuit").
- The last cluster is composed of incidents due to ATCo inability to both detect and resolve the LoS. This cluster is interestingly characterised by high values of the topic "alert";

B. Automatic Extraction of TOKAI Taxonomy Factors

The exploitation of Syntactic Analysis (see Section IV-C) enables the association of each CEANITA report to the TOKAI taxonomy factors. In particular, for the purposes of this research, only Part A of the TOKAI taxonomy was exploited, namely the one related to the Personnel, since the actions reported in the conclusions are usually more related to this subject. Table III reports Part A of the TOKAI taxonomy factors together with their specifications [11] and examples of sentences associated to the taxonomy by the developed tool.

At this point, the algorithm to link each CEANITA report to the TOKAI taxonomy factors can be presented (see Algorithm 1). Given the results of Algorithm 1, after grouping subjects into Flight elements (i.e., aircraft, pilot, etc.) and Ground elements (i.e., controller, sector, etc.), it is possible to estimate for each CEANITA report how the 5 factors are distributed, both in terms of positive and negative occurrences.

Figure 2 shows the global distribution of negative occurrences of each TOKAI-taxonomy factor by group of subjects. Figure 2 suggests that the main omissions for the Flight subjects are classified as factor A-4 and A-5 (e.g., problems with action or conformance with rules), while for the Ground

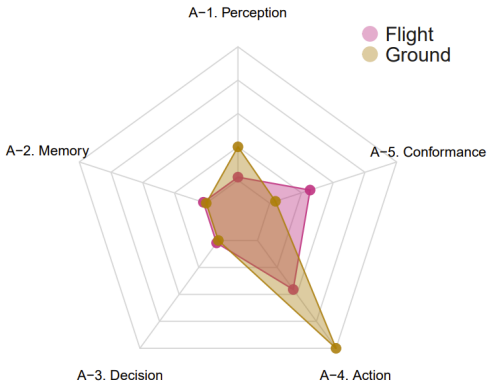

Figure 2. Global distribution of negative occurrences of each TOKAI-taxonomy factor by group of subjects.

subjects they are mostly classified as factor A-4 and A-1 (e.g., problems with action or perception). Interestingly, further analysing the data, it is possible to discover that all the problems with factor A-4 are relative to conveyance of information, for both Flight and Ground elements.

The proposed algorithm (Algorithm 1) could not be directly validated in the standard way since there is no ground truth. However, in order to validate it at least indirectly, a simple predictive model was developed to predict the main contribution (ATCo or Pilots) in an incident based on the extracted number of positive and negative occurrences of each taxonomy factor (i.e., the output of Algorithm 1). A good performance of this predictive model would indicate that the extracted information is reasonably accurate, since TOKAI taxonomy factors should well describe the ATCo's and Pilots' contribution to the event. Specifically, for each LoS, the goal was to predict:

- the Pilots' contribution, i.e., classified as direct or not;
- the ATCo's contribution, i.e., classified as direct or not;

based on:

- the number of positive and negative occurrences of each taxonomy factor (the outputs of Algorithm 1);
- the differences in prevalence between Flight and Ground elements for each taxonomy factor;
- the airspace class (in fact, similar behaviours of ATCo and Pilots can lead to different contribution assessments in different airspace classes due to different regulations).

Note that ATCo and Pilots can both have direct contribution to the incident.

Table IV reports the confusion matrices of the developed predictive models. In particular, a SVM with Gaussian Kernel

TABLE III. PART A OF THE TOKAI TAXONOMY FACTORS: SPECIFICATIONS AND EXAMPLES OF SENTENCES ASSOCIATED TO THE TAXONOMY BY THE DEVELOPED TOOL.

Factor	Specifications	Example
A-1. Perception	See - identification; See - detection; Hear - identification; Hear - detection; Perceive visual information - accuracy; Perceive auditory information - accuracy.	Sector CAO authorised aircraft 1 without detecting aircraft 2.
A-2. Memory	Remember to monitor or check; Remember to act; Remember previous actions; Recall information from working memory; Recall information from long-term memory.	Aircraft 2 was authorised by the Sector, not remembering presence of Aircraft 1.
A-3. Decision	Judge/Project; Decide/Plan.	APP LEMG planned the approximation sequence incorrectly.
A-4. Action	Select/Position manually; Convey/Record information.	Aircraft 1 did not communicate its position correctly.
A-5. Conformance	Deliberate or malicious act; Individual conformance with rules or procedures; Team conformance with rules or procedures.	Aircraft 2 did not comply with the instruction.

Algorithm 1: Algorithm to link each CEANITA report to the TOKAI taxonomy factors exploiting Syntactic Analysis.

Input: **1.** The sequences of verbs/actions in the base form for each factor (e.g., for factor A-1, the list "see", "identify", "detect", "hear", etc.). This sequences can be created directly by human operators, which can be supported by automatic tools. Possibly, two sequences can be created for each factor, a positive and a negative one (e.g., for A-2, "remember" is in the positive sequence, while "forget" in the negative one).
2. The text of the conclusive section of the CEANITA report of interest.

Output: For each of the factors (i.e., A-1, A-2, etc. in Table III) and for each subject (e.g., pilot or controller) the number of positive and negative occurrences.

1 The text of the report is processed via UDPipe (see Section IV-C and Table I as reference);
2 In the UDPipe output (i.e., the result of lemmatisation, Part-of-Speech tagging, and dependency parsing) we search, for each of the factors, the verbs in factor's lists (both for the positive and negative lists);
3 For each of the identified verb, the subject is retrieved, also taking into account passive forms where the subject is the agent;
4 A check for negative forms or adverbs (e.g., "incorrectly") is performed in the identified sentence to cope with the inversion of meaning (i.e., positive verbs become negative if a negative form or adverb is present) ;

TABLE IV. CONFUSION MATRICES (%) ON THE DUMMY PREDICTIVE PROBLEM (I.E., ESTIMATE ATCO'S AND PILOTS DIRECT CONTRIBUTION BASED ON OUTPUTS OF ALGORITHM 1) VIA SVM TO VALIDATE ALGORITHM 1.

(A) PILOTS CONTRIBUTION

		Pred. No	Pred. Yes
Truth	No	51.6±0.1	12.4±0.1
	Yes	4.5±0.3	31.5±0.3

(B) ATCO CONTRIBUTION

		Pred. No	Pred. Yes
Truth	No	25.8±0.2	3.4±0.2
	Yes	11.2±0.2	59.6±0.2

trained with the 89 CEANITA reports was used, performing accurate model selection (the kernel and the complexity hyperparameters were searched in $\{10^{-4.0}, 10^{-3.5}, \cdots, 10^{3.0}\}$ according to what described in Section IV-D). The confusion matrices computed on the test set are presented below.

Confusion matrices in Table IV appear to be reasonably balanced, especially considering that the classes are highly unbalanced. The global accuracy of the prediction is ≈83% for Pilots contribution and ≈85% for ATCo contribution. Therefore, it can be stated that:

- the proposed approach (Algorithm 1) is able to automatically link each CEANITA report to the TOKAI taxonomy factors exploiting Syntactic Analysis;
- an indirect validation performed with a dummy prediction problem showed promising performance;
- a side result of this indirect validation is that the extracted link between CEANITA reports and TOKAI taxonomy is actually a good proxy of the contribution assessment.

C. Automatic Contribution Assessment

After the EDA (Section V-A) and after linking CEANITA reports and TOKAI taxonomy (Section V-B), a data-driven model (see Section IV-D) was exploited to assess agents' contribution before (i.e., 10 minutes after the incident) human evaluation (which is a post-operation activity) based on the automatic analysis of ATC events and other contextual data (i.e., radar tracks of the aircraft and flight information). Furthermore, the analysis shows that this predictive model actually captured meaningful relations and not just spurious correlations from the data (see Section IV-D).

Specifically, for each incident, the goal was to predict:

- the Pilots' contribution, i.e., classified as direct or not;
- the ATCo's contribution, i.e., classified as direct or not;

based on:

- the flight type;
- the flight rule at the moment of the incident;
- the flight level at the moment of the incident;
- the airspace class at the moment of the incident;
- for each of the 15 classes of ATC events (see Section III-B) recorded from 30 minutes before to 10 minutes after the incident, their number of occurrences. Considering this time window is fundamental since the contributions of ATCo and Pilots depend both on what was done to prevent the potential LoS and on how it was managed when it became an actual LoS;

engineering a total of 19 features.

In this case, a Random Forest model was used (see Section IV-D), trained on the 70 incidents for which recorded ATC events were available (the number of trees was set to 1000 and the number of predictors to be randomly sampled during trees construction was searched in $\{5, 6, 7, 8, 9\}$ according to what was described in Section IV-D). Random Forests facilitate the generation of different optimal models changing the cut-off of the voting (i.e., how many trees need to agree to decide for a particular class). By doing so, it was possible to report different models, maximising respectively: the AUC, the Sensitivity, and the Specificity. Moreover, Random Forests provide the confidence of the prediction: this allows the user to trust the model only when its confidence is higher than a certain threshold.

Table V reports the confusion matrices of the developed predictive models (maximising AUC, Sensitivity, and Specificity) for both ATCos' and Pilots' contributions.

Table VI, instead, reports the confusion matrices of the predictive models (maximising the AUC, since they appeared

TABLE V. CONFUSION MATRICES OF THE DEVELOPED PREDICTIVE MODELS OF CONTRIBUTION BASED ON THE ATC EVENTS (MAXIMISING AUC, SENSITIVITY, AND SPECIFICITY) FOR ATCO AND PILOTS CONTRIBUTIONS.

(A) PILOTS CONTRIBUTION (MAXIMISING AUC)

		Pred. No	Pred. Yes
Truth	No	46.0±0.3	18.3±0.3
	Yes	6.9±0.3	28.8±0.3

(B) ATCO CONTRIBUTION (MAXIMISING AUC)

		Pred. No	Pred. Yes
Truth	No	20.8±0.3	4.9±0.3
	Yes	14.1±0.3	60.2±0.3

(C) PILOTS CONTRIBUTION (MAXIMISING SENSITIVITY)

		Pred. No	Pred. Yes
Truth	No	33.9±0.3	30.4±0.3
	Yes	0.1±0.2	35.6±0.2

(D) ATCO CONTRIBUTION (MAXIMISING SENSITIVITY)

		Pred. No	Pred. Yes
Truth	No	11.9±0.3	13.8±0.3
	Yes	1.4±0.2	72.9±0.2

(E) PILOTS CONTRIBUTION (MAXIMISING SPECIFICITY)

		Pred. No	Pred. Yes
Truth	No	64.3±0.0	0.0±0.0
	Yes	20.3±0.2	15.4±0.2

(F) ATCO CONTRIBUTION (MAXIMISING SPECIFICITY)

		Pred. No	Pred. Yes
Truth	No	24.9±0.2	0.8±0.2
	Yes	45.2±0.3	29.1±0.3

TABLE VI. CONFUSION MATRICES OF THE DEVELOPED PREDICTIVE MODELS BASED ON THE ATC EVENTS (MAXIMISING AUC) FOR BOTH ATCO AND PILOTS CONTRIBUTIONS WHEN PREDICTIONS ARE TRUSTED ONLY IF THEIR CONFIDENCE IS HIGHER THAN 60% AND 75%.

(A) PILOTS CONTRIBUTION (CONFIDENCE ≥60%)

		Pred. No	Pred. Yes
Truth	No	58.1±0.3	9.3±0.3
	Yes	4.7±0.3	27.9±0.3

(B) ATCO CONTRIBUTION (CONFIDENCE ≥60%)

		Pred. No	Pred. Yes
Truth	No	20.4±0.3	6.1±0.3
	Yes	8.2±0.3	65.3±0.3

(C) PILOTS CONTRIBUTION (CONFIDENCE ≥75%)

		Pred. No	Pred. Yes
Truth	No	60.0±0.0	00.0±0.0
	Yes	3.3±0.1	36.7±0.1

(D) ATCO CONTRIBUTION (CONFIDENCE ≥75%)

		Pred. No	Pred. Yes
Truth	No	29.0±0.1	3.2±0.1
	Yes	3.2±0.2	64.6±0.2

to be the most balanced ones) when predictions are considered only if their confidence is higher than 60% and 75%.

Table V shows that:

- when the AUC is maximised (i.e., assuming the user wants a balanced accuracy on both "Yes" and "No" classes), accuracy reaches ≈75% for Pilots contribution and ≈81% for ATCo; F1 score is ≈70% for Pilots and ≈86% for ATCo.
- when the Sensitivity is maximised, (i.e., assuming the user wants to be as sure as possible that if the Pilots/ATCo contribute to the LoS the algorithm classifies it as "Yes") the level of sensitivity reached is ≈100% for Pilots, with ≈70% of accuracy, and ≈98% for ATCo, with ≈85% of accuracy; F1 score is ≈70% for Pilots and ≈91% for ATCos.
- when the Specificity is maximised (i.e., assuming the user wants to be as sure as possible that if the Pilots/ATCo are not responsible, the algorithm classifies it as "No") the level of specificity reached is ≈100% for Pilots, with ≈80% of accuracy, and ≈96% for ATCo, at the price of a low accuracy, ≈54%. F1 score is ≈60% for Pilots and ≈56% for ATCos.

Furthermore, Table VI shows that:

- when just predictions with confidence ≥75% are considered, the accuracy reaches ≈97% for Pilots contribution and ≈94% for ATCo. With this threshold, only 43% of the predictions are trusted when assessing Pilots contribution and 44% when considering the ATCo;
- when, instead, the accepted confidence level is decreased from 75% to 60%, the accuracy reaches ≈86% for both Pilots and ACTo contributions. With this new confidence level, 62% of observations are classified when assessing Pilots contribution and 70% when considering ATCo.

Finally, the ranking of the features (see Section IV-D) produced by the Random Forest algorithm is computed. This allowed the authors to observe that, based on the experience of the domain experts, the models learned correctly the importance of features related to the separation responsibility, such as the Flight type, the Flight rules, or the Airspace Class, other than the relevance of interactions between the ATC and the CWP, such as Radar Contact, ETO Over Fix or Action on Flight Level, in order to identify ATM contributions. These are promising results as the model presents room for improvement, such as the inclusion of more surveillance information or operational indicators such as traffic load.

VI. CONCLUSIONS

The objective of this work was to facilitate an automatic extraction of meaningful and actionable information from LoS reports and to investigate how the information recorded by the systems can help estimating contribution assessment. For this purpose, the authors proposed a threefold approach based on (i) an EDA, (ii) an automatic classification of extracted knowledge considering a state-of-the-art safety taxonomy (TOKAI), and (iii) an Automatic Contribution Assessment model based solely on the information recorded by the systems and available a few minutes after the ASMTs' identification of the LoS. The approach was tested on the LoSs reported in the CEANITA public database and the related ATC events.

For EDA purposes, unsupervised NLP techniques were applied aiming at identifying latent topics. In addition, this exploration was complemented with a clustering analysis, which facilitated the identification of similar incidents. Results demonstrated the capacity of these techniques to effectively identify meaningful topics and group together incidents.

For the automatic extraction of the safety factors and their classification according to the TOKAI taxonomy, the authors leveraged Syntactic Analysis. This is pioneering work in the field, and the results showed an understanding of the potential that these methods bring to safety analysis as well as a Resilience Engineering perspective. Indeed, the classification of actions according to the TOKAI taxonomy (TOKAI factors are neither negatively nor positively oriented) enables reframing of human behaviour not as a sequence of errors that lead to an undesired outcome (i.e., only pointing out where people went wrong), but as emergent from the system, arising as a function of complex interactions . The results of this classification were validated by demonstrating the strong connection between the factors identified and the main contributor to the incident.

Finally, the last step was the generation of an Automatic Contribution Assessment model, able to provide a prior indication whether the pilots, the ATCo or both have contributed to an incident. In this sense, different performance metrics were considered for evaluating the validity of the result. The results show that when only high-confidence predictions are considered, the model output reaches approximately 97% of accuracy for pilots' contribution and 94% for ATCo.

Future work could validate these techniques on other databases of reports (e.g., UKAB AirProx Board) and, more-

over, these techniques could be tailored to identify factors to be included in safety taxonomies or hidden sources of Resilient performance (e.g., when not fulfilling a procedure was opportune [53]), based on their presence on the reports, and could help facilitating the analysis pointed out in [5]. Finally, integrating other sources of structured data (e.g., about weather phenomena, STCA or TCAS activation, or traffic load) to develop richer models could lead to further insights in the estimation of contributors and precursors.

REFERENCES

[1] SESAR Joint Undertaking, "European ATM master plan - executive view, 2015 edition," https://www.sesarju.eu/node/2865, 2015.

[2] ——, "European ATM master plan - executive view, 2020 edition," https://op.europa.eu/en/publication-detail/-/publication/8afa1ad9-aac4-11ea-bb7a-01aa75ed71a1, 2020.

[3] Performance Review Comission, EUROCONTROL, "Performance review report," https://www.eurocontrol.int/sites/default/files/2020-06/eurocontrol-prr-2019.pdf, 2020.

[4] EASA, "Report of the EASA SKPI RP3 S(K)PI working group," https://ec.europa.eu/transport/sites/transport/files/easa_rp3_skpi_working_group_-_final_report.pdf, 2016.

[5] CANSO, "Incidents investigation toolbox," https://canso.fra1.digitaloceanspaces.com/uploads/2021/04/CANSO-Incidents-Investigation-Toolbox.pdf, 2021.

[6] N. Oza, J. P. Castle, and J. Stutz, "Classification of aeronautics system health and safety documents," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, vol. 39, no. 6, pp. 670–680, 2009.

[7] J. Switzer, L. Khan, and F. B. Muhaya, "Subjectivity classification and analysis of the ASRS corpus," in *IEEE International Conference on Information Reuse & Integration*, 2011.

[8] S. Wolfe, "Wordplay: an examination of semantic approaches to classify safety reports," in *AIAA Infotech@Aerospace*, 2007.

[9] I. Persing and N. Vincent, "Semi-supervised cause identification from aviation safety reports," in *Joint Conference of the Annual Meeting of the ACL*, 2009.

[10] R. Patriarca, R. Cioponea, G. Di Gravio, and A. Licu, "Managing Safety Data: the TOKAI Experience for the Air Navigation Service Providers," *Transportation Research Procedia*, vol. 35, pp. 148–157, 2018.

[11] R. Patriarca, G. Di Gravio, R. Cioponea, and A. Licu, "Safety intelligence: Incremental proactive risk management for holistic aviation safety performance," *Safety science*, vol. 118, pp. 551–567, 2019.

[12] S. Ananyan and M. Goodfellow, "Example application of PolyAnalyst with IATA STEADES data," https://flightsafety.org/wp-content/uploads/2016/09/polyanalyst_application.pdf, 2004.

[13] L. Tanguy, N. Tulechki, A. Urieli, E. Hermann, and C. Raynal, "Natural language processing for aviation safety reports: From classification to interactive analysis," *Computers in Industry*, vol. 78, pp. 80–95, 2016.

[14] K. D. Kuhn, "Using structural topic modeling to identify latent topics and trends in aviation incident reports," *Transportation Research Part C: Emerging Technologies*, vol. 87, pp. 105–122, 2018.

[15] W. J. Irwin, S. D. Robinson, and S. M. Belt, "Visualization of large-scale narrative data describing human error," *Human factors*, vol. 59, no. 4, pp. 520–534, 2017.

[16] S. D. Robinson, "Temporal topic modeling applied to aviation safety reports: A subject matter expert review," *Safety science*, vol. 116, pp. 275–286, 2019.

[17] O. Sjöblom, "Data mining in promoting aviation safety management," in *International Conference on Well-Being in the Information Society*, 2014.

[18] S. Ravizza, J. Chen, J. A. D. Atkin, P. Stewart, and E. K. Burke, "Aircraft taxi time prediction: comparisons and insights," *Applied Soft Computing*, vol. 14, pp. 397–406, 2014.

[19] H. Lee, W. Malik, and Y. C. Jung, "Taxi-out time prediction for departures at Charlotte airport using machine learning techniques," in *AIAA Aviation Technology, Integration, and Operations Conference*, 2016.

[20] S. Ayhan and H. Samet, "Aircraft trajectory prediction made easy with predictive analytics," in *ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016.

[21] C. E. Verdonk Gallego, V. F. Gómez Comendador, M. A. Amaro Carmona, R. M. Arnaldo Valdés, F. G. Sáez Nieto, and M. García Martínez, "A machine learning approach to air traffic interdependency modelling and its application to trajectory prediction," *Transportation Research Part C: Emerging Technologies*, vol. 107, pp. 356–386, 2019.

[23] C. E. Verdonk Gallego, V. F. Gómez Comendador, F. J. Saez Nieto, and M. García Martinez, "Discussion on density-based clustering methods applied for automated identification of airspace flows," in *IEEE/AIAA Digital Avionics Systems Conference*, 2018.

[22] M. Conde Rocha Murca, R. DeLaura, R. J. Hansman, R. Jordan, T. Reynolds, and H. Balakrishnan, "Trajectory clustering and classification for characterization of air traffic flows," in *AIAA Aviation Technology, Integration, and Operations Conference*, 2016.

[24] N. Takeichi, R. Kaida, A. Shimomura, and T. Yamauchi, "Prediction of delay due to air traffic control by machine learning," in *AIAA Modeling and Simulation Technologies Conference*, 2017.

[25] S. Choi, Y. J. Kim, S. Briceno, and D. Mavris, "Prediction of weather-induced airline delays based on machine learning algorithms," in *IEEE/AIAA Digital Avionics Systems Conference*, 2016.

[26] G. Di Gravio, M. Mancini, R. Patriarca, and F. Costantino, "Overall safety performance of Air Traffic Management system: Forecasting and monitoring," *Safety science*, vol. 72, pp. 351–362, 2015.

[27] Á. Rodríguez-Sanz, F. Gómez, J. M. C. García, and L. Meler, "Analysis of saturation at the airport-airspace integrated operations," in *USA/Europe Air Traffic Management Research and Development Seminar*, 2017.

[28] F. Bati and L. Withington, "Application of machine learning for aviation safety risk metric," in *IEEE/AIAA Digital Avionics Systems Conference*, 2019.

[29] Z. Nazeri, D. Barbara, K. De Jong, G. Donohue, and L. Sherry, "Contrast-set mining of aircraft accidents and incidents," in *Industrial Conference on Data Mining*, 2008.

[30] S. D. Robinson, W. J. Irwin, T. K. Kelly, and X. O. Wu, "Application of machine learning to mapping primary causal factors in self reported safety narratives," *Safety science*, vol. 75, pp. 118–129, 2015.

[31] E. Hollnagel and D. D. Woods, "Epilogue: Resilience engineering precepts," *Resilience engineering: Concepts and precepts*, pp. 347–358, 2006.

[32] E. Hollnagel, *Safety-I and safety-II: the past and future of safety management*. CRC press, 2018.

[33] D. M. Blei and J. D. Lafferty, "Topic models," *Text mining: classification, clustering, and applications*, vol. 10, no. 71, p. 34, 2009.

[34] B. S. Duran and P. L. Odell, *Cluster analysis: a survey*. Springer Science & Business Media, 2013.

[35] M. Straka and J. Straková, "Tokenizing, pos tagging, lemmatizing and parsing ud 2.0 with udpipe," in *CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, 2017.

[36] S. Shalev-Shwartz and S. Ben-David, *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[37] I. C. A. Organization, "Annex 11," http://skyrise.aero/wp-content/uploads/2017/03/ICAO-Annex-11-Air-traffic-services.pdf, 2001.

[38] S. Das, K. Dixon, X. Sun, A. Dutta, and M. Zupancich, "Trends in transportation research: Exploring content analysis in topics," *Transportation Research Record*, vol. 2614, no. 1, pp. 27–38, 2017.

[39] L. Sun and Y. Yin, "Discovering themes and trends in transportation research using topic modeling," *Transportation Research Part C: Emerging Technologies*, vol. 77, pp. 49–66, 2017.

[40] F. Murtagh and P. Contreras, "Algorithms for hierarchical clustering: an overview," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 2, no. 1, pp. 86–97, 2012.

[41] B. Mirkin, "Choosing the number of clusters," *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, vol. 1, no. 3, pp. 252–260, 2011.

[42] J. Shawe-Taylor and N. Cristianini, *Kernel methods for pattern analysis*. Cambridge university press, 2004.

[43] L. Breiman, "Random forests," *Machine learning*, vol. 45, pp. 5–32, 2001.

[44] M. Fernández-Delgado, E. Cernadas, S. Barro, and D. Amorim, "Do we need hundreds of classifiers to solve real world classification problems?" *The journal of machine learning research*, vol. 15, no. 1, pp. 3133–3181, 2014.

[45] M. Wainberg, B. Alipanahi, and B. J. Frey, "Are random forests truly the best classifiers?" *The Journal of Machine Learning Research*, vol. 17, no. 1, pp. 3837–3841, 2016.

[46] L. Oneto, *Model Selection and Error Estimation in a Nutshell*. Springer, 2020.

[47] I. Orlandi, L. Oneto, and D. Anguita, "Random forests model selection," in *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning (ESANN)*, 2016.

[48] H. White, "A reality check for data snooping," *Econometrica*, vol. 68, no. 5, pp. 1097–1126, 2000.

[49] I. Guyon and A. Elisseeff, "An introduction to variable and feature selection," *Journal of machine learning research*, vol. 3, no. Mar, pp. 1157–1182, 2003.

[50] R. Guidotti, A. Monreale, S. Ruggieri, F. Turini, F. Giannotti, and D. Pedreschi, "A survey of methods for explaining black box models," *ACM computing surveys*, vol. 51, no. 5, pp. 1–42, 2018.

[51] Y. Saeys, T. Abeel, and Y. Van de Peer, "Robust feature selection using ensemble feature selection techniques," in *Machine Learning and Knowledge Discovery in Databases*, 2008, pp. 313–325.

[52] R. Genuer, J. M. Poggi, and C. Tuleau-Malot, "Variable selection using random forests," *Pattern recognition letters*, vol. 31, no. 14, pp. 2225–2236, 2010.

[53] NTSB, "Aircraft accident report NTSB/AAR-19/03," https://www.ntsb.gov/investigations/AccidentReports/Reports/AAR1903.pdf, 2018.