UNCECOMP 2023
5th ECCOMAS Thematic Conference on
Uncertainty Quantification in
Computational Sciences and Engineering
M. Papadrakakis, V. Papadopoulos, G. Stefanou (eds.)
Athens, Greece, 12-14 June 2023

AI TOOLS FOR HUMAN RELIABILITY ANALYSIS

Karl Johnson[1], Caroline Morais[2], Edoardo Patelli[1*]

[1]University of Strathclyde
Department of Civil and Environmental Engineering Glasgow, UK
{karl.johnson, edoardo.patelli}@strath.ac.uk

[2]Agency for Petroleum, Natural Gas and Biofuels (ANP), Brazil
cmorais@anp.gov.br

Abstract

Understanding and quantify human performance is an essential component to guarantee and control the safety of critical installations where human intervention can represent the ultimate safety defence. Human reliability analysis is a time consuming and tedious task usually performed by a human factor expert and therefore subjected to error and variability. In addition, within human reliability analysis there are numerous opportunities to learn from data. However, how data are gathered, presented, shared, and used is an area of continuous development and discussion.

In this work, we present a collection of artificial intelligence (AI) tools and methodologies developed to tackle different challenges within the field of human reliability. The aim is to automatise the process, learn from data and support the task of human reliability experts. The collection of tools includes: a tool to automatically classify human errors from accident reports and construct a Bayesian/Credal Networks. The developed works are freely available as part of the open source COSSAN software.

Keywords: Human Reliability Analysis, Human Error, Machine Learning, Natural Language Processing, Bayesian Networks, Software

1. INTRODUCTION

Human reliability analysis is a field that is concerned primarily with the human contribution to risk and the factors that influence human performance, known as performance shaping factors (PSFs) [1]. Understanding the contribution and interactions of PSFs on human reliability allows to aid design processes, prevent accidents, and improve overall safety. How PSFs affect human reliability can be learn from observations from real data of accident reports, near missing or from lab experiments and simulators [2].

For instance, databases such as the Multi-Attribute Technological Accidents Dataset (MATA-D) can be used to support the opportunity for learning from past accidents [3]. MATA-D is a collection of major accident reports, classified with a framework focused on the

relationship between human error and PSFs. However, manually read all reports, extract relevant information, and create a model for human error is a challenging and complex task.

This work introduces a collection of data-driven tools and methodologies that have been developed to address various challenges in human reliability analysis. These tools tackle data-gathering, quality-of-life, time-consuming tasks, and data driven modelling challenges across the field. The toolkit includes a virtual human factor classificator named "Virtual Raphael" that allows the expansion of the MATA-D without the direct need of a human factor expert. This tool is trained on accident reports previously labelled by human experts and allows an automated classification of human errors and the corresponding PSFs involved, and details of the methodology presented in [1], [4]. The original classificator based on bag-of-words approach [5]has been improved using a popular transformer-based machine learning, known as BERT (Bidirectional Encoder Representations from Transformers) [6]. An example of one of the ways to make use of the available data is then presented, with the discussion of a tool developed to elicit causal links between the performance shaping factors.

The use of probabilistic tools, such as Bayesian/Credal Networks has become a popular choice for modelling human error [7]. However, data has only been used to estimate the conditional probability tables [8]. Therefore, a tool that combines elements of information theory and structure learning algorithms for Bayesian Networks has been developed, to produce a network structure that can be tailored to the user's requirements based on the initial inputs [9].

The above-mentioned techniques for human reliability analysis have been implemented as a standard-alone tools and available as part of the COSSAN software [10]- [11] and freely available.

2. HUMAN RELIABILITY ANALYSIS

Human Reliability Analysis is a systematic approach used to evaluate and improve the reliability and safety of systems by assessing the potential contribution of human error. The process usually involves identifying potential human errors, analysing the causes of those errors, factors that influence these, and determining the likelihood and consequences of those errors. This analysis can then be used to identify opportunities to improve system design, training, procedures, or other factors that could reduce the likelihood of human error. Human error probability is measure of the likelihood of a human error occurring in a particular task or activity, this can be defined as the ratio between the number of performed errors and the number of given opportunities for error to occur [12]. However, it is also important to consider the influence of performance shaping factors on this risk.

To carry out quantitative human reliability analysis and calculate human error probability data from a range of sources is often considered, including expert judgements, data from simulators, from real operation, near misses and accidents. Data collected from real operations (i.e., incidents, near-misses, and accidents) is considered the most credible data for estimating human error probabilities, above simulators and expert judgment [13]. Data collected from simulators is often restricted to human-machine interfaces in control rooms. Often collected data needs to be calibrated by expert judgment adopting well known approaches. This approach is strong on detecting human errors, but weak on detecting all the performance shaping factors. This is due to the decontextualization of the studied tasks [14]; for instance, operators know that their actions will not have any consequence and often know that their actions are being observed [13]. Expert elicitation can be considered the least credible source of data. Their judgements can be oriented by different sources of bias [15], and be systematically overconfident about the accuracy of their judgments [16]. However. all approaches to human

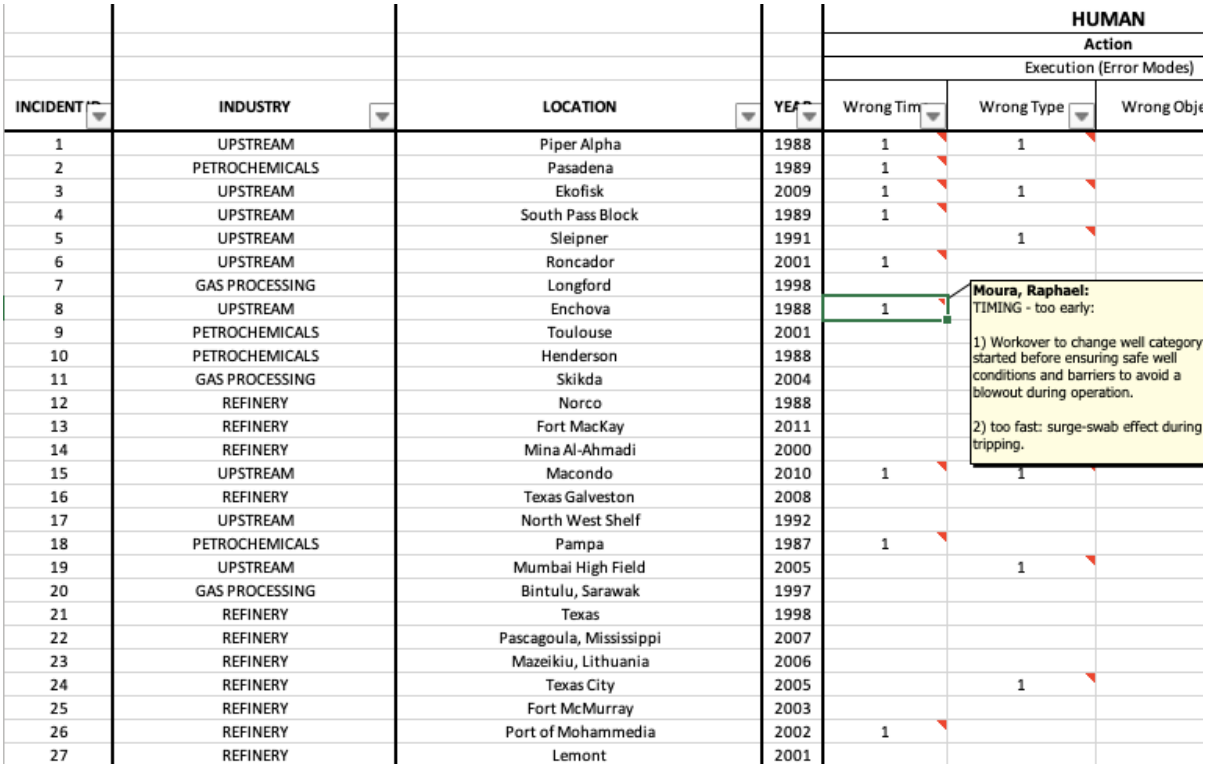
reliability analysis rely on expert judgment to some extent, as all methods start with a qualitative analysis of possible scenarios [17].

Data from major accident reports have the potential to better capture the interaction between human, machine, and organizational systems, providing additional contexts and scenarios not fully achieved by simulators and expert elicitation [3]. This is because detailed analysis of the causes that led to the accidents are required and performed [18]. Despite this the strategy of using major accident data to estimate performance shaping factors and human error probabilities had not been significantly explored, this led to the construction of the dataset used in this work, known as the Multi-Attribute Technological Accidents Dataset (MATA-D).

2.1 Multi-Attribute Technological Accidents Dataset

The Multi-Attribute Technological Accidents Dataset (MATA-D) [3] is a collection of 238 major accidents from a range of different industries considered to be of similar complexity, including aviation, chemical, oil & gas, nuclear, waste treatment etc., allowing the conceptual advantage of cross-learning from different industrial sectors. The accident reports for these incidents were then analysed by an expert focused on the contributing human factors, classified using the CREAM (Cognitive Reliability and Error Analysis Method) framework [19]. The CREAM taxonomy is comprised of human errors and performance shaping factors including organizational, technological, and individual factors [19].

This accident dataset provides researchers and practitioners with a simple interface and straighforward access to human factors data from accidents from a range of industrial sectors. The binary classification for the evaluated factors (presence or absence) allows data interpretation using uncomplicated statistical methods or sophisticated mathematical models, depending on the user's requirements. The MATA-D has been coded in Excel, and can be accessed and downloaded at https://datacat.liverpool.ac.uk/1018/ [3]. Each present factor also comes with a brief description, explaining the expert's decision to select these factors, as seen in Figure 1. As well as the CREAM classification and description, details regarding the location, industry and year of the accident are stored within the MATA-D.



				HUMAN		
				Action		
				Execution (Error Modes)		
INCIDENT	INDUSTRY	LOCATION	YEA	Wrong Tim	Wrong Type	Wrong Obje
1	UPSTREAM	Piper Alpha	1988	1	1	
2	PETROCHEMICALS	Pasadena	1989	1		
3	UPSTREAM	Ekofisk	2009	1	1	
4	UPSTREAM	South Pass Block	1989	1		
5	UPSTREAM	Sleipner	1991		1	
6	UPSTREAM	Roncador	2001	1		
7	GAS PROCESSING	Longford	1998			
8	UPSTREAM	Enchova	1988	1		
9	PETROCHEMICALS	Toulouse	2001			
10	PETROCHEMICALS	Henderson	1988			
11	GAS PROCESSING	Skikda	2004			
12	REFINERY	Norco	1988			
13	REFINERY	Fort MacKay	2011			
14	REFINERY	Mina Al-Ahmadi	2000			
15	UPSTREAM	Macondo	2010	1	1	
16	REFINERY	Texas Galveston	2008			
17	UPSTREAM	North West Shelf	1992			
18	PETROCHEMICALS	Pampa	1987	1		
19	UPSTREAM	Mumbai High Field	2005		1	
20	GAS PROCESSING	Bintulu, Sarawak	1997			
21	REFINERY	Texas	1998			
22	REFINERY	Pascagoula, Mississippi	2007			
23	REFINERY	Mazeikiu, Lithuania	2006			
24	REFINERY	Texas City	2005		1	
25	REFINERY	Fort McMurray	2003			
26	REFINERY	Port of Mohammedia	2002	1		
27	REFINERY	Lemont	2001			

Figure 1 Screenshot of the MATA-D showing the factors responsible for the accident and detail justification.

3. VIRTUAL RAPHAEL: A VIRTUAL HUMAN FACTORS CLASSIFIER

The collection of empirical data is a time-consuming and expensive, especially in the field of human reliability, where data collection and classification are usually done by experts in their fields. For example, the MATA-D database was constructed through the extensive reading and manual classifying of 238 accident investigation reports, a task that took approximately one year to be completed. Alongside the time-consuming nature of this task, a specialised knowledge is required, as the assessors had to be minimally trained on the taxonomy used to pursue the classification. The expansion of the dataset is necessary not only to reduce epistemic uncertainty, by collecting and classifying more accident reports to increase the chance of describing more human-machine-organization interactions, but also to account for changes in human behaviour and technological advancements over time.

3.1. SVM Approach

The presented work shares the development of an automated approach to analyse new accident reports and identify the present human factors, trained using the classified reports that make up the MATA-D. The proposed methodologies have been applied to analyse the preliminary accident report of the 2018 Lion Air accident showing the feasibility of the machine-learning proposed approach [4].

Shown in Figure 2 is a simplified workflow of the machine learning based classification approach while details of the algorithm and its performance are available in Ref. [1]. In the first module, the accident investigation reports are processed to check if the text in the files is recognized by the machine and, if not, an optical character recognition software (OCR) is used to convert them to text files. After this pre-treatment, the text is pre-processed to clean punctuation, stop words, and reducing words to their stem (e.g., ‘testing’ is reduced to ‘test’). The tool takes each accident report’s file name and finds the corresponding entry in the MATA-

D. This gives the machine-learning component the desired input for each accident report, a combination of the accident report text and their known human factors features. Then, the selected text is converted into bag-of-words objects, that is the vocabulary of words and their frequency of occurrence from the text, forming the input for the model together with the features extracted from the MATA-D which serves as the output of the model. The third module of the approach is based on Support Vector Machine (SVM) [20]. The SVM is trained and tested using data processed by the previous two modules, and calculate the performance metrics such as accuracy, precision, recall and F1.

The fourth and final module of the tool is concerned with the application of the model learnt through the previous modules. Users are able to add a new report that is not yet part of the MATA-D. This report then goes through a similar series of steps as set out before and tested using the learnt SVM model. Finally, a list of the human reliability features identified by the tool (an array of the predicted positive factors), a small table with all positives and negatives predictions (the 53 features of the chosen taxonomy), and a word cloud of the most relevant words in the report, is produced.

The classifier has been coded in Matlab and supported by the *text analytics toolbox*, which uses the bag-of-words model to extract text strings from files and prepare data for the machine-learning algorithm. The *MATLAB statistics and the machine-learning toolbox* is used to transform text inputs into binary classification adopting the Support Vector Machine [1]. The Virtual Raphael classificator runs the entire process in approximately one minute on a standard laptop computer, (which may be further reduced by saving and storing the SVM model, rather than retraining it each time a new report is tested), compared with the approximately three days it takes to manually read and classify one 200-page accident report.

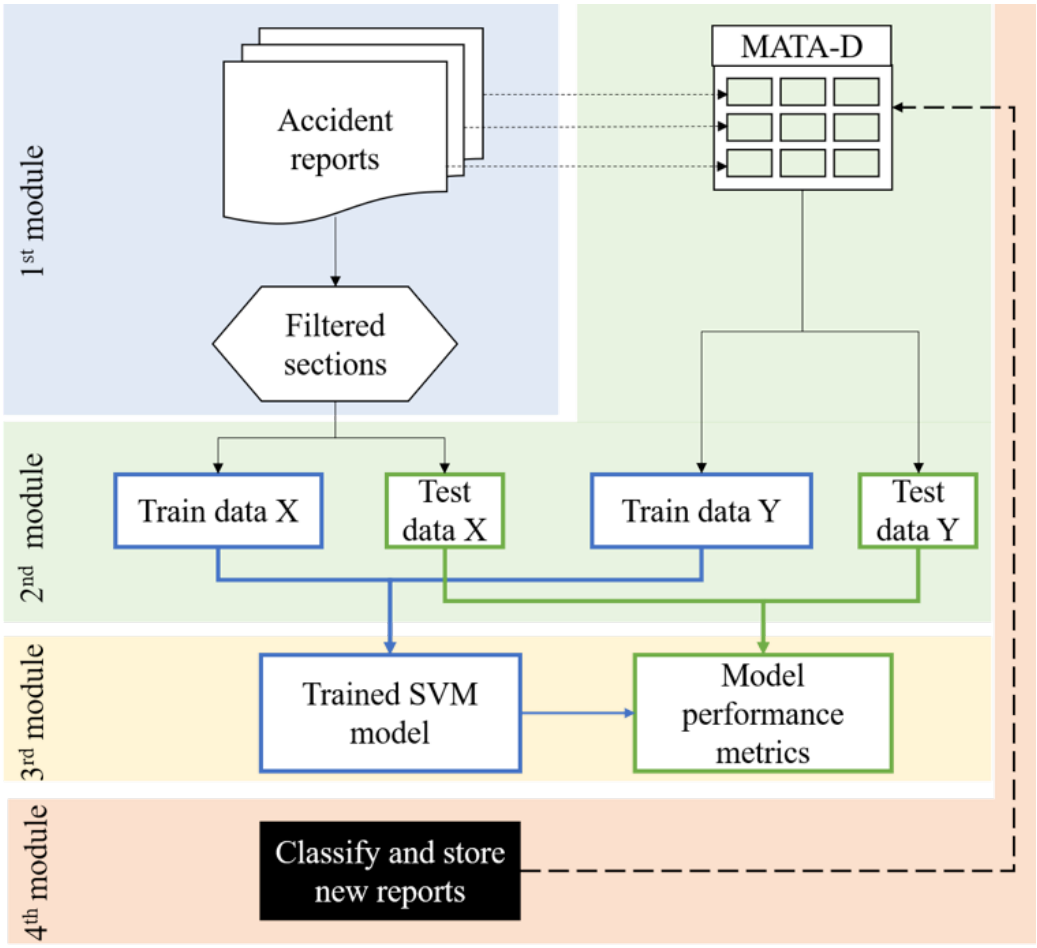


Figure 2 Workflow of the "Virtual Raphael" Human Factors Virtual Classifier.

Error! Reference source not found. shows the average performance metrics for the virtual classifier based on SVM approach on a 10% test set of the available reports excluded from the training set. The results obtained are similar to other studies and demonstrate there is value in this approach for aiding experts in the evaluation of accident reports.

Metric	
Accuracy	86%
Precision	60%
Recall	46%
F1 score	52%

Table 1 Performance Metrics Human Factors Virtual Classifier SVM Approach, adapted from [1].

3.2 BERT Approach

Although the developed tool based on SVM (see Section 3.1) performs sufficiently well and it is quite robust, an improved version has been proposed for making use of more recent progression in the field of Natural Language Processing (NLP). The tool has the same aim and applicability of the original Virtual Human Factors Classifier but with an improvement in the performance metrics.

This proposed approach makes use of BERT (Bidirectional Encoder Representations from Transformers), which is a pre-trained language model that is designed to understand natural human written or spoken language [6]. BERT has proven to be highly effective when used for text classification tasks. BERT makes use of a Transformer, an attention mechanism, that learns contextual relations between words in a text. Transformers include two separate mechanisms, an encoder that reads the text input and a decoder that produces a prediction for the task [6]. As opposed to directional models, which read the text input sequentially (left-to-right or right-to-left), the transformer encoder reads the entire sequence of words at once. Therefore, it is considered bidirectional, this allows the model to learn the context of a word based on all of its surroundings. By better considering the context of a word than other approaches, BERT can lead to a more accurate classification of the text. Alongside this, BERT is pre-trained on a large corpus of text data, that allows the model to better learn general structures of language. This transfer learning approach can therefore reduce the amount of pre-labelled data required for training a model, which is one of the major advantages for this proposed approach due to the limited availability of pre-labelled accident reports in the MATA-D and in other databases relevant for human reliability analysis.

Similarly, to the original Virtual Human Factors Classifier, the accident investigation reports are processed to check if the text in the files is recognised by the machine and, if not, an optical character recognition software (OCR) is used to convert them to text files. To prepare the text data to be used by the model, it must first be tokenized using the same scheme used to pretrain the BERT model, which is called WordPiece tokenization [21]. This works by splitting words either into the full forms (e.g., one word becomes one token) or into word pieces — where one word can be broken into multiple tokens. Using word pieces allows BERT to easily identify related words as they will usually share some of the same input tokens, which are then fed into the first layers of BERT. This step is completed using the pre-trained BERT tokenizer from Hugging Face Transformers library on Python. The text data is then combined with the labels from MATA-D [21]. This is used to train the classification layer added on top of the pre-trained BERT model imported from the same library. Which is then optimized using stochastic gradient descent, with binary cross-entropy loss function, where the hyperparameters were fine-tuned based on performance on the validation set. The model is then saved, so that new accident reports, which must first be processed through the same tokenizer, can be classified, with the final tool outputting the identified factors, and a binary array corresponding to all 53 factors that can be used to add the incident to the MATA-D.

Significant improvements in the performance metrics have been obtained and reported in Table 2 that shows the average of the performance metrics based on a 10% test set of the available reports excluded from the training set. However, the increased performance come at the prize of a greater complexity that reduces usability of this tool when compared with the original based on SVM. This tool will allow the expansion of the MATA-D at a faster rate (approximately one minute per report) whilst maintaining a performance and accuracy, that can be considered more in line with the performance (in terms of classification) from a human expert.

Metric	
Accuracy	91%
Precision	88%
Recall	77%
F1 score	82%

Table 2 Performance Metrics Virtual Human Factors Classifier BERT Approach.

3.3 Validation

In order to demonstrate and test the application of the above tools, a new accident report (not yet part of MATA-D) from an incident in the oil & gas industry was selected to be analysed and classified by the same experts that classified the accident reports for MATA-D assuring consistency in the classification (avoiding the natural and unavoidable variability across experts). The results of the automated classification were not shown to the human expert before the task, to avoid biased [1].

The case study selected is the explosion onboard of floating production, storage and offloading unit (FPSO) Cidade de São Mateus in Brazil occurred in February 2015. The FPSO was operated by BW Offshore in gas fields under concession to Petróleo Brasileiro S.A (Petrobras) in Brazilian waters [57]. The accident killed nine workers and injured 26, as well as caused damage to the installation, and the production was halted at two gas production fields up to 2020. The Brazilian Oil & Gas regulator (ANP) has included in their investigation report root causes from the design phase to the emergency response.

Table 3 shows the classification of human factors obtained by the human experts and by the "Virtual Raphael" tool with the SVM and BERT version, respectively. Based on the BERT classification tool, the following features are observed in the oil & gas installation incident of the FPSO Cidade de Sao Mateus: human errors of execution of wrong place, of interpretation with faulty diagnosis and wrong reasoning, and permanent person related functions due to cognitive bias; the technological factors of incomplete information (related to temporary interfaces); the organisational factors of communication failure, missing information, maintenance failure, inadequate quality control, design failure, inadequate task allocation, insufficient skills and knowledge, and excessive demand related to working conditions. As well as two human errors due to Temporary Person Related Functions, memory failure and inattention, not identified by the expert.

Table 4 shows the performance of "Virtual Raphael" for the validation case. Both approaches demonstrate a high precision and good accuracy scores, however there is a significant improvement in recall and therefore F1 Score from the SVM to the BERT. Recall can be considered the most important performance metric for human reliability classifier. The cost of a human error or a performance shaping factor that goes undetected (and consequently with no resource allocated to its risk reduction) seems to be more severe than spotting errors that are non-existent. On the other hand, when considering resource allocation, precision is also of high importance, as this increases the chance that the factors predicted are mostly correct and thus investment is not wasted. F1 Score symmetrically represents both precision and recall in one metric given by the formula F1=2*Recall*Precision/Recall+Precision. The significant improvement in the F1 Score seen in the average performance metrics (Table 1, Table 2) and here in this validation task, justifies the development of the BERT method tool.

				Expert	SVM	BERT
HUMAN	**Action**	Execution (Error Modes)	Wrong Time	0	0	0
			Wrong Type	0	0	0
			Wrong Object	0	0	0
			Wrong Place	**1**	0	1
	Specific Cognitive Functions	Observation	Observation Missed	**1**	0	0
			False Observation	0	0	0
			Wrong Identification	0	0	0
		Interpretation	Faulty diagnosis	**1**	0	1
			Wrong reasoning	**1**	0	1
			Decision error	0	0	0
			Delayed interpretation	0	0	0
			Incorrect prediction	0	0	0
		Planning	Inadequate plan	**1**	0	0
			Priority error	0	0	0
	Temporary Person Related Functions		Memory failure	0	0	1
			Fear	0	0	0
			Distraction	0	0	0
			Fatigue	0	0	0
			Performance Variability	0	0	0
			Inattention	0	0	1
			Physiological stress	0	0	0
			Psychological stress	0	0	0
	Permanent Person Related Functions		Functional impairment	0	0	0
			Cognitive style	0	0	0
			Cognitive bias	1	0	1
TECHNOLOGY	**Equipment**		Equipment failure	0	0	0
			Software fault	0	0	0
	Procedures		Inadequate procedure	1	1	0
	Temporary Interface		Access limitations	0	0	0
			Ambiguous information	0	0	0
			Incomplete information	1	0	1
	Permanent Interface		Access problems	0	0	0
			Mislabelling	0	0	0
ORGANISATION	**Communication**		Communication failure	1	0	1
			Missing information	1	0	1
	Organisation		Maintenance failure	1	1	1
			Inadequate quality control	1	1	1
			Management problem	0	0	0
			Design failure	1	1	1
			Inadequate task allocation	1	1	1
			Social pressure	1	0	0
	Training		Insufficient skills	1	0	1
			Insufficient knowledge	1	0	1
	Ambient Conditions		Temperature	0	0	0
			Sound	0	0	0
			Humidity	0	0	0
			Illumination	0	0	0
			Other	0	0	0
			Adverse ambient conditions	0	0	0
	Working Conditions		Excessive demand	1	0	1
			Inadequate workplace layout	0	0	0
			Inadequate team support	0	0	0
			Irregular working hours	0	0	0

Table 3 Classification of accident report by human expert and Virtual Raphael.

Performance	SVN	BERT
Sum of true negatives	35	33
Sum of false positives	0	2
Sum of false negatives	13	4
Accuracy	75%	89%
Precision	100%	88%
Recall	28%	78%
F1 Score	43%	83%

Table 4 Performance of the Virtual Raphael for the validation case study.

4. VIRTUAL RAPHAEL: IDENTIFYING CAUSAL LINKS - HUMAN ERRORS & PSFs

Various human reliability analysis methods have been developed to aid in the incorporation of the human contribution to risk into overall system safety analysis. In particular, performance shaping factors are factors which may have positive or negative influence on human performance, and these include organizational, technological, and personal factors [1]. The identification and quantification of the effects of these factors is a key step in the process of various human reliability analysis approaches [22]. Such factors are interdependent, therefore there is also the need to consider the interrelationships among these factors [8]. A popular choice for producing these models has been with the use of Bayesian and Credal networks [7]. Their graphical structure allows the causal links between PSFs and events to be easily recognized by those not involved directly in the model building [7]. The level of influence that these factors have on each other can be stored within the conditional probability tables included in the model [23].

Various models have been proposed where analysts have been able to combine information from several sources, including empirical data and expert opinion. However, there is a reliance on expert opinion to identify the causal links between the PSFs, and thus determine the structure of such models. Due to this there may be some potential links between factors not considered due to various types of bias [8]. Therefore, in this section a tool that can be used to elicit the causal links between PSFs from data is presented. The tool proposed allows the integration of expert opinion into the model through the use of structure learning algorithms for Bayesian networks. These causal links give the structure of Bayesian network for human reliability analysis.

4.1. Implementation details

The flowchart of the tool for identifying the structure of a Bayesian or Credal network is presented in Figure *3*. The tool is based on 4 steps. In the first step, the user needs to select the most appropriate options for the models. In particular, the user needs to determine whether to use the ungrouped or grouped PSFs, i.e. the 53 factors originally discussed in MATA-D, or a grouped version of these factors into 15 categories, respectively. For example, they may have identified a set of factors at play in the operation space and are now looking to evaluate how these factors impact human performance.

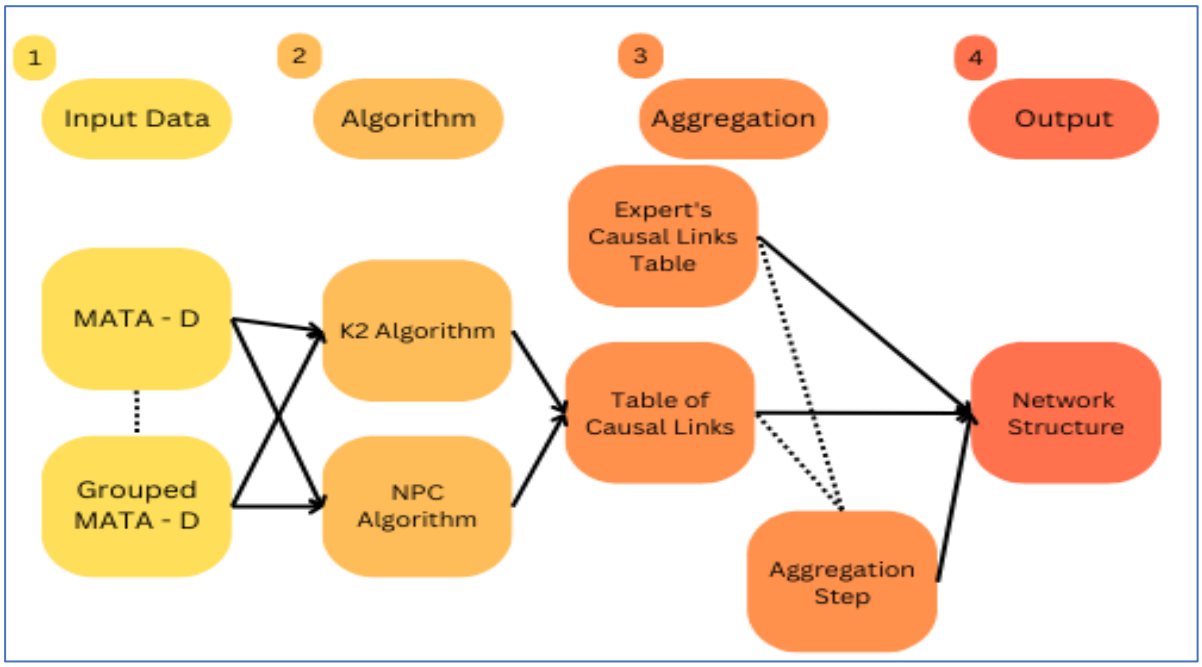


Figure 3 Human reliability analysis Structure Learning Tool Steps

In step 2, the user is presented with the option of selecting the most appropriate methodology for learning the structure of a direct acyclic graph: these are the K2 algorithm and NPC algorithm [24]. K2 and NPC are popular example of structure learning algorithms for Bayesian Networks (score-based and constraint-based respectively). Depending on the selection some further inputs may be required, for example, max number of parents, confidence level, etc.

In step 3 the aggregation approach is selected. This can be based on "Expert" or "Aggregated". The Expert option is based upon a table presented in Hollnagel's original work regarding the CREAM human reliability analysis method [19]. The "Aggregated" option combines the information learnt from the K2 and NPC structure learning approaches, where the arcs appearing in both learnt networks can be considered to be part of the core structure of the network [25]. The aggregation method triggers the use of both the learning algorithms. If the expert option is also than the network is expanded through the inclusion of the other arcs that are present in the expert's table, and in one of the two learnt structures.

Step 4 represents the output of the tool. The output is presented in a table format where the presence of a one denotes a causal link between the corresponding factors. The tool also provides a visual representation of this table, as would be usually presented in a Bayesian Network model.

The main application of this tool is to support users in their construction of causal models, by providing a suggested network structure that can be directly input into a Bayesian Network model and used to inform and provide evidence in the user's choices. A benefit of using this methodology, is that causal relationships between factors that experts have not previously considered may be identified by the structure learning algorithms. Such relationships may have been previously overlooked due to underlying bias or there may be situations that would benefit from further research and data gathering efforts. As well as aiding in the construction of causal models, the user can generate multiple different networks selecting different sets of PSFs. This could be used to highlight which parent factors need to be focused on to reduce specific risks or other factors that affect human performance. This would allow management to make more informed decisions on where to invest resources to mitigate human risk [9].

4.2 Algorithm Testing

In order to validate the tool, the algorithms behind it were tested on various sized datasets. To achieve this, a random network as shown in Figure 4 was created. Then, datasets were generated using the process presented by Oehm in 2015 [26]. Three sample datasets were created with 50, 150 and 1500 entries, respectively. The proposed algorithm is then used to generate network using the NPC algorithm and K2 algorithms form the artificial dataset created.

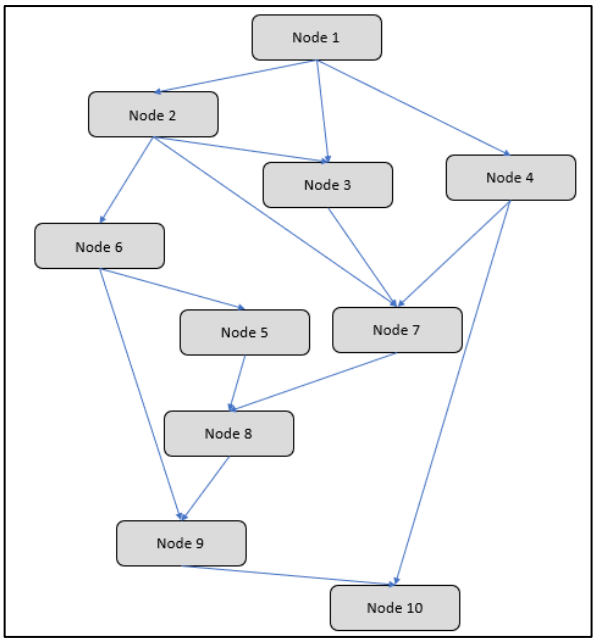


Figure 4 Test Network Structure

To evaluate the performance, the generated networks were then compared with the Test Network of *Figure 4* used to generate the dataset. If an arc exists between a parent and child node is a binary problem, where we have 1 if a node exists from parent node A to child node B, and 0 if ones does not. Therefore, using the following,

- True positives (TP) occur when the true value is 1 and the model correctly predicts 1
- False negatives (FN) occur if the true value is 1 but the model wrongly predicts 0
- True negatives (TN) occur when true value is 0 and the model correctly predicts 0
- False positives (FP) occur when true value should be 0 but the model predicts 1.

The resulting metrics *accuracy, precision, recall and F_1 Score*, are presented in *Table 5*.

	Metrics	50 samples	200 samples	2000 samples
NPC Algorithm	Accuracy	73%	86%	91%
	Precision	7%	52%	55%
	Recall	14%	72%	77%
	F_1 Score	9%	60%	64%
K2 Algorithm	Accuracy	77%	83%	87%
	Precision	14%	52%	55%
	Recall	15%	72%	74%
	F_1 Score	15%	60%	63%

Table 5 K2 and NPC Performance Metrics for Test Network

The test network demonstrated the capacity of the algorithm to identify the most appropriate links in a network that explain the dataset. The quality of the network depends on the number of sample available but with good result also with a limited number of samples. The 200 samples results are the most significant as this is approximately how many samples are in MATA-D currently being used by the tool. The algorithms performed better in the recall metric than precision, suggesting the algorithms are more likely to produce a false positive than a false negative, which could then be pruned out using expert judgement in practice. It is expected that with further expansion of the MATA-D, aided by the virtual human classifier, the increase in available data will improve the overall tools performance, as is suggested by these test results.

4.3 Causal Links Tool Case Study

To demonstrate an application of this tool, the tools is applied to identify the potential structure of a Bayesian network model able to explain the links between the human errors and PSFs for the FPSO Cidade de São Mateus accident analysed in Section 3.3.

The analysis starts by selecting the tools options of the ungrouped set of factors (i.e. *Wrong Place, Wrong reasoning, Memory failure, Inattention, Cognitive bias, Incomplete information, Communication failure, Missing information, Maintenance failure, Inadequate quality control, Design failure, Inadequate task allocation, Insufficient skills, Insufficient knowledge, Excessive demand*). Then, the learning algorithm and aggregation technique are selected. In this example, the "Aggregated" option is selected together with a cap of four parent nodes for the K2 algorithm, and a 5% significance level for the NPC algorithm. The produced network structure is shown in *Figure 5*.

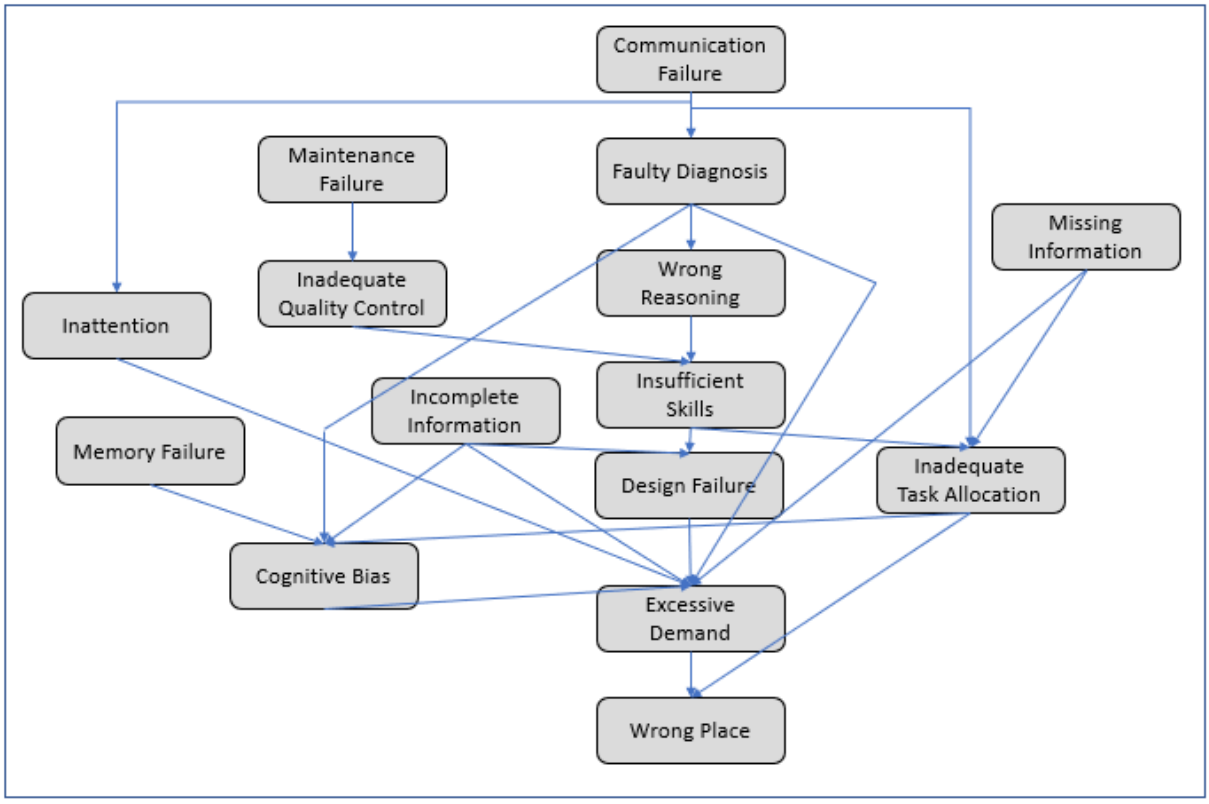


Figure 5 Causal Links obtained for the FPSO Cidade de São Mateus accident using the aggregation method.

This network structure suggests that there were several of the elicited factors that led to *excessive demand* on the operator, which in turn led to the human error action of '*wrong place'*. There is still some user interpretation when considering the produced network. For example, there is a suggested link from '*Insufficient Skill*s' to '*Design Failure'* to '*Excessive Demand'*, that could cause some confusion. It is important to understand that '*Design Failure'* is not an action in the incident, but instead an organizational factor. The operator's *insufficient skill* highlighted in this example that led to an '*Excessive Demand'* is due to a organizational factor. This network structure could then be combined with conditional probability tables calculated from available data, whether this be from the MATA-D or other sources (other accident datasets, simulator data, expert judgement), to produce a complete Bayesian network, that can be used to calculate the HEP and quantify the influence between factors.

6. CONCLUSIONS

This work demonstrates through the use of machine learning tools, an efficient process to gather human reliability data from accident reports, and how to model the causal links between the identified performance shaping factors. The developed "Virtual Raphael" allows to automatise the expert dependent multiple day process of evaluating accident reports for human errors and PSFs, and complete the task within a few minutes, with good to great performance. These tools should therefore aid in increasing the rate of expansion of the MATA-D, by reducing the time taken and the need for a continuous supervision of the classification task by an experienced expert and allowing an increasing of the amount of highly valuable real operation data on human error and PSFs. This in turn feeds back into the causal links/structure learning tool, by providing more data, increasing the opportunities for learning for past accidents and the trust in the models produced. It is also addressing the fields desire to move more towards data driven/support models and choices.

The current code/tools are available on https://github.com/cossan-working-group, and freely available as part of the COSSAN software [10], [11], [27]. Alongside this, a web interface to access MATA-D and the AI tools for human reliability analysis is current under development making these tools directly available to analysists and practicians.

8. ACKNOWLEDGMENTS

This work was partially supported by the EPSRC grant EP/T517938/1. All data underpinning this publication are openly available from **https://datacat.liverpool.ac.uk/1018/**

REFRENCES

[1] C. Morais, K. Yung, K. Johnson, R. Moura, M. Beer and E. Patelli, "Identification of human errors and influencing factors: A machine learning approach," vol. 146, p. 105528, 2022.

[2] C. Morais, S. Ferson, R. Moura, S. Tolo, M. Beer and E. Patelli, "Handling the Uncertainty with Confidence in Human Reliability Analysis'," in *31st European Safety and Reliability Conference*, Anger, France, 2021.

[3] R. Moura, M. Beer, E. Patelli, J. Lewis and F. Knoll, "Learning from major accidents to improve system design," *Safety Science,* vol. 84, pp. 37-45, 2016.

[4] C. Morais, K. L. Yung and E. Patelli, "Machine-Learning Tool for Human Factors Evaluation – Application to Lion Air Boeing 737-8 Max Accident'.," in

UNCECOMP 2019. 3rd ECCOMAS Thematic Conference on International Conference on Uncertainty Quantification, Crete, 2019.

[5] W. Qader, M. Ameen and B. Ahmed, "An Overview of Bag of Words;Importance, Implementation, Applications, and Challenges," in *2019 International Engineering Conference (IEC)*, Erbil, Iraq, 2019.

[6] R. Horev, "BERT Explained: State of the art language model for NLP," 2018. [Online]. Available: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270.

[7] C. Morais, H. Estrada-Lugo, S. Tolo, T. Jacques, R. Moura, M. Beer and P. E, "Robust Data-Driven Human Reliability Analysis Using Credal Networks," *Reliability Engineering & System Safety,* vol. 218, 2022.

[8] K. M. Groth and A. Mosleh, "Derivingcausal Bayesian networks from humanreliability analysis data: A methodologyand example model," *Journal of Risk and Reliability,* vol. 226, no. 4, pp. 361-379, 2012.

[9] K. Johnson, C. Morais, L. Walls and E. Patelli, "A data driven approach to elicit causal links between performance shaping factors andhuman failure events," in *32nd European Safety and Reliability Conference*, Dublin, Ireland, 2022.

[10] E. Patelli, "COSSAN: A Multidisciplinary Software Suite for Uncertainty Quantification and Risk Management," *Handbook of Uncertainty Quantification,* 2016.

[11] E. Patelli, M. Broggi, M. d. Angelis and M. Beer, "OpenCossan: An Efficient Open Tool for Dealing with Epistemic and Aleatory Uncertainties'. In Vulnerability, Uncertainty, and Risk," *American Society of Civil Engineers,* p. 2564–73, 2014.

[12] F. Khan, P. Amyotte and D. DiMattia, "HEPI: A new tool for human error probability calculation for offshore operation," *Safety Science ,* vol. 44, no. 4, 2006.

[13] B. Kirwin, A Guide To Practical Human Reliability Assessment, 1994.

[14] D. Gertman, H. Blackman, J. Marble, C. Smith, R. Boring and P. O'Reilly, "THE SPAR-H HUMAN RELIABILITY ANALYSIS METHOD," in *Fourth American Nuclear Society International Topical Meeting on Nuclear Plant Instrumentation, Controls and Human-Machine Interface Technologies (NPIC&HMIT)*, Columbus, Ohio, 2004.

[15] A. Mosleh, V. Bier and G. Apostolakis, "A Critique of Current Practice for the Use of Expert Opinions in Probabilistic Risk Assessment," *Reliability Engineering and System Safety ,* vol. 20, pp. 63-865, 1988.

[16] S.-W. Lin and V. Bier, "A study of expert overconfidence," *Reliability Engineering & System Safety ,* vol. 93, no. 5, pp. 711-721 , 2008.

[17] K. Laumann, M. Rasmussen and R. L. Boring, "A literature study to explore empirically: what is the scientific discipline of human factors and what makes it distinct from other related fields," in *Advances in Human Error, Reliability, Resilience, and Performance*, 2018.

[18] T. Hywel, "ACCIDENT AND INCIDENT ROOT CAUSE ANALYSIS," 2018.

[19] E. Hollnagel, Cognitive Reliability and Error Analysis Method (CREAM), Oxford: Elsevier Science Ltd, 1998.

[20] T. Joachims, *Text Categorization with Support Vector Machines: Learning with Many Relevant Features,* Dortmund, Germany, 2005.

[21] C. Khanna, "WordPiece: Subword-based tokenization algorithm," 2021. [Online]. Available: https://towardsdatascience.com/wordpiece-subword-based-tokenization-algorithm-1fbd14394ed7.

[22] C. D. Griffith and S. Mahadevan, "Inclusion of fatigue effects in human reliability analysis," *Reliability Engineering & System Safety,* vol. 96, no. 11, pp. 1437-1447, 2011.

[23] C. Morais, R. Moura, M. Beer and E. Patelli, "Analysis and Estimation of Human Errors From Major Accident Investigation Reports," *ASME J. Risk Uncertainty Part B. ,* vol. 6, no. 1, pp. 011014-011030, 2020.

[24] M. Scutari, C. E. Graafland and J. M. Gutierrez, "Who Learns Better Bayesian Network Structures: Accuracy and Speed of Structure Learning Algorithms," *International Journal of Approximate Reasoning,* vol. 115, pp. 235-253, 2019.

[25] S. Acid, "A comparison of learning algorithms for Bayesian networks: A case study based on data from an emergency medical service.," *S.Artificial Intelligence in Medecine,* vol. 30, no. 3, pp. 215-232, 2004.

[26] D. Oehm, "Simulating data with Bayesian networks," 2020. [Online]. Available: https://www.r-bloggers.com/2020/02/simulating-data-with-bayesian-networks-by-daniel-oehm/.

[27] E. Patelli, H. George-Williams, J. Sadeghi, R. Rocchetta, M. Broggi and M. de Angelis, "OpenCossan 2.0: An Efficient Computational Toolbox for Risk," in *Joint ICVRAM ISUMA UNCERTAINTIES Conference*, 2018.

[28] M. Galarnyk, "Visualizing Decision Trees with Python (Scikit-learn, Graphviz, Matplotlib)," 2020. [Online]. Available: https://towardsdatascience.com/visualizing-decision-trees-with-python-scikit-learn-graphviz-matplotlib-1c50b4aa68dc.

[29] R. Horev, "BERT Explained: State of the art language model for NLP," 2018. [Online]. Available: https://towardsdatascience.com/bert-explained-state-of-the-art-language-model-for-nlp-f8b21a9b6270. [Accessed March 2023].

[30] E. Patelli, "COSSAN: A Multidisciplinary Software Suite for Uncertainty Quantification and Risk Management," in *Handbook of Uncertainty Quantification*, Cham, Springer International Publishing, 2016, pp. 1-59.

[31] H. Zerrouki, H. D. Estrada-Lugo and E. Patelli, "Workplace Accident Analysis in the Algerian Oil and Gas Industry," *Process Safety Progress,* 2023.