ALMA MATER STUDIORUM
UNIVERSITÀ DI BOLOGNA

in cotutela con Norwegian University of Science and Technology

DOTTORATO DI RICERCA IN

INGEGNERIA CHIMICA E DI PROCESSO

Ciclo 36

Settore Concorsuale: 09/D3 – Impianti e processi industriali chimici

Settore Scientifico Disciplinare: ING-IND/25 – Impianti chimici

Development of Data-Driven Methods for Dynamic Risk Management in the Chemical Industry

Presentata da: Nicola Tamascelli

Coordinatore Dottorato

Prof. Alessandro Tugnoli

Supervisore

Prof. Valerio Cozzani

Co-Supervisore

Prof. Nicola Paltrinieri

Esame finale anno 2024

Preface

This thesis is submitted to the University of Bologna (UNIBO) for partial fulfillment of the requirements for the degree of Philosophiae Doctor.

This Ph.D. study was carried out in the framework of a cotutelle agreement between UNIBO and the Norwegian University of Science and Technology (NTNU). The work was carried out at the Department of Mechanical and Industrial Engineering at NTNU, in Trondheim, Norway, and at the department of Civil, Chemical, Environmental, and Material Engineering at UNIBO, in Bologna, Italy. Professor Nicola Paltrinieri from NTNU and Professor Valerio Cozzani from UNIBO co-supervised the activities. NTNU and UNIBO jointly funded the doctoral work.

The target audience of this work includes researchers and practitioners interested in the following areas: Risk Assessment and Management, Dynamic Risk Management, Machine Learning methods in chemical and process safety, evaluation and monitoring of safety barriers, predictive and diagnostic methods to improve the performance of industrial alarm systems, consequence prediction of accidents involving dangerous substances, frequency evaluation in domino scenarios.

Summary

Large amounts of hazardous substances are handled and stored in chemical facilities, elevating the risk of accidental releases with potentially disastrous consequences. Over the past three decades, there has been a significant evolution in the domain of safety science, leading to the standardization and widespread implementation of Risk Managment (RM) frameworks designed to identify, quantify, evaluate, control, and manage the risk associated with industrial activities involving hazardous substances. However, canonical RM techniques suffer several limitations, such as their inherent staticity and inability to update the risk picture in evolving and degrading systems. To overcome these limitations, recent research has proposed to move toward a more dynamic and proactive approach to process safety, named Dynamic Risk Management (DRM), which aims at capturing risk variations in industrial facilities, taking into account the performance of the control system, safety barriers, inspection and maintenance activities, and the human factor. This paradigm shift raises the need for dynamic and inherently updatable tools to capture the intricate dynamics between risk-influencing factors. In this context, Machine Learning (ML) techniques emerge as valuable tools due to their inherent ability to make predictions under uncertainty and model complex nonlinear relationships between features. However, the potential of these techniques in the context of DRM is still scarcely explored. Therefore, this Ph.D. study seeks to contribute to the development of ML methods to enhance and support DRM.

Specifically, this investigation formulates and presents practical ML-based methods to address critical tasks in DRM, namely consequence prediction, frequency evaluation, and monitoring of safety barriers. In addition, this study delves deep into the broader implications of adopting ML technologies, such as the intricate relationship between human expertise and AI, critically examining their respective contributions in the future of Risk Management. Different ML algorithms have been explored, including classification, clustering, regression, and Natural Language Processing. Diverse data sources have been utilized, such as alarm data, process data, and accident data. The contributions of this research include:

- the assessment of recent advancements of ML in the domain of safety and reliability;
- the development of classification models to predict the consequences of major accidents;
- the investigation of ML models to aid Risk Based Inspection of hydrogen systems;
- the use of regression models to predict the Time-To-Failure of tanks exposed to external fire;
- the exploration of classification models and natural language processing algorithms to monitor and improve the performance of industrial alarm systems;
- the integration between traditional risk assessment tools, data-driven models, and resilience analysis to evaluate safety barriers in environmental-critical facilities;
- the analysis of the involvement between ML and human actors in Risk Management.

Proposed methods have been tested on real-world case studies to demonstrate their efficacy. The results indicate that ML methods can be used to take advantage of the wealth of heterogeneous data made available by the widespread digitalization of industrial sectors in order to extract safety-relevant knowledge and provide critical support to DRM. Limitations and challenges have been acknowledged and discussed, including the challenges linked to imbalanced and cost-sensitive classification, the importance of data quality and sound preprocessing procedures, model interpretability, quantification of prediction uncertainty, and the challenges related to model selection and hyperparameters tuning. While the trajectory of progress suggests an increasing adoption of AI tools, domain knowledge and human expertise remain pivotal, ensuring effective oversight of intelligent systems, understanding the limitations of ML models, and contextualizing their predictions.

Acknowledgments

The completion of this Ph.D. journey is a culmination of relentless effort, patience, and profound interactions with many who have enriched both my research and my personal growth. It is my pleasure to express my deep gratitude to everyone who has contributed to this journey.

I would like to begin by expressing my sincere appreciation to Professors Nicola Paltrinieri and Valerio Cozzani. Your belief in my potential and your guidance through the intricate challenges of research have consistently surpassed expectations. The breadth and depth of your knowledge in process safety and engineering have been an inexhaustible source of inspiration. Beyond academia, your kindness and rectitude have provided invaluable life lessons. I owe a significant portion of my personal and professional growth to your mentorship.

A special thanks to Professor Tongwen Chen for the opportunity to deepen my understanding in alarm management and for the warm welcome during my stay at the University of Alberta. Your expertise has been pivotal in enhancing my research. I also wish to thank the Advanced Alarm Management team at UoA. Special thanks to Hari and Mani for their company and for making me feel at home.

To my colleagues in Bologna, I feel extremely fortunate to be part of such an outstanding team; thank you! Giordano and Alessandro, your contributions to my research have been invaluable. To my office mates, Carmen and Leonardo, thank you for the moments we have shared. To Federica, your support and friendship through the ups and downs of academic life have been truly comforting.

I am blessed to have my long-standing friends in Ferrara. Your presence has provided countless laughs, stimulating conversations, and cherished memories. Thank you for being an example of true friendship.

To Chiara, my heartfelt thanks for illuminating my days with joy, unforgettable memories, and unyielding faith in my abilities. Your unwavering support and understanding have been the foundation of my strength.

Last but certainly not least, my deepest gratitude goes to my family. To my brother, your companionship and intellectual curiosity is a constant source of inspiration. Thank you for the enlightening conversations and the shared enthusiasm for knowledge; your like-minded spirit and constant presence have been invaluable. To my mother, grandmother, father, and all the members of my family, your guidance, wisdom, rectitude, and inspiration have molded me into the person I have become today. Thank you for being my anchor, keeping me grounded when storms try to sway me.

Table of contents

List of Figures

List of Tables

Acronyms

AF	Alarm Flood
AI	Artificial Intelligence
CBOW	Continuous Bag-of-Words
CPS	Chemical Process Safety
CFD	Computational Fluid Dynamics
CE	Cross Entropy
CoF	Consequence of Failure
DOE	Design of Experiments
DRA	Dynamic Risk Assessment
DRM	Dynamic Risk Management
FEM	Finite Element Method
FMEA	Failure Mode and Effect Analysis
FN	False Negative
FP	False Positive
HAZOP	Hazard and Operability study
IoT	Internet of Thing
ISOMAP	Isometric Mapping
LASSO	Least Absolute Shrinkage and Selection Operator
LOPA	Layer of Protection Analysis
MHIDAS	Major Incident Data Source
ML	Machine Learning
MSE	Mean Squared Error
NN	Neural Network

NR	Narrative review
NLP	Natural Language Processing
PCA	Principal Component Analysis
RBI	Risk-Based Inspection
RL	Reinforcement Learning
RM	Risk Management
RMSE	Root Mean Squared Error
OSM	Process Safety Management
SR	Systematic Review
TN	True Negative
TP	True Positive
TTF	Time-To-Failure

Thesis structure

This doctoral thesis is a collection of articles and it is structured in two main parts:

- Part I, the main report, interrelates the articles and summarises the research performed during the entire PhD study;
- Part II, where the articles published within the Ph.D. study are collected.

It is suggested to read the two parts in the proposed order. However, they are independent and can be read in any order.

Publications

The scientific publications produced in this Ph.D. study are listed in Table 1. Full-text articles are presented in Part II.

Table 1. List of published articles. "J" indicates a publication in a peer-reviewed journal, "C" refers to a publication in the proceedings of a peer-reviewed conference.

Article no.	Type	Title
I	J	Artificial Intelligence for Safety and Reliability: A Systematic Review Focusing on Machine Learning
II	J	Learning From Major Accidents: A Machine Learning Approach
III	J	Learning From Major Accidents: A Meta-Learning Perspective
IV	C	Predicting the Consequences of Hydrogen Releases: How a Machine Learning Approach May Improve Risk-Based Inspection Planning
V	J	A Neural Network Approach to Predict the Time-to-Failure of Atmospheric Tanks Exposed to External Fire
VI	C	A Machine Learning Approach to Predict Chattering Alarms
VII	J	Predicting Chattering Alarms: A Machine Learning Approach
VIII	C	A Data-Driven Approach to Improve Control Room Operators' Response
IX	C	Online Classification of Alarm Floods Using a Word2vec Algorithm
X	C	Integration Between Data-Driven Process Simulation Models and Resilience Analysis to Improve Environmental Risk Management in the Waste-To-Energy Industry
XI	J	Assessment of Safety Barrier Performance in Environmentally Critical Facilities: Bridging Conventional Risk Assessment Techniques with Data-Driven Modelling
XII	C	Are we Going Towards "No-Brainer" Risk Management? A Case Study on Climate Hazards

Additional details about the articles and the contributions made by the authors are available below.

Article I:

Tamascelli, N., Campari, A., Parhizkar, T., Paltrinieri, N. (2023). Artificial Intelligence for Safety and Reliability: A Systematic Review Focusing on Machine Learning. Journal of Loss Prevention in the Process Industries. *Revised version submitted for publication*.

Contribution of authors:

The first author wrote the original draft, initiated the research ideas, performed formal analyses, curated data, and participated in the conceptualization of the study. The second author wrote the original draft, designed the methodology, participated in the conceptualization of the study, and curated data. The third author wrote the original draft, participated in the conceptualization of the study, and curated data. The fourth author provided supervision, participated in the conceptualization of the study, reviewed the manuscript, and provided critical feedback.

Article II:

Tamascelli, N., Solini, R., Paltrinieri, N., Cozzani, V. (2022). Learning from major accidents: A machine learning approach. Computers & Chemical Engineering, 162, 107786. https://doi.org/10.1016/j.compchemeng.2022.107786.

Contribution of authors:

The first author wrote the original draft, conducted formal analyses, participated in the study conceptualization, and tested/improved the algorithms. The second author initiated the research idea, developed the methodology, conducted formal analyses, and developed the algorithms. The third author initiated the research idea, contributed to the methodology, supervised, reviewed, and edited the manuscript. The fourth author reviewed the manuscript and provided critical feedback.

Article III:

Tamascelli, N., Paltrinieri, N., & Cozzani, V. (2023). Learning From Major Accidents: A Meta-Learning Perspective. Safety Science, 158, 105984. https://doi.org/10.1016/j.ssci.2022.105984.

Contribution of authors:

The first author wrote the original draft, conducted formal analyses, participated in the study conceptualization, wrote the algorithms, and contributed to the development and design of the methodology. The second author initiated the research idea, developed the methodology, and supervised, reviewed, and edited the manuscript. The fourth author reviewed the manuscript and provided critical feedback.

Article IV:

Giannini, L., Tamascelli, N., Salzano, E., Paltrinieri, N. (2023). Predicting the Consequences of Hydrogen Releases: how a Machine Learning Approach May Improve Risk-Based Inspection Planning. Proceedings of the PSAM 2023 Topical Conference on AI & Risk Analysis for Probabilistic Safety/Security Assessment & Management. *Accepted for publication*.

Contribution of authors:

The first author wrote the original draft, initiated the research ideas, and participated in conceptualizing and developing the methodology. The second author provided supervision, conducted formal analyses, curated data, reviewed and edited the manuscript. The third author reviewed the manuscript and provided critical feedback. The fourth author initiated the research ideas, provided supervision, reviewed the manuscript and provided critical feedback.

Article V:

Tamascelli, N., Scarponi, G.E., Amin, M. T., Sajid, Z., Paltrinieri, N., Khan, F., Cozzani, V., (2023). A Neural Network Approach to Predict the Time-to-Failure of Atmospheric Tanks Exposed to External Fire. Reliability Engineering & System Safety. *Revised version submitted for publication.*

Contribution of authors:

The first author wrote the original draft, initiated the research ideas, developed the methodology and the algorithms, participated in the formulation of research goals, curated and validated the data. The second author initiated the research ideas, participated in the formulation and evolution of research goals, produced the data, supervised, reviewed, and edited the manuscript. The third, fourth, fifth, and sixth authors reviewed the manuscript providing feedback and improving the quality. The seventh author initiated the research ideas, participated in the formulation of research goals, supervised, reviewed, and edited the manuscript.

Article VI:

Tamascelli, N., Arslan, T., Shah, S.L., Paltrinieri, N., Cozzani, V. (2020). A Machine Learning Approach to Predict Chattering Alarms. Chem. Eng. Trans. 82. https://doi.org/10.3303/CET2082032.

Contribution of authors:

The first author wrote the original draft, initiated the research ideas, developed the methodology and the algorithms, participated in the formulation of research goals, curated and validated the data. The second author supported the development of the algorithm, reviewed and edited the manuscript. The third author reviewed and edited the manuscript, improving the overall quality. The fourth author participated in formulating research goals and developing the methodology, provided supervision, reviewed and edited the manuscript. The fifth author reviewed the manuscript and provided critical feedback.

Article VII:

Tamascelli, N., Paltrinieri, N., & Cozzani, V. (2020). Predicting Chattering Alarms: A Machine Learning Approach. Computers & Chemical Engineering, 143, 107122. https://doi.org/10.1016/j.compchemeng.2020.107122.

Contribution of authors:

The first author wrote the original draft, initiated the research ideas, developed the methodology and the algorithms, participated in the formulation of research goals, curated and validated the data. The second author participated in formulating research goals and developing the methodology, provided supervision, reviewed and edited the manuscript. The third author reviewed the manuscript and provided critical feedback.

Article VIII:

Tamascelli, N., Scarponi, G.E., Paltrinieri, N., Cozzani, V., (2021). A data-driven approach to improve control room operators' response. Chem. Eng. Trans. 86, 757–762. https://doi.org/10.3303/CET2186127.

Contribution of authors:

The first author wrote the original draft, initiated the research ideas, participated in developing the methodology, curated data, and performed formal analyses. The second author reviewed and edited the manuscript. The third author provided supervision, initiated research ideas, participated in developing the methodology, reviewed and edited the paper. The fourth author reviewed the paper and provided feedback.

Article IX:

Tamascelli, N., Rao, H. R. M., Cozzani, V., Paltrinieri, N., Chen, T. (2023). Online Classification of Alarm Floods Using a Word2vec Algorithm. IECON 2023 – 49th Annual Conference of the IEEE Industrial Electronics Society, Singapore, 2023. https://doi.org/10.1109/IECON51785.2023.10312435.

Contribution of authors:

The first authors wrote the original draft, initiated the research ideas, developed the methodology and the algorithms, performed formal analysis, curated and validated data. The second author wrote the original draft, participated in developing the algorithms and curating data. The third and fourth authors reviewed the manuscript and provided critical feedback. The fifth author provided supervision and financial support, reviewed the manuscript, and delivered critical feedback.

Article X:

Tamascelli, N., Dal Pozzo, A., Liu, Y., Cozzani, V., Paltrinieri, N. (2022). Integration between data-driven process simulation models and resilience analysis to improve environmental risk management in the Waste-

to-Energy industry. Proceedings of the 32nd European Safety and Reliability Conference (ESREL 2022), Dublin, Ireland, 2022. 1409–1416. https://doi.org/10.3850/978-981-18-5183-4_R23-03-206-cd.

Contribution of authors:

The first author wrote the original draft, initiated the research ideas, developed part of the algorithms, curated and validated data, and performed formal analysis. The second author initiated the research ideas, provided supervision, reviewed and edited the manuscript. The third author provided supervision, reviewed the paper, and provided critical feedback. The fourth and fifth authors initiated the research ideas, provided supervision, reviewed the manuscript, and delivered critical feedback.

Article XI:

Tamascelli, N., Dal Pozzo, A., Scarponi, G.E., Paltrinieri, N., Cozzani, V. (2024) Assessment of Safety Barrier Performance in Environmentally Critical Facilities: Bridging Conventional Risk Assessment Techniques with Data-Driven Modelling. Process Safety and Environmental Protection. https://doi.org/10.1016/j.psep.2023.11.021.

Contribution of authors:

The first author wrote the original draft, initiated the research ideas, developed part of the algorithms, curated and validated data, and performed formal analysis. The second author wrote part of the original draft, initiated the research ideas, provided supervision, and curated data. The third author provided supervision, reviewed the paper, and provided critical feedback. The fourth and fifth authors initiated the research ideas, provided supervision, reviewed the manuscript, and delivered critical feedback.

Article XII:

Tamascelli, N., Nakhal Akel, A.J., Patriarca, R., Paltrinieri, N., Cruz, A.M. (2022). Are we going towards "no-brainer" risk management? A case study on climate hazards. Proceedings of the 16th Probabilistic Safety Assessment & Management Conference (PSAM16), Honolulu, Hawaii, 2022. ISBN: 9781713863755.

Contribution of authors:

The first author wrote the original draft, contributed to the conceptualization of the study, curated data, and performed formal analyses. The second and third authors reviewed and edited the manuscript. The fourth author initiated the research ideas, provided supervision, and wrote part of the original draft. The sixth author reviewed the manuscript and provided critical feedback.

Part Ⅰ

Main report

1. Introduction

Major accidents involving dangerous substances can severely impact human health, the environment, and company finances. Within the chemical and process industry, large quantities of hazardous substances are stored and handled during daily operations, elevating the risk of accidental releases with potentially disastrous consequences. However, while industrial activities have always carried significant risks, there were no standardized measures in place to manage the risk posed by hazardous substances before the second half of the sixties (Kletz, 2012; Pasman et al., 1992). At that time, handling and storing dangerous substances were regulated by traditional occupational safety and good engineering practices (Abdul Aziz and Mohd Shariff, 2017). Later, a series of terrible accidents – including Woodbine (1971), Seveso (1976), Bhopal (1984), and Pasadena (1989) – highlighted the need to go beyond the existing standard and develop a different approach to prevent major accidents and their consequences. Those unfortunate events were the driving force for the formulation and development of modern safety management programs (Abdul Aziz and Mohd Shariff, 2017), leading to the establishment of what is now called Process Safety Management (PSM) (Khan et al., 2016).

During the last three decades, the discipline of Risk Management (RM) – defined as the study of methods, tools, and techniques aimed at identifying, quantifying, and controlling risk – has witnessed extraordinary growth, leading to its adoption and codification in international standards and regulations. Notable examples include the ISO 31000 standard (International Organization for Standardization (ISO), 2018) providing guidance on the selection and application of various techniques to model uncertainty and manage risk, the API 750 RP and Process Safety Management standards (American Petroleum Institute, 1991; Canadian Society for Chemical Engineering, 2012), focusing on process hazards, and the European Directive 2012/18/EU (European Parliament Council of the European Union, 2012), providing guidelines on the control of major-accident hazards and defining regulatory requirements for industrial facilities handling dangerous substances.

In spite of the progress made in the field of RM, accidents involving dangerous substances still occur (Pasman and Fabiano, 2021), suggesting that traditional risk management techniques may not be sufficient in controlling risk. For example, consider the terrible ammonium nitrate explosion that occurred in Beirut (2020), causing more than 200 fatalities and 6000 injuries (El Zahran et al., 2022; Pasman et al., 2020), or the chlorine gas leak that occurred in Aquaba, Jordan (2022) taking the life of 13 and injuring more than 260 (Gritten, 2022). In fact, several authors have highlighted that canonical risk management techniques have inherent limitations (Villa et al., 2016), such as the inability to capture the risk variations resulting from changes in operative conditions. In other words, RM methods apply static reasoning to describe a dynamic environment, capturing a static risk picture that does not reflect the intricate dynamics between Risk Influencing Factors. In order to solve these limitations, recent research has proposed to go beyond traditional RM frameworks toward a more dynamic and proactive approach to process safety, called Dynamic Risk

Management (DRM). This new discipline advocates for methods and techniques that can capture risk variations in evolving and degrading systems, taking into account the performance of the control system, safety barriers, inspection and maintenance activities, the human factor, and procedures (Khan et al., 2016). DRM relies on the ability to monitor operative conditions and update the risk picture as new observations become available (Paltrinieri et al., 2014a). To this end, new and inherently updatable methods are needed, in opposition to the static tools currently in use in traditional risk management frameworks. In this context, most research has focused on Bayesian Networks and Petri nets to model the relationships between Risk Influencing Factors and update the risk picture (Kabir and Papadopoulos, 2019). Other approaches have proposed the use of Dynamic Fault Trees (Gascard and Simeu-Abazi, 2018), Monte Carlo simulation (Rabiti et al., 2013), and Markow models (Sievers and Madni, 2022). However, the research on DRM is still in its infancy. The literature appears scattered and lacks cohesion; there are few contributions on the topic, highlighting an urgent need for new, proactive tools to capture the dynamics of unsafe interactions in increasingly complex and interconnected systems (Zio, 2018).

The digitalization of industrial systems has revolutionized the manufacturing industry. The widespread use of remote sensing, Internet of Things technologies, and cloud storage has tremendously increased the ability to monitor and control industrial processes (Lee et al., 2019). At the same time, advancements in computing technologies and the advent of new data analysis tools open interesting opportunities to learn from data and extract safety-relevant knowledge. For example, the advent of Artificial Intelligence (AI) and Machine Learning (ML) is catalyzing a profound transformation in our world. These technologies are not only reshaping our daily interactions and digital experiences but are also announcing a new era for industrial processes (Peres et al., 2020). ML methods hold great potential to advance the research on DRM (Paltrinieri et al., 2019), allowing the development of inherently updatable, proactive, and dynamic techniques that can capture the intricate dynamics of complex phenomena and predict how changes in process conditions affect the risk picture. However, the idea of integrating ML into a DRM framework is relatively recent, and the topic is still largely unexplored, missing the chance to push the research on DRM methods, which seems to stagnate and not keep the peace with technological advancements (Pasman and Fabiano, 2021).

In this context, this thesis aims to contribute to the research on investigating ML methods to support the ambitious objectives of DRM within the chemical and process industry. This entails exploring the potentials and limitations of ML techniques, suggesting practical solutions for addressing various DRM tasks, and examining the implications and prospective roles of ML and humans in the future of DRM. After defining the state of the art of ML methods in safety and reliability, this research explores practical ML-based methodologies to support different phases of DRM, such as consequence evaluation, frequency evaluation, and monitoring of safety barriers. In addition, the investigation is complemented with considerations on the future of risk management, specifically focusing on the roles of humans and machines in the future of process safety.

This manuscript is organized as follows. Section 2 describes the research background, providing a brief introduction to Risk Management, Dynamic Risk Management, and Machine Learning. Section 3 and 4 describe the aim and scope of this investigation, defining research questions, and revealing the connection between publications and objectives. Section 5 focuses on the research methodology, providing details on the characteristics of the research approach adopted and on quality assurance criteria. Section 6 describes the methods utilized within this Ph.D. study. The contributions of this PhD study are reported in Section 7 and discussed in Section 8. Finally, conclusions are drawn in Section 9, along with ideas for future research.

2. Research background

This section offers a comprehensive background for the Ph.D. study. It aims to establish the context of the investigation by emphasizing the present state of research on the subject, underscoring existing limitations, and identifying knowledge gaps. Specifically, Section 2.1 delves into Risk Management and Dynamic Risk Management, detailing their frameworks, activities, and associated challenges. Meanwhile, Section 2.2 focuses on Machine Learning, outlining its fundamentals, describing various ML approaches, and highlighting its potential in supporting and enhancing DRM.

2.1. Risk Management and Dynamic Risk Management

Risk management (RM) refers to the set of "coordinated activities to direct and control an organization with regard to risk" (International Organization for Standardization (ISO), 2018). In other words, it concerns quantifying, evaluating, prioritizing, and controlling the risk associated with a particular activity. The development and establishment of RM as a scientific discipline began approximately 30 to 40 years ago, driven by the need to regulate the design and operations of high-risk industrial activities (Aven, 2016), such as nuclear, chemical, and process industries. Today, RM is a widely-accepted and established process, described and codified in international standards (International Organization for Standardization (ISO), 2018), and adopted by international regulations, such as the Directive 2012/18/EU of the European Parliament and Council (European Parliament Council of the European Union, 2012).

The RM process comprises five main activities, as represented in Figure 1 and discussed in Section 2.1.1.

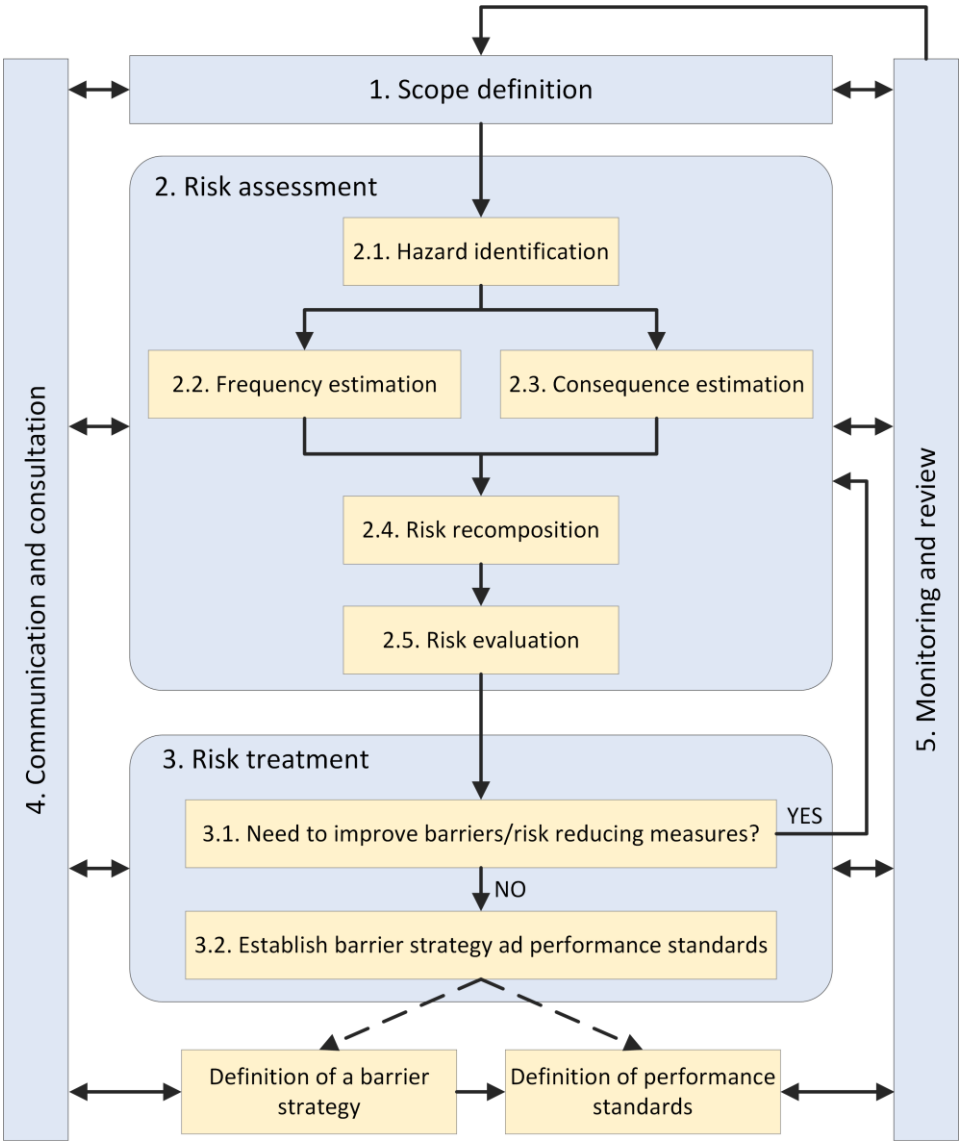


Figure 1. Overview of the Risk Management framework highlighting the role of Safety Barriers in risk treatment. Adapted from (Petroleum Safety Authority, 2013)

2.1.1. The Risk Management process

The Risk Management process comprises the following activities:

1. Scope definition;
2. Risk assessment;
3. Risk treatment;
4. Communication and consultation;
5. Monitoring and review.

The first activity involves specifying the boundaries of the analysis. Since the risk management process can be applied at various levels within the plant lifecycle, such as design, construction, and operation, clarity regarding the scope is paramount. This involves understanding the objectives to be addressed and aligning them with organizational objectives. Also, organizations must define the level and type of risk they are willing to accept in relation to their objectives and external constraints (e.g., regulatory requirements).

The second activity, namely risk assessment, aims to quantify the risk levels related to a particular activity. This involves identifying the hazards (step 2.1 in Figure), estimating the frequency and the potential consequences of unwanted events (steps 2.2 and 2.3 in Figure), calculating a risk measure (step 2.4 in Figure) and comparing the risk levels with the established risk criteria to determine whether the risk is acceptable (step 2.5). Risk assessment may be considered one of the most challenging and time-consuming parts of the RM process (Lees, 2012). It requires specialized knowledge and the use of articulated techniques, such as Hazard and Operability study (HAZOP), Failure Mode and Effect Analysis (FMEA), Layer of Protection analysis (LOPA), Bow-Tie analysis, dispersion models, fire and explosion models. Also, the analysis must consider the presence and effects of preventive and protective measures, also called safety barriers, installed to prevent, mitigate, or control unwanted events (Sklet, 2006). In fact, barrier assessment and management is an integral part of Risk management (Petroleum Safety Authority, 2013) and requires an in-depth understanding of technical, operational, and organizational measures installed to reduce risk, including safety-critical equipment, such as the alarm system and fire protection devices, as represented in Figure 2.

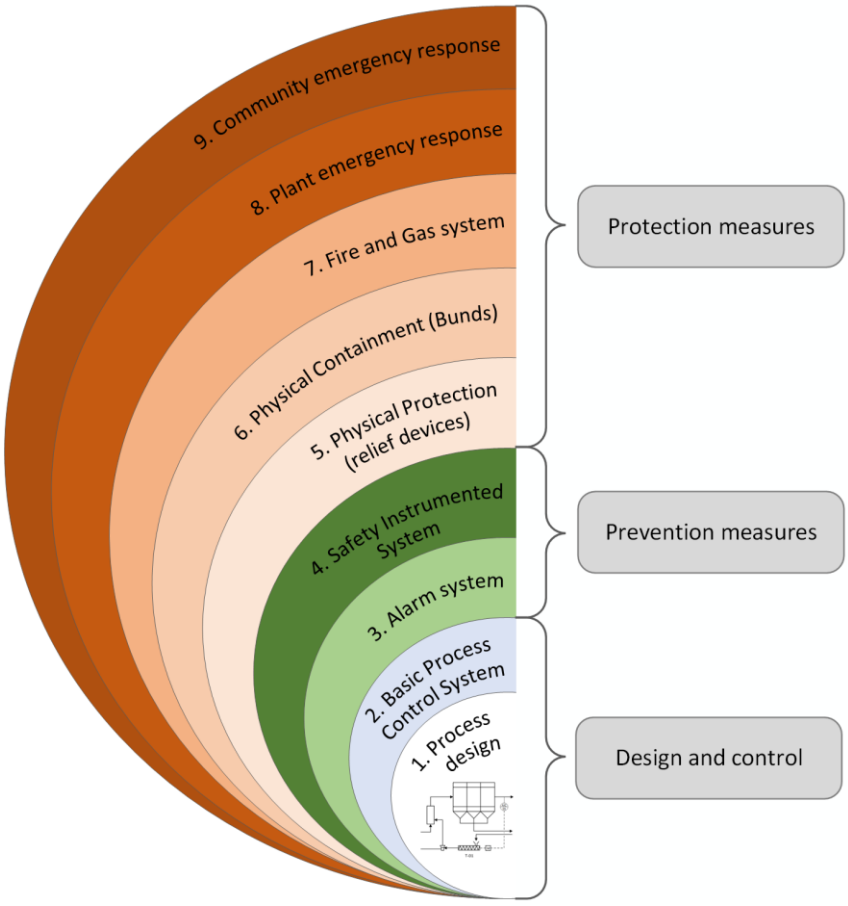


Figure 2. Layers of defense against a possible accident. Adapted from (Center for Chemical Process Safety, 2010).

The third activity, called Risk treatment, allows the selection and implementation of additional preventive and protective measures to lower the risk and meet the risk acceptance criteria. When the residual risk is deemed acceptable, this activity guides the formulation of a strategy to manage and operate safety barriers, along with performance criteria to measure their performance.

The fourth activity involves recording, reporting, and communicating the outcome of the analysis to internal and external stakeholders in order to (i) ensure that risk management activities are communicated and disseminated across the organization and (ii) enable risk-informed decision-making.

The last activity, called "Monitoring and review", ensures that the risk management process and its outcomes are periodically reviewed and improved to guarantee their quality. This is a continuous activity that permeates the entire Risk Management process. It assumes particular importance in the operational phase, ensuring that the plant operations align with assumptions, requirements, and technical conditions. The monitoring and review activity aims to ensure that conditions within the plant do not deviate from the assumptions made during the risk assessment activity, guaranteeing that estimated risk levels remain valid through the entire plant lifecycle. This multifaceted activity involves monitoring the performance levels of safety barriers, coupled with the formulation of precise protocols that solidify technical integrity, covering areas like startup, shutdown, inspection, and maintenance procedures. The activity acts as a bulwark against potential deviations and loss of control over risk-influencing factors, effectively preventing the potential transformation of unforeseen deviations into incidents or accidents. Furthermore, the monitoring and review phase ensures that valuable lessons are derived from incidents, should they occur, preventing their reoccurrence and improving the effectiveness of the RM process.

2.1.2. Limitations of traditional risk management

While playing a pivotal role in controlling risks and ensuring safe operations, RM remains a relatively young and ever-evolving discipline with some inherent limitations (Villa et al., 2016). Firstly, traditional RM appears not to offer and support an effective learning strategy, resulting in the reoccurrence of similar accidents. In fact, several authors have pointed out that "the chemical industry as a whole does not learn from its past mistakes" (Chung and Jefferson, 1998). After more than 10 years, Pasman (2009) and Le Coze (2013) noted that little progress has been made; similar accidents keep happening, and organizations struggle to derive and apply lessons from past accidents. Another limitation of RM is that it produces static results, while the risk is dynamic and varies as the plant ages and operative conditions change (Kalantarnia et al., 2009). In fact, Risk Assessment – one of the core activities of risk management – returns a static risk picture, which is limited to reflect the risk level in a very specific plant configuration, but cannot be used to estimate how the risk varies with time. Due to the substantial efforts demanded by the techniques used to assess risk levels, the Risk Assessment activity is typically re-iterated only every five years, in conjunction with major changes in the plant configuration or operation, or after major accidents (European Parliament Council of the European Union, 2012). However, conditions within process plants are dynamic and constantly evolving (Paltrinieri and Reniers, 2017). Equipment ages and degrades, and operational conditions can be altered due to technical failures, feed variability, wrong settings, improper methods, and human actions (Hashemi et al., 2014), such as the

misinterpretation of an alarm. In this context, traditional RM lacks in comprehensively addressing potential issues, as its instruments are not designed to be easily updatable and capture fluctuations in risk arising from process deviations (Kalantarnia et al., 2009).

Another criticism of the RM process is that formal probabilistic risk analysis techniques, such as fault trees and event trees, rely on "hard" logic (e.g., "AND" and "OR" gates in bow-tie diagrams) to represent causal connections that are "soft" and "partial" in nature (Vatn, 2012). In other words, traditional RM techniques lack the capacity to effectively depict uncertain causal connections among risk-influencing factors, particularly when it comes to human and organizational factors, which are notably challenging to describe in probabilistic contexts (Vinnem et al., 2012). Furthermore, the increasing complexity of industrial facilities presents substantial obstacles, as conventional RM methods struggle to consider the interplay among the risk-influencing factors; as their number increases, they become arduous to model and incorporate into the traditional framework (Villa et al., 2016).

2.1.3. Dynamic Risk Management

In response to the acknowledged limitations and criticisms leveled at the conventional RM process, recent research has shifted its focus toward investigating and shaping what is referred to as "Dynamic Risk Management" (DRM). This new discipline aims to proceed beyond the canonical RM process toward the definition of tools and methods that can capture the dynamic evolution of process conditions and their effect on risk levels (Paltrinieri et al., 2014a).

Two of the main characteristics that differentiate DRM from its static counterpart are (Khan et al., 2016):

- the utilization of a Dynamic Risk Assessment (DRA) activity, in opposition to the static Risk Assessment process discussed in Section 2.1.1;
- strengthened and more efficient monitoring and review practices that allow for (i) simultaneous monitoring of numerous interconnected process variables, (ii) continuous tracking of process and operational modifications, and (iii) extraction and application of insights from previous failures and incidents.

Dynamic Risk Assessment refers to a set of methods "that update estimated risk of a deteriorating process according to the performance of the control system, safety barriers, inspection and maintenance activities, the human factor, and procedures" (Khan et al., 2016). The difference between traditional RA and DRA lies in the tools used to produce the risk picture. DRA makes use of tools and techniques that are designed to be updatable, while the methods used in RA are static and difficult to update. This distinction becomes clearer with an analogy borrowed from the realm of photography: consider RA as resembling an older Polaroid camera. While capable of capturing multiple images, the intervals between shots are constrained by the camera's mechanical limitations, rendering it unsuitable for capturing dynamic subjects. In contrast, Dynamic Risk Assessment (DRA) can be likened to a modern digital camera. Involving fewer moving components and

minimized downtime, it facilitates rapid image capture, akin to capturing even the most nuanced subject movements. Another important distinction is that DRA methods should be designed to consider information from the plant operations (e.g., from sensor measurements, maintenance activities, alarm logs) and about past failures and incidents. Such information is made available thanks to an improved monitoring and review phase. Therefore DRM involves a continuous exchange of information between the plant (e.g., from sensor measurements, maintenance activities, alarm logs) and DRA techniques in order to update the risk picture as conditions changes. As such, DRM strongly relies on data, both present – reflecting current process conditions – and past – shedding light into accident precursors and enabling proactive countermeasures.

In recent years, research into Dynamic Risk Management (DRM) has experienced remarkable growth, largely propelled by technological advancements that have become deeply established within the manufacturing industry. Notably, the progress in remote sensing and the Internet of Things (IoT) has enabled the collection of a multitude of process variables. Concurrently, enhancements in storage capacities have facilitated the retention of extensive process data spanning numerous years. The development of accident databases has made information about past accidents easily accessible. Meanwhile, improved computational capabilities have opened doors to deploying advanced techniques such as Computational Fluid Dynamics (CFD), Finite Element Method (FEM), Digital Twins, Montecarlo simulations, Bayesian Networks (BN), and Machine Learning. Exploiting these enabling technologies, many researchers have harnessed their potential to craft tools suited for DRM. Notable examples include the use of Bayesian Networks for Dynamic Risk Assessment (Dimaio et al., 2021; Kalantarnia et al., 2009; Khakzad et al., 2018; Vinnem et al., 2012; Zeng et al., 2020; Zeng and Zio, 2018), the so-called DyPASI method for the dynamic identification of hazards (Paltrinieri et al., 2013), the Risk Barometer (Hauge et al., 2015; Paltrinieri et al., 2014b) for the monitoring of safety barriers and assessment of early deviations and accident precursors, and early attempts to leverage Machine Learning techniques and support DRA (Paltrinieri et al., 2019).

However, despite the notable advancements, DRM may still be considered in its early stages. It is an ambitious and largely unexplored topic, facing challenges in garnering industrial backing and widespread adoption (Taleb-Berrouane and Pasman, 2022). Overcoming obstacles and limitations of existing DRM methods is imperative, including addressing the intricacies of modeling complex dependencies among risk factors and moving beyond point-based probability values, which overlook the inherent uncertainty tied to probability estimations (Khan et al., 2016).

2.2. AI and Machine Learning

Artificial Intelligence (AI) may be defined as “the part of computer science concerned with designing intelligent computer systems” (Barr and Feigenbaum, 1981). As the definition suggests, AI is a vast and ever-changing field with different domains, methods, and applications such as Natural Language Processing, Computer Vision, Expert Systems, Robotics, and more (Finlay and Dix, 2020). Among the different branches

of AI there is Machine Learning (ML), which refers to "a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kind of decision making under uncertainty" (Murphy, 2012). In other words, Artificial Intelligence indicates the broader set of technologies designed to mimic the capabilities of the human mind, while Machine Learning may be seen as a subset of AI focusing on computer programs that improve their performance through experience.

Machine Learning methods can be broadly divided into four macro categories: Supervised Learning, Unsupervised Learning, Semi-supervised Learning, and Reinforcement Learning. The choice of the most appropriate method depends on many factors, such as the nature of the problem under assessment, data availability, and the expected output. A brief description of each ML category is provided below and summarized in Table 2.

Supervised Learning makes use of a set of existing observations to uncover the system dynamic and learn a mapping between input and output variables (Sammut and Webb, 2017a). The approach is used when the problem involves the prediction of an output measure, also called the "label", given a set of input variables, also called the "features". This approach is called "supervised" because it involves training an ML model using a labeled dataset, where each observation is associated with a corresponding output label. In other words, during the training process, the algorithm is "supervised" by providing the correct answers for a set of input data. Depending on whether the label is categorical or real-valued, there are two primary examples of Supervised Learning: Classification and Regression. Classification is used when the problem requires the categorization of observations into one or more classes. Instead, regression approaches are used when the problem involves the prediction of a real-valued label.

Unsupervised Learning is used when no output measure is available or when the aim of the analyses is knowledge discovery rather than knowledge-based learning (Sammut and Webb, 2017b). In this approach, unlabeled observations are fed to the learner, which seeks to find latent structure in the feature space with little or no supervision. In other words, the model is not constrained to fit an expected output. Instead, it is free to explore the feature space in search of correlations and hidden structures between features, hence the term "Unsupervised". Common examples of Unsupervised approaches are Clustering and Dimensionality Reduction. The former is used to group observations into clusters so that similar observations fall within the same cluster (Han et al., 2011). The latter is used to find low-dimensional structures hidden in high-dimensional observations (Van Der Maaten et al., 2009).

Semi-supervised Learning is used when there exists partial labeled information (van Engelen and Hoos, 2020). In other words, only a limited number of the samples have associated a label. This situation is frequent in real-world scenarios, where observations are abundant, but the labeling process is both time-consuming and resource-intensive, resulting in a scarcity of labeled data and a surplus of unlabeled data (Zhu, 2017). Semi-supervised algorithms typically use labeled data to guide the learning process, helping the model understand the relationships between input features and their corresponding outputs. Meanwhile, unlabeled data aids

in capturing the underlying structure and patterns within the data. By incorporating information from both labeled and unlabeled examples, the model can make more accurate predictions and better generalize to new, unseen data points.

Supervised, Unsupervised and Semi-supervised learning algorithms aim to map data from input features to outputs by minimizing a relevant cost function. Reinforcement Learning (RL) stands apart by introducing an agent and environment (Li, 2017). The agent makes decisions and takes actions in an environment to optimize rewards or penalties. Unlike other machine learning paradigms, RL typically doesn't require labeled pairs or training samples. RL involves the definition of environmental and agent states, possible actions, and transition probabilities. Rewards or penalties characterize state transitions after actions. The agent's goal is to maximize cumulative rewards while interacting with the environment, honing effective strategies to achieve specific tasks.

Table 2. Summary of the five ML macro-categories.

ML category	Learning strategy	Required input	Tasks
Supervised	By examples	Labeled examples	• Classification • Regression
Unsupervised	By exploration	Unlabeled examples	• Clustering • Dimensionality reduction • Anomaly detection • Discover hidden patterns
Semi-supervised	Hybrid	Labeled and unlabeled examples	• Classification • Regression • Clustering
Reinforcement	Trial and error	Agent-environment interaction	• Skills acquisition • Optimize decision-making • Control strategies

Machine Learning offers several potential advantages to overcome the challenges and limitations of actual RM strategies and proceed further to achieve the ambitious goals of DRM. Particularly:

- ML is designed to make predictions under uncertainty. Considering that risk is uncertainty regarding events and their consequences (Aven and Renn, 2009), it appears evident that ML may offer valuable tools and techniques to support Risk Management.
- ML can model complex nonlinear relationships among features, which is crucial considering the increasing complexity and interconnectedness of industrial environments.
- ML models can generalize to previously unseen scenarios, which is highly relevant to process safety, where rare and unexpected events are common challenges.

- ML appears to be a valuable tool to take advantage of the wealth of heterogeneous data made available by the widespread digitalization of industrial processes in order to extract and retain safety-relevant information.
- ML learning models offer fast predictions, making them well-suited for handling large volumes of high-frequency data, effectively addressing the dynamic nature demanded by DRM techniques.

The potential utilization of ML to enhance risk management has captivated safety researchers for the past two decades. In 1992, Diekmann (1992) stated that "future approaches to risk analysis will certainly rely more on the advances being made in artificial intelligence and the cognitive sciences". While ML has been effectively employed in tasks like fault detection, diagnosis, and anomaly detection (Xu and Saleh, 2021), its application in the domain of DRM remains relatively underexplored, particularly in the context of chemical and process engineering (Hegde and Rokseth, 2020). Although recent contributions have showcased notable progress (Paltrinieri et al., 2019), the landscape remains fragmented, with numerous unaddressed topics and untapped potentialities.

3. Research questions

This Ph.D. project is motivated by the need to overcome the constraints of conventional RM methods and the requirement for the development of proactive strategies in process safety. AI and Machine Learning present a promising opportunity for crafting tools suited for DRM. Nevertheless, this area remains relatively underexplored.

Dynamic Risk Management presents ambitious goals and necessitates addressing several challenges, including the requirement for dynamic methodologies capable of modeling intricate, nonlinear relationships between risk-influencing factors. In this context, Machine Learning holds a significant potential to improve and support DRM, particularly in the light of the prevailing digitization trends, the advances in computational capabilities, and the evolution of sophisticated data analytics algorithms. Despite these promising prospects, the exploration of ML within DRM is advancing at a measured pace. It is still unclear how and to what extent ML can contribute to the DRM paradigm. In this context, the research question that this project aims to answer is:

> *"How can advancements in digital technologies and Machine Learning be harnessed to effectively address the objectives of Dynamic Risk Management?"*

Moreover, given the criticality of the "Risk Assessment" and "Monitoring and review" activities, and considering the burden associated with the tools traditionally employed to address these tasks, main research question can be segmented into three distinct subquestions, each delving into a specific facet of DRM. In particular:

Question 1.1. "How can machine learning algorithms be developed or adapted to analyze historical accident data for more accurate consequence evaluation?

Question 1.2. "What methodologies can be employed to use machine learning techniques for frequency estimation in risk assessment?"

Question 1.3. "How can machine learning models be designed and implemented to continuously monitor, evaluate, and enhance the effectiveness of safety barriers?"

The three research subquestions deal with the tangible impact of ML in enhancing and supporting DRM activities. However, in a rapidly advancing technological age, investigating the coexistence and synergies between ML and human expertise, as well as understanding the potential and limitations of ML approaches in guiding decision-making, stands as a crucial point of exploration in paving the way forward for effective and innovative Risk Management strategies. Therefore, in addition to exploring potential applications, it is crucial to investigate the broader implications of adopting this technology. Hence, a fourth research subquestion is considered:

Question 1.4. "What are the potential and limitations of ML in supporting risk-based Decision-Making?"

4. Objectives and scope

The primary goal of this thesis is to contribute to the development of ML methods to enhance and support DRM within the chemical and process industry. This entails investigating the potentials and limitations of ML techniques, suggesting practical solutions for addressing various DRM tasks, and examining the implications and prospective roles of ML and humans in the future of DRM. To accomplish this goal, the following objectives are outlined, derived from the research questions presented in Section 3.

Objective 1. Evaluate the current state of the art, identify gaps, potential areas, and limitations of ML techniques in the domain of Risk Management.

Objective 2. Develop ML-based methods to support and promote Dynamic Risk Management.

- *Objective 2.1. Explore the use of ML to extract safety-relevant knowledge from heterogeneous data sources and predict the consequences of major accidents;*
- *Objective 2.2. Investigate Machine Learning-based models to support the frequency evaluation of accidents involving dangerous substances;*
- *Objective 2.3. Explore ML methods to monitor, evaluate, and improve the performance of safety barriers.*

Objective 3. Investigate the potential and limitations of ML in supporting risk-based decision-making.

Throughout this Ph.D. research, innovative tools and methodologies have been developed, highlighting the potential of ML in Dynamic Risk Management. However, it is crucial to highlight that the project does not address the practical challenges of deploying these techniques in real-world scenarios, especially the potential constraints on storage and computational resources. Moreover, the potential risks and cybersecurity concerns arising from these algorithms are not delved into in this study. Often, the proposed models aim to showcase the feasibility of the approach rather than being the definitive best in their category. Before any considerations can be made toward their full-scale integration into industrial IT systems, it is imperative that they undergo extensive testing in a real-world environment.

4.1. Overview of publications in relation to the research objectives

The connections between the publications achieved during the Ph.D. project (reported in Part II of the present document) and the objectives of the Ph.D. study are illustrated in Figure 3. Specifically, Article I contributes to Objective 1 by offering a review of the existing literature concerning the utilization of ML techniques to enhance the safety and reliability of engineered systems.

Articles II to IV cover Objective 2.1 by proposing the use of classification models for predicting the severity of major accidents involving dangerous substances. In addition, Article IV partially addresses Objective 3 by describing the potential of ML techniques to aid risk-based inspection and maintenance activities. Article V proposes regression models to predict the Time-To-Failure of atmospheric tanks exposed to external fire,

supporting frequency evaluation in domino scenarios and, thus, addressing Objective 2.2. In addition, the methods described in Article V are specifically developed to account for the effect of safety barriers, providing a practical contribution to Objective 2.3. Articles VI to IX contribute to Objective 2.3 by focusing on ML methods to improve and monitor industrial alarm systems, providing proactive tools to address alarm chatter, evaluate the response of control room operators, and perform online classification of alarm floods. It is also worth noting that Article VIII partially addresses Objective 3 because it discusses the use of classification models to support and guide control room operators by providing live feedback on the efficacy of their actions. Articles X and XI further contribute to Objective 2.3 by describing a hybrid approach, featuring traditional risk assessment techniques, regression models, and resilience analysis, aimed at evaluating alternative safety barrier configurations in environmentally critical facilities.

The last contribution, Article XII, delves deep into Objective 3, exploring the intricate interplay between human expertise and AI, critically examining their respective contributions. The discussion is supported by an interesting case study involving the use of unsupervised learning to categorize countries based on their similarity toward natural disaster exposure.

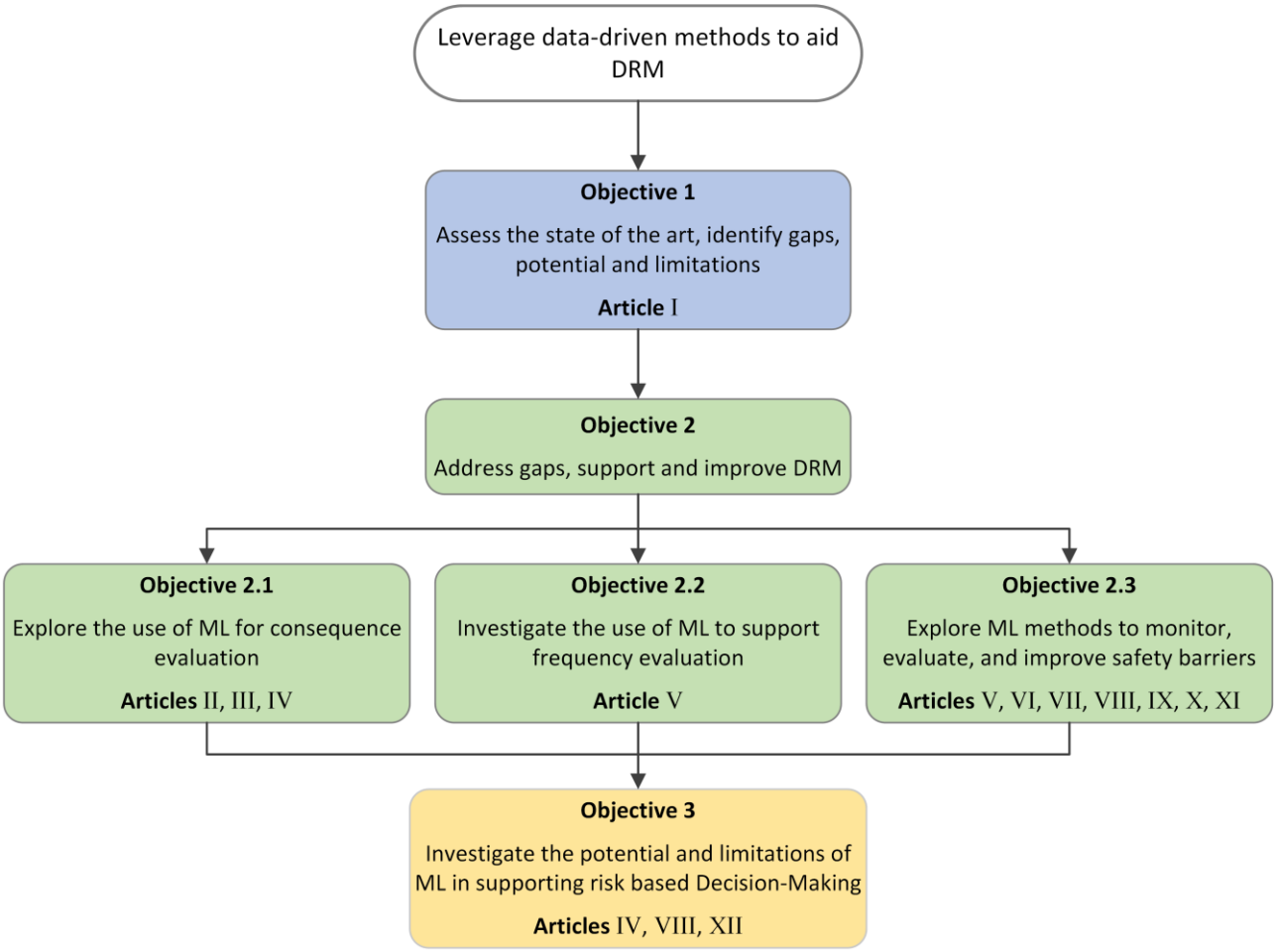


Figure 3. Illustration of the objectives of this thesis and their link with the publications reported in Part II.

5. Research methodology

Research methodology refers to a systematic approach that guides the resolution of a research problem. It can be viewed as the scientific study of how research is conducted (Kothari, 2004). In essence, research methodology involves understanding how research is systematically planned, executed, and validated.

This section describes the methodology adopted in this Ph.D., providing details about the type of research conducted (Section 5.1), interdisciplinarity characteristics (Section 5.2), type of research approaches utilized in each publication (Section 5.3), and criteria utilized to ensure the quality of the scientific production (Section 5.4). Table 3 provides a summary of the research approaches and quality assurance methods adopted in each article.

Table 3. Overview of the research approach and quality assurance criteria for each article included in this Ph.D. thesis.

Article no.	Research approach	Quality assurance
I	Mixed	- Submitted to a peer-reviewed journal - Expert judgment
II	Quantitative	- Publication in a peer-reviewed journal - Expert judgment - Test on a real case study
III	Quantitative	- Publication in a peer-reviewed journal - Expert judgment - Test on a real case study
IV	Qualitative	- Publication in a peer-reviewed conference - Expert judgment
V	Quantitative	- Submitted to a peer-reviewed journal - Expert judgment - Test on a real case study
VI	Quantitative	- Publication in a peer-reviewed conference - Expert judgment - Test on a real case study
VII	Quantitative	- Publication in a peer-reviewed journal - Expert judgment - Test on a real case study
VIII	Quantitative	- Publication in a peer-reviewed conference - Expert judgment - Test on a real case study
IX	Quantitative	- Publication in a peer-reviewed conference - Expert judgment - Test on a real case study
X	Quantitative	- Publication in a peer-reviewed conference - Expert judgment - Test on a real case study

XI	Quantitative	- Publication in a peer-reviewed journal - Expert judgment - Test on a real case study
XII	Mixed	- Publication in a peer-reviewed conference - Expert judgment - Test on a real case study

5.1. Research types

The Frascati Manual (OECD, 2015) defines three types of research activities:

- Basic research;
- Applied research;
- Experimental development.

Basic research refers to experimental or theoretical work undertaken primarily to acquire new knowledge of the underlying foundations of phenomena and observable facts, without any particular application or use in view. Applied research indicates original investigation undertaken in order to acquire new knowledge. It is, however, directed primarily towards a specific, practical aim or objective. Finally, experimental development refers to systematic work, drawing on knowledge gained from research and practical experience and producing additional knowledge, which is directed to producing new products or processes or to improving existing products or processes.

Furthermore, research activities must possess five fundamental criteria (OECD, 2015); these include:

- Novelty;
- Creativity;
- Uncertainty;
- Systematic approach;
- Transferability and/or reproducibility.

The research activities carried out within this Ph.D. predominantly align with the concept of applied research, particularly evident in the context of Objective 2, which is notably specific and pragmatic. The majority of the published articles are supplemented with real-world case studies, underscoring their explicit applicability. Nonetheless, some investigations also transcend into the domain of fundamental research due to their broader scope and engagement with underlying principles. For instance, the publications supporting Objective 3 delve into the potential and limitations of ML techniques in the context of RM and risk-based decision-making, tackling foundational topics such as expanding digitalization, the role of data-driven methodologies in enhancing safety and reliability, and the interplay of humans within the digital safety paradigm. Therefore, the research pursuits of this Ph.D. exhibit a hybrid character, positioning them at the confluence of applied and basic research.

Moreover, the research activities carried out within this Ph.D. adhere to the five fundamental criteria outlined above. Specifically, the research aimed to address gaps in knowledge and to propose innovative techniques and methodologies (novelty criterion). The methods employed and their applications showcase originality and are grounded in non-obvious concepts (creativity). The outcomes of the analyses were not predetermined, often necessitating adjustments in the research goals and a certain degree of trial and error to achieve the desired results (uncertainty). Furthermore, the research activities were meticulously planned and executed with a systematic approach. The findings have been disseminated through publication in peer-reviewed conferences and journals, contributing to their transferability. Additionally, the approaches, methods, and algorithms employed have been thoroughly documented, ensuring their reproducibility.

5.2. Interdisciplinarity

Interdisciplinarity stands as an essential requirement for addressing intricate phenomena present in reality (Stoop et al., 2017). Safety in the chemical and process industry is certainly a multifaceted field that deals with complex phenomena and requires a broad set of knowledge, including an understanding of the design and operations of industrial plants, physical phenomena, behavioral science, regulations, and statistical modeling. Furthermore, this research concerns the use of Machine Learning algorithms, increasing the complexity and introducing elements of data science and programming. Therefore, this Ph.D. study meets the requirements of interdisciplinary research (Pruzan, 2016)

5.3. Research approach

Research activities follow three primary approaches (Creswell, 2014):

- quantitative;
- qualitative;
- mixed.

These approaches, although distinct, can sometimes overlap. Quantitative and qualitative methods both involve using data to enhance understanding, with the former relying on numbers and the latter on words (Creswell, 2014). Actually, quantitative research focus on confirming relationships between measurable variables, often through statistical analysis. On the other hand, qualitative research interprets textual data gathered from interviews and observations. The mixed research approach combines both quantitative and qualitative methods. This synthesis aims to offer a thorough understanding of research problems, capitalizing on the strengths of each approach. By merging statistical rigor with interpretive insight, mixed research bridges the insights from pure quantitative and qualitative approaches (Johnson and Christensen, 2015).

Most of the research activities carried out in this Ph.D. follows a quantitative approach since they offer formal analyses supported by case studies. Only three articles (i.e., I, XII, and IV) follow mixed and qualitative approaches. Specifically, article I follows a mixed method since it couples narrative review (NR) with the

statistical findings of a systematic review (SR). Similarly, Article XII may be considered mixed since it provides a narrative discussion on foundational issues supported by data from a specific case study. On the contrary, Article IV follows a qualitative approach since it discusses how the methods described in Articles II and III may be used to support Risk-Based Inspection (RBI) methodologies.

5.4. Quality assurance

The Articles presented in Part II have either been published or are currently undergoing revision in peer-reviewed journals or conferences. This rigorous process ensures that the research meets stringent scientific standards and undergoes meticulous evaluation by an impartial team of subject matter experts. The constant guidance of the two supervisors has ensured the quality of results and methodologies, as they have not only provided crucial support but also delivered critical feedback that has refined the outcomes of this research. Furthermore, collaborative efforts with co-authors from other departments and institutions have substantially enhanced the significance and overall quality of the research. Additionally, all the proposed methodologies were rigorously tested through real-world case studies, affirming their quality and applicability.

6. Research method

This section provides a brief overview of the methods utilized during the Ph.D. study. Additional details on these methods can be found in the publications (Article I – Article XII).

6.1. Narrative and systematic review

Literature reviews play an important role in research, offering insights into previous work, avoiding redundant efforts, and highlighting potential areas for new research. Two main methodologies for conducting literature reviews are the Narrative Review (NR) and the Systematic Review (SR). Ferrari (2015) describes a Narrative Review as a summarization of previous literature, focusing on identifying existing research, preventing redundancy, and pinpointing uncharted study domains. Typically, an NR introduces the content, scope, and objectives, followed by a structured search of literature based on chosen criteria, and culminates in a comprehensive discussion and conclusion, similar to conventional scientific publications. In contrast, a Systematic Review adopts a more rigorous and reproducible approach, seeking to answer well-defined research questions through systematic identification, selection, and critical appraisal of pertinent studies. The PRISMA Statement (Preferred Reporting Items for Systematic Review and Meta-Analysis) (Liberati et al., 2009) describes a comprehensive framework for conducting a systematic review. The process involves four phases:

1. Identification of relevant articles through database screening.
2. Screening of the records' title and abstract to remove duplicates and unrelated articles
3. Eligibility assessment of screened records based on manual analysis of the full-text articles.
4. Inclusion and analysis of the eligible records.

This method mandates transparent documentation of every step, from the exact search queries to the chosen filters, databases, and exclusion criteria. Moreover, the systematic exploration of broad databases in SRs typically yields a substantial number of articles. Consequently, dedicated software and tools are regularly utilized to analyze these data.

While both methodologies aim to synthesize existing literature, the NR offers a broader, more descriptive perspective, whereas the SR is more structured, focused, and exhaustive in its approach. It must be noted that despite the thoroughness of SR, the possibility of omitting relevant articles always exists, emphasizing the need for meticulousness and acknowledging inherent limitations.

In Article I, a hybrid methodology is adopted, aiming to merge the advantages of both narrative and systematic reviews. More details on the specific methods and findings can be found in Section 7.1.

6.2. Data preparation

Data preparation encompasses a range of activities focused on gathering, cleaning, transforming, and manipulating data to facilitate subsequent analyses and ensure that data are well-suited for ML algorithms (Abdallah et al., 2017). It serves as the foundation upon which all subsequent analyses are built, significantly impacting the validity and reliability of the outcomes. Poorly prepared or preprocessed data can introduce noise, bias, or inaccuracies, ultimately skewing the results, affecting the model performance, and leading to potentially erroneous conclusions (Jain et al., 2020). Therefore, a meticulous approach to data preparation and preprocessing is not just advisable but essential for the integrity of any study making use of ML techniques.

Typically, data preparation involves the following activities:

1. Data gathering;
2. Feature selection;
3. Data cleaning;
4. Data transformation;
5. Data partitioning.

A brief overview of each activity is provided in the following paragraphs.

6.2.1. Data gathering

Data must be extracted from relevant data sources or created through simulations. Specifically, data may come from one of three sources:

a. Extracted from historical databases at the real plant: In this scenario, data are extracted directly from the plant storage systems. The type and amount of data to extract are generally guided by experience, although some guidelines may be found in the literature (Stanula et al., 2018). Constraints of the specific system under study must also be considered, as many plant components may have limitations on the volume of data that can be extracted. It is crucial to ensure that the data adequately represent the phenomena being studied. If the plant allows for real-world testing and experimentation, a meticulously planned Design of Experiments (DOE) should be implemented to maximize the information obtained while minimizing the data and experimental needs. Real alarm data from a plant for ammonia production were used in Articles VI, VII, and VIII, while process data from a waste incineration facility were used in Articles X and XI.

b. Extracted from digital repositories: In this case, data have already been sourced and made available online. The data might have undergone some level of preprocessing or manipulation, which could improve their quality and reduce the time needed for further preprocessing. However, these datasets are generally less customizable, limiting the ability to acquire additional or different types of data. Articles II, III, and XII deal with data extracted from digital accident databases.

c. Generated through computer simulations: Data can also be produced using computational methods such as CFD simulations, FEM, or process simulators. This option offers the highest degree of customizability, as it allows the researcher to explore different aspects of the studied phenomena. Nevertheless, the simulation setup should be carefully designed to maximize information yield while minimizing computational load. Simulated alarm data are used in Article IX, while data from a lumped-parameter model are used in Article V.

Regardless of the source, the data-gathering process usually results in a tabular-like dataset $\mathcal{D}$, where each row represents an individual observation and each column indicates a characteristic or feature of the studied phenomena. Depending on the data collection methods, the dataset may contain numerous redundant or irrelevant features that require removal, a topic covered in the subsequent section.

6.2.2. Feature selection

Feature selection is a crucial step in the machine learning pipeline, aimed at identifying the most relevant variables, or "features", for a given problem. This step is critical because choosing the right features can drastically improve model performance, while irrelevant, redundant, and noisy features can significantly affect the model performance (Wang et al., 2016). Techniques for feature selection often include statistical methods, like mutual information or chi-squared tests, as well as algorithmic approaches, such as recursive feature elimination and LASSO (Least Absolute Shrinkage and Selection Operator) (Li et al., 2016). These techniques help in removing redundant or uninformative features, thus simplifying the model, speeding up training, and improving interpretability. Feature selection is often a manual and somewhat experimental process (Witten et al., 2011). It typically involves trial and error, where different subsets of features are evaluated for their effectiveness in prediction, and the model is refined accordingly. This iterative nature makes feature selection both an art and a science, requiring a deep understanding of the domain, the data at hand, and the nuances of various selection techniques.

6.2.3. Data cleaning

Data cleaning is the process of detecting and rectifying corrupt or inaccurate records within a dataset (Chu, 2018). this practice addresses various data inconsistencies that could adversely affect the learning process, such as missing values, duplicated records, erroneous entries, and non-standardized representations of identical data (García et al., 2015). One common issue in data cleaning is the removal of duplicated or irrelevant observations. Duplicate records can distort statistical analyses and yield inaccurate representations of the data. Likewise, irrelevant observations can dilute the information content of the dataset and introduce noise into the ML models. Therefore, it is crucial to meticulously identify and remove these types of data to maintain the dataset integrity. Handling missing values is another critical aspect of data cleaning. The strategy for dealing with missing data depends on the type of data (e.g., numerical or categorical) and the nature of missing data (e.g., missing at random or not). In cases where missing values do

not significantly affect the dataset size or representativeness, entire observations may be removed. Likewise, if a feature is largely affected by missing values, it may be excluded from the dataset. As a general guideline, a missing rate lower than 15 % allows the elimination of missing instances without impacting the learning procedure (Strike et al., 2001). If the rate is higher, more advanced methods are recommended for managing the missing information (Acuña and Rodriguez, 2004). For instance, missing values could be replaced with a constant value, such as the most frequent feature value for categorical data or the mean for numerical data. More sophisticated techniques, like interpolation or imputation, can also be employed to estimate missing values based on other dataset features (Donders et al., 2006).

6.2.4. Data transformation

Data transformation involves altering and consolidating raw data to enhance the efficiency of mining processes and facilitate easier pattern recognition by ML algorithms (Han et al., 2012). The transformation serves not only to enhance data interpretability but also to adapt the data to the requirements of the chosen machine learning methodologies. For example, Neural Networks and many other ML algorithms require numerical input. If the dataset contains categorical features, these must be converted into a numerical format using techniques like one-hot and multi-hot encoding, label encoding, and hashing (Hancock and Khoshgoftaar, 2020).

Additionally, numerical data may be categorized or discretized to make data more interpretable or meaningful (Ramírez-Gallego et al., 2016). An example would be converting a continuous numerical attribute like time of day into discrete intervals (“8-12 A.M.") or descriptive categories (“Morning” or “First work shift”). Discretization offers several advantages, such as reducing computational complexity and potentially improving algorithmic efficiency (Witten et al., 2011). However, it must be noted that discretization can also lead to loss of information. Careful consideration must be given to the number and range of categories used, as poorly chosen bins could hide significant trends or patterns in the data.(Witten et al., 2011). However, it must be noted that discretization can also lead to loss of information. Careful consideration must be given to the number and range of categories used, as poorly chosen bins could hide significant trends or patterns in the data.

Normalization is another crucial step in data transformation, especially for algorithms sensitive to the scale of input features. Methods like min-max scaling and z-score normalization are used to standardize numerical attributes to fit within a specified range, thus preventing any single feature from disproportionately influencing the model’s performance. Normalization is particularly important for algorithms that utilize distance metrics, like nearest-neighbor classifiers, to ensure that features with a larger scale do not dominate those with a smaller scale. (Han et al., 2012). For example, if one feature is in the range of 0 to 1 and another is in the range of 0 to 1000, the latter feature could disproportionately affect the distance calculations, leading the algorithm to give it more importance than it may deserve. Also, feature scaling is essential for

algorithms that rely on gradient descent, like Neural Networks, because it facilitates convergence and speeds up the learning phase (Sola and Sevilla, 1997).

Other transformation techniques can be highly beneficial depending on the nature of the problem at hand, such as dimensionality reduction techniques like Principal Component Analysis (PCA) or ISOMAP (Vlachos, 2017), and noise reduction techniques (Xiong et al., 2006).

6.2.5. Data splitting

The dataset $\mathcal{D}$ needs to be split into a set of independent datasets in order to allow for a fair and unbiased evaluation of the model's performance and assess its ability to generalize over unseen data. Data splitting is predominantly done in supervised learning frameworks because the main objective is to train a model on one subset of the data (training set) and then evaluate its performance on another, unseen subset (testing set) using labeled outcomes. In unsupervised learning, where the primary tasks often include clustering, there are no predefined labels to predict or evaluate against, making data splitting for performance evaluation less applicable.

The simplest data-splitting method is the so-called Holdout technique (Raschka, 2018), which involves splitting the dataset into two mutually disjoint subsets $\mathcal{D}_{train}$ and $\mathcal{D}_{test}$

$$\mathcal{D} = \begin{vmatrix} \mathcal{D}_{train} \\ \mathcal{D}_{test} \end{vmatrix} \tag{1}$$

$\mathcal{D}_{train}$ is used to train the model, while $\mathcal{D}_{test}$ is utilized to evaluate its performance. Conventionally, $\mathcal{D}_{train}$ encompasses 80 % of the total observations, leaving the remaining 20 % for $\mathcal{D}_{test}$. Before the split, the observations (i.e., rows of $\mathcal{D}$) are shuffled to increase randomness and support an unbiased evaluation. The splitting procedure must be designed to prevent any information leakage from $\mathcal{D}_{train}$ to $\mathcal{D}_{test}$. For instance, when normalization is part of data preprocessing, it is essential to ensure that data are normalized after splitting, with $\mathcal{D}_{test}$ being normalized solely using the statistical parameters (e.g., min, max, mean, and variance) from $\mathcal{D}_{train}$. In other words, $\mathcal{D}_{test}$ should mimic an independent set of observations, resembling new data input during real-world applications.

Beyond the basic holdout approach, there are more sophisticated splitting methodologies that offer a deeper and more balanced model evaluation. One such method is the "k-fold cross-validation" (Hastie et al., 2009), which divides the dataset k mutually disjoint subsets. In this method, a single subset is used for testing, while the remaining k-1 subsets are employed for training. This process is replicated k times, each instance utilizing a different test subset in order to provide a more comprehensive evaluation of the model's performance.

6.3. Machine Learning

The following paragraphs describe the Machine Learning algorithms utilized throughout this Ph.D. project. Specifically, Section 6.3.1 describes classification and regression techniques, Section 6.3.2 focuses on clustering, and Section 6.3.3 illustrates Natural Language Processing.

6.3.1. Classification and regression

Classification and regression algorithms aim to build a mathematical model that approximates the relationship between the features of an observation (X_i) and its label (Y_i). This mathematical model may be approximated to a function (f) with learnable parameters (θ) that convert the features of an observation into its label:

$$Y_i \approx f(X_i, \theta) \tag{2}$$

The nature of the label depends on the specific problem at hand: Y_i is categorical in classification tasks, while in regression, Y_i is numerical. The internal structure of the function f, also called the "model", depends on the specific ML algorithm chosen.

Classification algorithms are utilized in Articles II, III, VI, VII, and VIII, while regression models are used in Articles V, X, and XII. Specifically, Articles II and III, aligning with Objective 2.1, propose the use of classification algorithms to predict the severity of accidents associated with hazardous substances. Articles VI, VII, and VIII, relevant to Objective 2.3, introduce classification models for monitoring and improving industrial alarm systems (see Sections 7.5 and 7.6 for more details). Article V, supporting Objective 2.2, describes a regression model to predict the Time-To-Failure of atmospheric tanks subjected to external fires (see Section 7.4). Finally, Articles X and XI, addressing Objective 2.3, describe how regression algorithms may be used to build data-driven process simulation models and support the evaluation of safety barriers in environmental-critical facilities (see Section 7.8).

6.3.1.1. Training and evaluation

The development of the ML model revolves around two phases: training and evaluation. In the training phase, the algorithm is provided with observations from $\mathcal{D}_{train}$, where each observation consists of a set of features (X_i) and an associated label (Y_i). The model adjusts its internal weights to map the features to their corresponding labels based on the provided training data. The tuning procedure aims to identify the optimal set of weights (θ^*) that minimizes the error between true labels (Y) and the predictions made by the model ($\hat{Y}$)

$$\theta^* = \underset{\theta}{argmin}\left[\ell\left(Y, f(X, \theta)\right)\right] \tag{3}$$

where $f(X, \theta) = \hat{Y}$ represents the model predictions, and ℓ indicates the loss function –i.e., a measure that quantifies the error between true and predicted labels. The choice of the most appropriate loss function depends on the specific problem at hand. A widely used loss function for regression tasks is the Mean Squared Error (MSE)

$$MSE = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i - \hat{Y}_i\right)^2 \tag{4}$$

where N indicates the number of observations included in the training dataset. Instead, the Cross-Entropy (CE) is often used for classification tasks

$$CE = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{M} y_{ij}\log(p_{ij}) \quad (5)$$

where M represents the number of classes, y_{ij} is a binary indicator that takes value 1 if observation i belongs to class j, and p_{ij} represents the model's predicted probability that observation i belongs to class j.

After training, the model must be evaluated on a new, independent set of observations to test its prediction and generalization capabilities. The aim of the evaluation phase is to ensure that the model offers accurate and robust predictions, confirming that it can generalize the lesson learned during training to new, unseen data, and identifying issues like overfitting and underfitting (Goodfellow et al., 2016). During the evaluation stage, the observations included in $\mathcal{D}_{test}$ are fed to the model, which predicts their labels according to the knowledge extracted from the training dataset. Predicted labels are compared to true labels to evaluate the model performance. A variety of performance metrics may be utilized to quantify the prediction capabilities. For example, some of the most widely used performance metrics for regression tasks include the MSE (Eq. 4), the Root Mean Squared Error (RMSE), and the coefficient of determination (R^2)

$$RMSE = \sqrt{MSE} \quad (6)$$

$$R^2 = 1 - \frac{\sum_{i=1}^{N}(Y_i - \hat{Y}_i)^2}{\sum_{i=1}^{N}(Y_i - \mu)^2} \quad (7)$$

where μ indicates the mean of the true labels.

Instead, the performance of a classifier is evaluated based on the number of correct and wrong predictions. For example, consider a binary classification problem where observations belong to two mutually exclusive classes, namely "1" and "0". When the model makes a prediction for a given observation, there are four possible outcomes:

- TP = True Positive –i.e., $\hat{Y}_i$ = 1, Y_i = 1;
- TN = True Negative –i.e., $\hat{Y}_i$ = 0, Y_i = 0;
- FP = False Positive –i.e., $\hat{Y}_i$ = 1, Y_i = 0;
- FN = False Negative –i.e., $\hat{Y}_i$ = 0, Y_i = 1.

The sum of True Positives and True Negatives yields the number of correct predictions, whereas the sum of False Positives and False Negatives provides the number of incorrect predictions. True Positives, True Negatives, False Positives, and False Negatives are typically condensed into more meaningful metrics:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (8)$$

$$Precision = \frac{TP}{TP + FP} \quad (9)$$

$$Recall = \frac{TP}{TP + FN} \quad (10)$$

The choice of the most appropriate metrics largely depends on the problem under assessment. Different metrics offer various perspectives, and multiple metrics might be considered depending on the specific context or objectives. For example, suppose the problem involves the prediction of rare events (classes). In that case, the accuracy might not truly reflect the model's performance since a high accuracy can be obtained by always predicting the most frequent class. Here, precision and recall should be considered. Likewise, if the classes have unequal misclassification costs, meaning that misclassifying observations that belong to class "1" has more severe consequences than misclassifying observations of the other class, more emphasis should be put on the recall. Often, Precision and Recall are considered together in the so-called F-score measure (F_β)

$$F_\beta = (1+\beta^2)\frac{Precision \cdot Recall}{(\beta^2 \cdot Precision) + Recall} \tag{11}$$

where $\beta \in \mathbb{R}^+$ is a coefficient that determines the tilt toward precision or recall. If $\beta > 1$, the metric is recall-oriented, meaning that the Recall is considered to be β times more important than Precision. On the contrary, if $\beta < 1$, the metric is precision oriented. Also, it is worth mentioning that specific training and evaluation strategies are available to address class imbalance (Hasib et al., 2020) and cost-sensitive classification (Fernández et al., 2018).

6.3.1.2. Models

With the rapid evolution of machine learning, there has been a proliferation of models for classification and regression. The model defines the mathematical structure of the function f in Eq. (2) and the specific optimization procedure in Eq. (3). There are two main categories of models: parametric and non-parametric. Parametric models make an assumption on the functional form of f (e.g., linear, exponential) and have a fixed number of internal parameters (i.e., weights). Instead, non-parametric models do not make any assumption on the form of f, and their number of parameters is flexible and dependent on the number of training samples.

Three parametric models have been used and applied to address different problems in this Ph.D. project: linear models, Neural Networks (NNs), and Wide&Deep models, which are hybrid models comprising a linear part and a Neural Network part.

Linear models represent the labels as a linear combination of the features:

$$Y_i = \beta_0 + \sum_{j=1}^{K} x_j \cdot \beta_j \tag{12}$$

Where $K \in \mathbb{N}$ represents the total number of features, β_0 and β_j indicate the model weights, and x_j represents the jth feature of the ith observation. Linear models are considered one of the simplest models in ML, and they are often used as a baseline to compare and evaluate the performance of more complex models. However, although simple, linear models are still widely used because they are fast, robust, interpretable, and perform well on large datasets (Brink et al., 2016).

Neural Networks are directed acyclic graphs that describe the relationship between features and labels through a set of nonlinear transformations of linear combinations. They can be described as a series of interconnected layers, where each layer contains multiple nodes or "neurons", as shown in Figure 4.

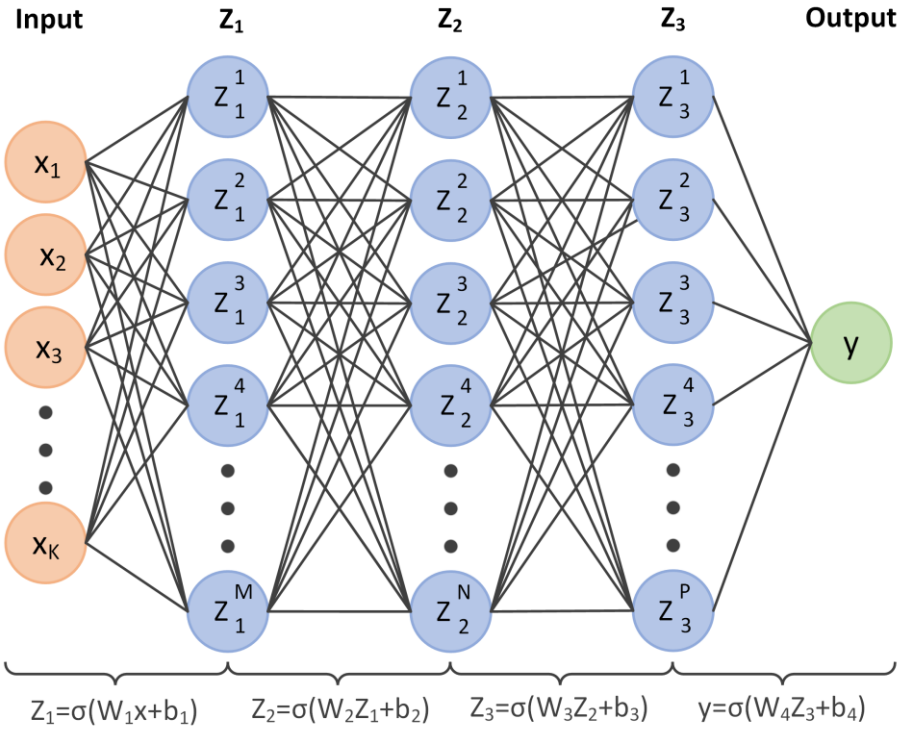


Figure 4. Graphical representation of a Neural Network with K inputs (orange), three hidden layers (blue), and one output (green). Z_i^j indicates the j-th neuron in the i-th hidden layer, x_i represents the i-th input, and y indicates the output. Calculations are shown on the bottom, where σ represents a nonlinear function, W_i indicates the matrix of the weights of the i-th layer, and b_i represents the vector of the biases of the i-th layer.

The first layer (orange in Figure 4), also called the input layer, receives the raw input data, with each neuron representing a single feature of an observation. After the input layer, there are one or more hidden layers (blue in Figure 4). Neurons in hidden layers are real-valued entities that are calculated through a nonlinear transformation of the linearly combined units in the previous layers. Specifically, a generic layer $Z_i \in \mathbb{R}^{M\times 1}$ with M neurons is computed as follows:

$$Z_i = \sigma(W_i \cdot Z_{i-1} + b_i) \tag{13}$$

where σ is a nonlinear function, $W_i \in \mathbb{R}^{N\times M}$ is the matrix of the weights, $Z_{i-1} \in \mathbb{R}^{N\times 1}$ is the layer preceding Z_i, and $b_i \in \mathbb{R}^{M\times 1}$ is the vector of the biases. Finally, the output layer returns the model predictions. Neural networks excel at modeling complex, non-linear relationships. Their ability to generalize to new data and automatically extract relevant features from raw input sets them apart from many traditional algorithms. However, their intricate structures render them "black boxes," obscuring their decision-making processes and raising concerns about interpretability. Moreover, they demand significant computational resources, especially during training, which might necessitate specialized hardware. Also, Neural Networks are prone to overfitting and typically require large amounts of labeled data to perform optimally.

Wide&Deep models feature a linear part and a Neural Network part, as shown in Figure 5.

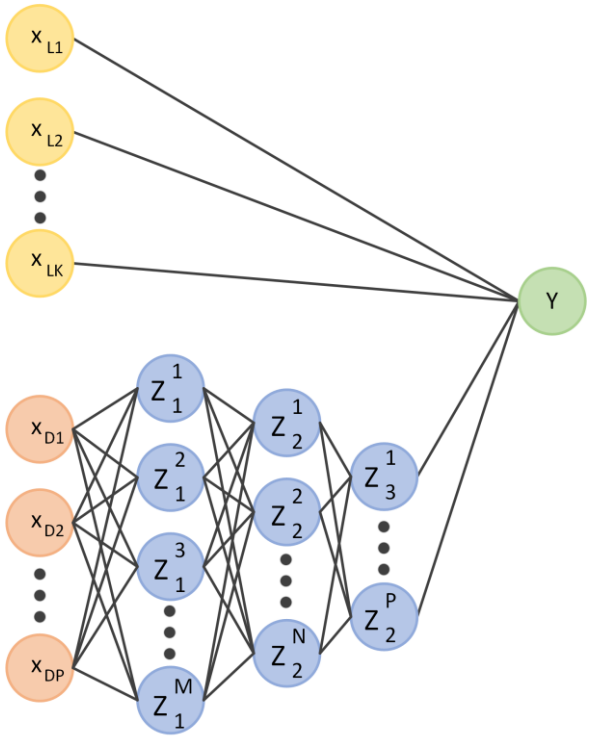


Figure 5. Illustration of the Wide&Deep model. x_{Li} indicates the i-th input the linear part of the model (yellow), while x_{Dj} represents the j-th input of the Neural Network part (blue). Adapted from (Cheng et al., 2016).

During the training phase, linear and NN parts are jointly trained, meaning that the weights of the models are optimized simultaneously (Cheng et al., 2016). This hybrid model has proven to combine the advantages of the linear and Neural Network models, minimizing their drawbacks and enhancing their qualities.

6.3.2. Clustering

Clustering algorithms aim to group observations into clusters in such a way that observations in the same clusters are similar to each other and different from the observations included in different clusters (Shultz et al., 2011). Classification and clustering both involve grouping observations. However, in classification, observations are provided with their true labels, and the goal is to learn the relationship between features and labels; it is a supervised technique with a clear map of how data should be categorized. Clustering, however, ventures into the territory of unsupervised learning. Without any knowledge of the true labels associated with observations, clustering algorithms explore the dataset, identifying patterns and groups based on inherent similarities. In essence, while classification operates with established categories, clustering discovers potential categories within the data itself. However, it is worth mentioning that some clustering methodologies have been adapted to utilize labeled examples (Qin et al., 2019). These approaches are not covered in this section since they were not used during this Ph.D. research.

The development of a clustering algorithm is different from their supervised counterparts. The concepts of training and evaluation persist, but they manifest differently compared to classification tasks. Firstly, in clustering there is no requirement to split the data into training and evaluation sets, since the primary focus is on understanding the existing data, and there is no information regarding the true labels. Therefore, the

entire dataset $\mathcal{D}$ is used to train the model, which identifies patterns and groupings without explicit external guidance or predefined labels. Secondly, the evaluation phase typically involves assessing the quality of the learned clusters using various metrics or visual examinations rather than measuring the accuracy against known ground truth labels, as is done in classification.

The evaluation of a clustering relies on metrics that quantify the "quality" of a cluster without any apriori information. Therefore, the process is inherently more challenging than its supervised counterpart (Palacio-Niño and Berzal, 2019). One widely used quality metric is the Silhouette index (S) (Rousseeuw, 1987), which estimates how close each point in one cluster is to the points in the neighboring clusters. The Silhouette index is computed for each data point and is given by the formula

$$S(i) = \frac{b(i) - a(i)}{\max\{a(i), b(i)\}} \tag{14}$$

where $a(i)$ is the average distance between the ith observation in a cluster and the other observations in the same cluster, and $b(i)$ is the smallest average distance between the ith observation in a cluster and an observation in a different cluster. The Silhouette index ranges between -1 and 1. A large value indicates that an observation aligns well with its assigned cluster and is distinctly separate from other clusters. Further, the average silhouette index provides insight into the overall quality of the clustering configuration. Specifically, an average silhouette value near 1 suggests optimal clustering. A value nearing 0 indicates potential overlap between clusters, while a value close to -1 implies that observations may have been grouped incorrectly.

The realm of data clustering is rich with a variety of algorithms. The K-means clustering algorithm is among the most widely employed clustering methods. This model was used in Article XII to group countries based on their similarities in natural disaster exposure (see Section 7.9 for more details). It evaluates the similarity between observations based on their distance to the nearest centroid (Mannor et al., 2011). Prior to training, the number $K \in \mathbb{N}^+$ of centroids (i.e., clusters) must be specified by the user. During training, the temporary position of centroids is randomly initialized, and observations are assigned to the nearest centroid. Consequently, a cluster consists of observations that are closest to a particular centroid. Subsequently, the algorithm follows these main steps:

1. For each of the k clusters, compute the new centroid coordinates, which are obtained as the mean of the features of the observations assigned to a cluster.
2. Reassign observations to the nearest centroid, leading to potential changes in cluster memberships.

This process is repeated until either (i) no observation shifts from one cluster to another after step 2 or (ii) the number of iterations exceeds a user-defined threshold. In other words, K-means clustering aims to group data by minimizing the within-cluster-sum-of-squares, which represents the distance between each data point and the cluster centroid. Often, the Euclidean distance is used to measure the distance between observations; therefore, the problem of finding the best clustering configuration $\{C_1, \cdots, C_K\}$ may be formulated as

$$\underset{\{C_1,\cdots,C_K\}}{minimize}\left(\sum_{k=1}^{K}\frac{1}{|C_k|}\sum_{i,i'\in C_k}\sum_{j=1}^{p}\left(x_{i,j}-x_{i'j}\right)^2\right) \tag{15}$$

where $k \in \{1,2,\cdots,K\}$, C_k indicates the observations included in the kth cluster, $|C_k|$ represents the total number of observations in C_k, i and i' indicate two distinct observations in C_k, and $p \in \mathbb{N}^+$ is the number of features of an observation (James et al., 2021).

K-means is a popular, efficient, and interpretable clustering algorithm. However, it is not robust to outliers and noise. Additionally, the results of the clustering procedure are affected by the initialization of centroids, thus leading to potential convergence to local minima rather than a global solution. Also, one of the main disadvantages of K-means is that the number of clusters needs to be specified in advance. Determining an appropriate value for K is not trivial and often necessitates iterative exploration. For example, the clustering procedure may be repeated N times with an increasing number of clusters (e.g., from 2 to 10). Subsequently, the optimal number of clusters might be identified by examining the average Silhouette index, seeking the number of clusters that maximizes this metric (Rousseeuw, 1987). An alternative evaluative approach involves plotting the number of clusters against their associated within-cluster sum of squares. Here, the point at which the rate of decline attenuates, often referred to as the "elbow", can provide an indication of the optimal number of clusters (Shi et al., 2021).

6.3.3. Natural Language Processing

Natural Language Processing (NLP) refers to the broad set of methods and algorithms that aims to learn, understand, and produce human language (Hirschberg and Manning, 2015). The capabilities of such models have tremendously increased during the last few decades thanks to the advancements in Deep Learning and computational capabilities (Otter et al., 2021). Much progress has been made from the first neural network for Next Word Prediction and vectorial representation of words proposed by Bengio et al. (2003) to the ground-breaking OpenAI's GPT-4 (OpenAI, 2023). Today, NLP models have become ubiquitous in our lives in many ways, such as text autocompletion (Chen et al., 2019), translation engines (Wang et al., 2022), fraud and fake news detection (Chen et al., 2017; de Oliveira et al., 2021), sentiment analysis (Solangi et al., 2018), and chatbots (Caldarini et al., 2022).

NLP algorithms were used in Article IX, specifically addressing Objective 2.3, to detect hidden correlations between alarms and perform online identification of alarm floods. More details on the specific application are provided in Section 7.7, while the following focuses on the algorithm used for the analysis, namely word2vec.

Word2vec (Mikolov et al., 2013) belongs to the branch of NLP that studies methods to represent words as vectors of real numbers in a high-dimensional space. This task, also called word representation and embedding, aims to represent words and their contextual relationships in a machine-understandable manner. In fact, as discussed in Section 6.2.4, most ML models cannot process categorical data (i.e., words), which must be converted into numerical features prior to the analysis. Simple techniques such as one-hot

encoding may work well with tabular data, where features typically assume a relatively limited number of unique categorical values. However, human language is highly articulated, the vocabulary comprises thousands of words, leading to a sparse and inefficient representation. Also, one-hot encoding and other simple encoding techniques cannot represent semantic information, which is a substantial limitation when dealing with human language. Therefore, the dynamic and rich nature of natural language necessitates more sophisticated methods, like word2vec, to capture the depth and breadth of linguistic relationships.

Word2vec takes a large corpus of text as input and learns to predict the context in which words occur based on their co-occurrence patterns. It offers two variants: Continuous Bag-of-Words (CBOW) and Skip-Gram. The CBOW model predicts the target word given the surrounding context words, while the Skip-Gram model predicts the context words given the target word. In Article IX, the Skip-gram architecture was used. However, before describing the model functioning, it is worth clarifying the meaning of "target" and "context" words.

In general, the context of a word in a sentence may be described by the words preceding and succeeding it. Considering that alarm floods are ordered sequences of alarms, and alarms are presented to operators as words (i.e., LI201.LL to indicate a low-level alarm triggered by the level indicator 201), we assume that the context of an alarm in an alarm flood sequence may be defined by the alarms preceding and succeeding it. Specifically, consider an alarm flood $\mathcal{F}$ made of N alarms, $\mathcal{F} =< a_1, a_2, \cdots, a_N >$, where a_i indicates the ith alarm, and let $\omega \in \mathbb{N}^+$be a user-defined variable. The "context" alarms of the target alarm a_i are the ω alarms that precede and follow a_i, as shown in Figure 6.

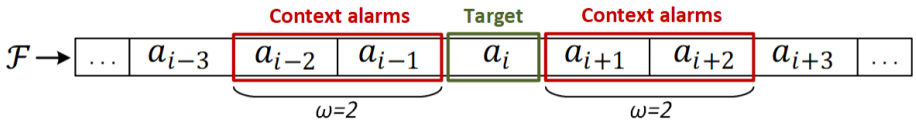


Figure 6. An example of target and context alarms in a flood. Here, a_i represents the ith alarm, $\mathcal{F}$ represents the incoming alarm flood, and ω is the user-defined parameter to determine the number of context alarms. Adapted from (Tamascelli et al., 2023)

The Skip-Gram architecture of the word2vec model tries to predict the context alarms (red in Figure 6) that appear in a given window around a target alarm (green in Figure 6). The structure of Skip-Gram is shown in Figure 7.

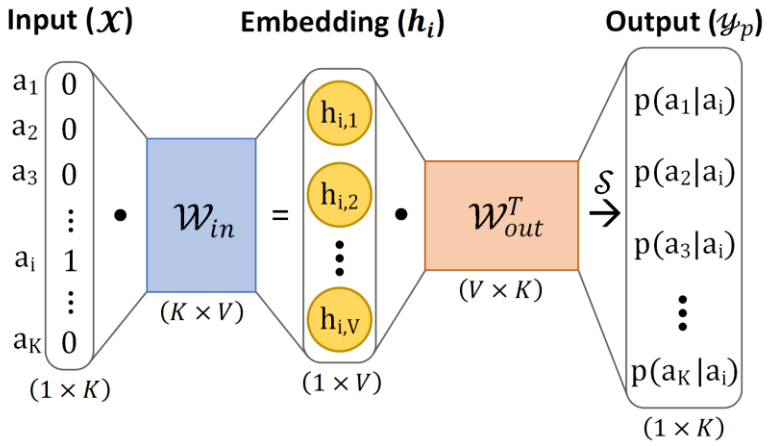


Figure 7. Illustration of Skip-gram model, where the input layer is a K dimensional one-hot encoded representation of a_i, the embedding layer h_i is the V dimensional representation of a_i, and the output layer is the conditional probability $p(a_k|a_i)$, $k = 1, \cdots, K$. $\mathcal{W}_{in}$ and $\mathcal{W}_{out}$ are the internal weights of the model. $\mathcal{S}$ is the softmax transformation function.

The model is a single-layer neural network that transforms a word (i.e., an alarm) (Input layer in Figure 7) into a vector of reals of dimension $V \in \mathbb{N}^+$ (Embedding layer in Figure 7) and returns the conditional probability of each alarm in the vocabulary being a context alarm (Output layer in Figure 7). Here, the vocabulary contains all the unique alarms configured in the plant. Specifically, the model takes as an input the one hot-encoded representation ($\mathcal{X}$ in Figure 7) of an alarm, say a_i in Figure 7, and performs the following calculations

$$h_i = \mathcal{X} \cdot \mathcal{W}_{in} \tag{11}$$

$$\mathcal{y} = h_i \cdot \mathcal{W}_{out}^T \tag{12}$$

$$\mathcal{y}_p = Softmax(\mathcal{y}) = \begin{vmatrix} p(a_1|a_i) \\ p(a_2|a_i) \\ \vdots \\ p(a_K|a_i) \end{vmatrix} \tag{13}$$

Where h_i represents the embedding layer, $\mathcal{X}$ indicates the one-hot encoding representation of a_i, $\mathcal{W}_{in}$ and $\mathcal{W}_{out}$ are $K \times V$ matrices of the model's internal weights, and $\mathcal{y}_p$ represents the model output. Here, $K \in \mathbb{N}^+$ indicates the number of alarms in the vocabulary, and V is a user-defined parameter that determines the size of the word embedding. Each row of $\mathcal{W}_{in}$ contains word embedding of a specific alarm, whereas rows of $\mathcal{W}_{out}$ represent the contextual relationships between alarms. Together, $\mathcal{W}_{in}$ and $\mathcal{W}_{out}$ represent the model's internal weights θ

$$\theta = [\mathcal{W}_{in}, \mathcal{W}_{out}] \tag{14}$$

During the training process, the model iterates through alarms in an alarm flood, and θ is updated iteratively using backpropagation and stochastic gradient descent to minimize the negative log-likelihood of the observed context alarms given the target alarm. That is

$$\hat{\theta} = \underset{\theta}{argmin}\left(-log \prod_c p(a_c|a_i)\right) \tag{15}$$

Where $\hat{\theta}$ indicates the updated model weights and c represents the index of the true context alarm of a_i. In other words, referring to the example shown in Figure 6, $c = \{i-2, i-1, i+1, i+2\}$.

After training, the vectors learned by word2vec capture semantic and syntactic similarities between alarms. Furthermore, the model can be used to predict the most probable contextual alarms given a target alarm as described (Eq. 13). These interesting properties have been leveraged to develop an alarm flood classification framework, where an ensemble of word2vec models is trained to learn contextual similarities between alarms generated by different fault conditions, and eventually used to identify the root cause of new alarm floods.

7. Contributions

This section provides a concise overview of each Article, encapsulating key insights and principal findings. The connection between the Articles and the objectives outlined in Section 4 and represented applying a top-down approach in Figure 3 is represented using a bottom-up approach in Figure 8, in order to better evidence the contribution of each single publication to the objectives outlined above. In the following paragraphs, articles are presented and discussed according to their topic.

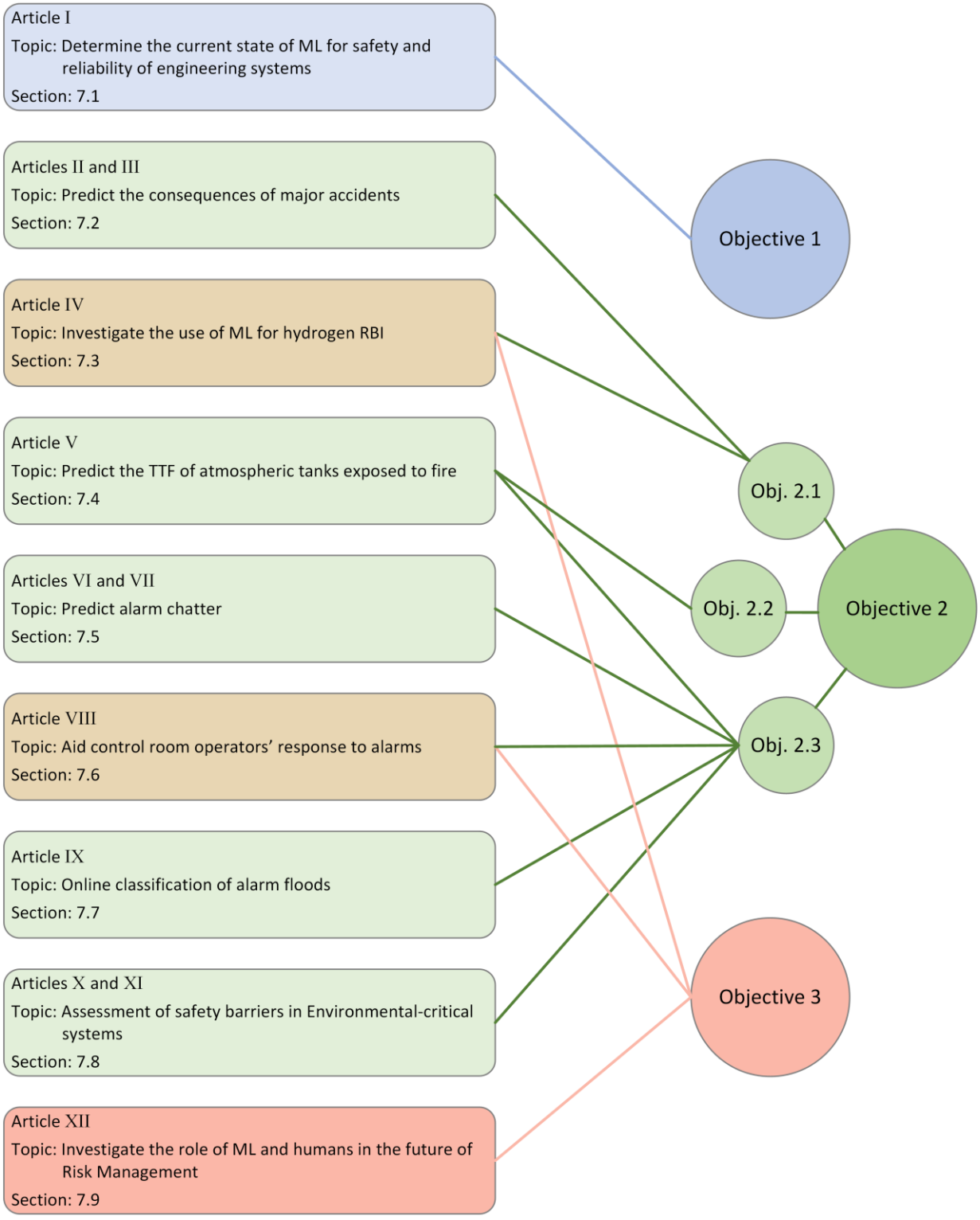


Figure 8. Bottom-up approach to the link with research objectives of the Articles published within the Ph.D. study (reported in Part II). Articles are grouped into nine contributions according to their topic.

7.1. Article I: determine the current state of ML for safety and reliability of engineering systems

This article addresses the first objective by offering a review of the existing literature concerning the utilization of ML techniques to enhance the safety and reliability of engineered systems. The analysis

integrates a narrative review with statistical findings from a systematic review. The primary goal of this study is to give a comprehensive overview of the subject, elucidating trends, gaps, and potential opportunities for advancement. The examination delves into ML methods applied to diverse safety and reliability topics, encompassing fault detection and diagnosis, anomaly detection, system prognosis, reliability analysis, and risk assessment. By extracting and filtering relevant manuscripts from the Web of Science Core Collection database (Clarivate, 2022), a total of 273 pertinent articles were identified and analyzed. The narrative review provides comprehensive insights into each topic, including methodologies, challenges, and limitations. In parallel, the systematic review offers a statistical-oriented perspective, shedding light on the distribution of ML categories, tasks, data types, preprocessing methods, and algorithms.

This hybrid approach, combining narrative and systematic analyses, permits to take advantage of the benefits of the two approaches. The narrative section provides readers with an in-depth overview of existing ML techniques for addressing the selected topics, guiding them toward the most relevant methods. In contrast, the systematic part supplies a statistical overview of current trends, providing interesting insights into the current landscape, highlighting gaps, and offering valuable perspectives that could shape the future role of ML in the realm of safety and reliability.

Some of the most relevant findings of this study are as follows:

- Supervised Learning is the most utilized approach for safety and reliability, primarily due to its ease of implementation and interpretability.
- Unsupervised and semi-supervised approaches, although perceived as more challenging, have the potential to address complex dynamics and rare event patterns.
- Most papers focus on fault detection, diagnosis, and location, followed by anomaly detection. Less emphasis has been placed on Risk Assessment, system prognosis, and reliability analysis.
- Different ML categories are favored for different domains. For instance, supervised classification is prevalent in fault detection, while unsupervised and semi-supervised methods are prominent in anomaly detection and risk assessment.
- Support Vector Machines, Neural Networks and Decision Trees are the most commonly used models. Decision Trees offer interpretability, while more complex models like Neural Networks tend to achieve better performance at the expense of being less interpretable.
- Industrial, experimental, and simulated data are the most common, reflecting the increasing digitalization of industrial processes and IoT technologies.
- Regulatory frameworks are evolving, with the European Commission proposing regulations on AI in safety-critical applications. Ensuring regulations keep pace with technology advancements is crucial.
- The availability and quality of data play a pivotal role in accurate predictions and model reliability.

Overall. The results confirm the growing prominence of ML in safety and reliability research, showcasing its potential to enhance safety and reliability in high-risk applications.

7.2. Articles II and III: predict the consequences of major accidents

Articles II and III align with Objective 2, specifically addressing Objective 2.1, which concerns the development of ML methods to estimate the consequences of major accidents. Traditional methods for consequence estimation are ill-suited for dynamic frameworks as they require specialized models, such as emission, dispersion, fire, and explosion models, often requiring computationally intensive techniques (e.g., Computational Fluid Dynamics and Finite Element Analysis) for accurate estimates. In this context, articles II and III investigate the use of classification models (see Section 6.3.1) to (i) extract information from accident databases and (ii) use the acquired knowledge to predict the severity of major accidents in terms of number of fatalities and injuries. These contributions aim to showcase the potential of ML models in utilizing historical accident data and develop intuitive and computationally inexpensive models for consequence estimation. Article II establishes the foundational principles for applying ML to this task, while Article III takes a step forward by considering the issue of Transfer Learning (Pan and Yang, 2010), illustrating how knowledge extracted from a broad accident database can be harnessed to forecast the outcomes of technology-specific accidents.

Article II introduces a generic framework for the development of ML models for the categorization of accidental events based on the number of people involved. The models are trained on historical accident data to learn the relationship between accident characteristics and accident consequences. The algorithms take as an input a set of intuitive accident features (e.g., the type and quantity of substance released, the accident type, the cause and origin of the release, the population density) and output the number of people involved in the accident. Five consequence categories (Table 4) are considered to reflect severity categories used by risk matrices and other risk analysis methods (ARAMIS project team, 2004). A One-vs-Rest approach is adopted.

Table 4. Accident consequence categories.

Severity Category	Description
NO	no killed/injured
1 – 10	from 1 to 10 killed/injured
10 – 100	from 10 to 100 killed/injured
100 – 1000	from 100 to 1000 killed/injured
> 1000	more than 1000 killed/injured

The article comprehensively encompasses all essential methodological phases, ranging from the extraction and preprocessing of data to the model evaluation. Also, hyperparameter tuning is discussed. The proposed approach is rigorously tested through a practical case study, utilizing accident data from the Major Incident Data Source (MHIDAS) database (AEA Technology, 1999). This contribution introduces innovative and valuable insights that hold potential in the creation of dynamic tools that may be employed to (i) assess the severity of different accident scenarios based on a set of readily available features, (ii) discriminate between

different severity levels and direct efforts to prevent/mitigate high criticality scenarios, and (iii) estimate the consequences of new accident scenarios without relying on computation-intensive techniques and detailed modeling.

Article III builds upon the foundation established in Article II, expanding its horizons to assess the potential of ML algorithms in transferring knowledge extracted from a broad accident database to predict the consequences of accidents involving specific substances. This study delves into the realm of Meta-Learning, often referred to as "learning to learn", which involves techniques aimed at emulating human capacity for generalizing and recalling prior experiences to enhance the efficiency of learning new tasks (Vanschoren, 2018). While humans inherently possess transfer learning abilities, drawing on past experiences to tackle novel challenges, ML algorithms do not exhibit the same proficiency in this area (Pan and Yang, 2010). Thus, this research investigates Transfer Learning capabilities in extending ML applications for learning from historical accident data. A novel methodology is introduced, leveraging insights from general accident databases to forecast outcomes in technology-specific accident scenarios. The efficacy of the approach is evaluated using the Major Hazard Incident Data Service (MHIDAS) database for initial learning, while a tailored database capturing accidents within ammonia production is utilized for evaluating the Transfer Learning capabilities. In addition to expediting the development of consequence prediction models by minimizing data requirements, this approach enhances the generalization capabilities of machine learning algorithms.

7.3. Article IV: investigate the use of ML for hydrogen Risk Based Inspection

Article IV investigates the potentiality of ML methods to aid Risk Based Inspection (RBI) of equipment working in pure hydrogen environments. Specifically, it focuses on ML methods to estimate the consequences of accidents involving hydrogen, leveraging the findings of Articles II and III, and proposing a practical application of the methods described in those studies. Similar to Articles II and III, this contribution addresses Objective 2.1. In addition, by providing ML tools to aid maintenance activities, this contribution partially addresses Objective 3, which concerns the interplay between ML and human activities, exploring potential and limitations of utilizing ML techniques to support risk-based decision-making.

The study focuses on addressing safety challenges associated with the growing adoption of hydrogen technologies. The peculiar thermophysical characteristics of hydrogen and its potential use in densely populated areas pose new challenges to ensure safety through the entire hydrogen value chain. In this context, the significance of inspection and maintenance activities cannot be overstated, as they play a pivotal role in safeguarding the physical integrity of equipment functioning with pure hydrogen. These efforts are vital to avoiding potential catastrophic consequences arising from hydrogen releases.

Within the RBI framework, determining the Consequence of Failure (CoF) associated with specific equipment is essential for estimating risk levels and, subsequently, prioritizing inspection and maintenance activities

toward the most critical apparatuses. Existing methods for the estimation of the CoF require the incorporation of numerous input parameters or assumptions tailored to each distinct case. Article IV describes a simplified approach, where ML is used to take advantage of the historical accident data and provide a reliable estimate of the CoF without the need for intricate methodologies. The discussion maintains a qualitative outlook, highlighting the benefit of ML techniques to aid RBI but also stressing limitations, such as the lack of data regarding hydrogen accidents. In fact, while accident databases focusing on hydrogen have recently been developed (Wen et al., 2022), there still appears to be a significant data shortage, prompting the exploration of data from diverse sources and the potential application of Transfer Learning techniques.

7.4. Article V: predict the Time-To-Failure of atmospheric tanks exposed to external fire

Article V aligns with Objective 2 and specifically addresses Objective 2.2, which focuses on the development of ML methods to estimate the frequency of undesired events. Specifically, this contribution proposes ML-based tools for estimating the Time-To-Failure (TTF) of atmospheric tanks exposed to external fire.

Atmospheric tanks, commonly used for storing flammable liquids, present a significant risk due to the large quantities of hazardous substances involved. As industrial facilities grow more complex and densely packed, the risk of fire-triggered domino scenarios increases, emphasizing the need for effective escalation probability quantification. Established methodologies for the estimation of the escalation probability take as input the TTF of the unit targeted by fire (Cozzani et al., 2005). Unfortunately, the quantification of the TTF through rigorous modeling (i.e., by CFD and FEM simulations) is often excluded from risk assessment studies due to the need for highly specialized knowledge and computational resources. Empirical correlations have been proposed for faster calculation, enabling the Dynamic Risk Analysis of domino scenarios triggered by fire. However, such correlations suffer several issues, such as their restrictive assumptions and inability to consider the effect of safety barriers (e.g., deluge systems).

Article V addresses these limitations by introducing ML models that can (i) provide an accurate and fast estimate of the TTF of atmospheric tanks exposed to external fire, and (ii) consider the effect of mitigation barriers. To this end, a lumped parameter model called RADMOD (Landucci et al., 2009) was used to simulate a large number of fire scenarios, encapsulating diverse atmospheric tanks subject to various fire conditions. The resulting failure data were used to train a Neural-Network model for the prediction of the TTF. The proposed ML model takes as an input the tank geometry (i.e., shell thickness, tank height, diameter, filling level), the fire characteristic (i.e., the total heat flux targeting the tank), and the characteristics of the safety measures (i.e., the barrier activation time and effectiveness) and outputs the TTF. Hyperparameter tuning strategies were implemented to ensure optimal performance. In addition, a model-agnostic method for the estimation of confidence intervals was implemented to enrich the model output and enable better-informed decision-making. The model achieved an RMSE equal to 1.66s for unmitigated scenarios (mean TTF = 269s),

and 71s for mitigated scenarios (mean TTF = 645s). The results were compared with existing simplified correlations for the calculation of the TTF (Landucci et al., 2009; Yang et al., 2023), proving the superior performance of the proposed NN model.

7.5. Articles VI and VII: predict alarm chatter.

Articles VI and VII align with Objective 2 and specifically address Objective 2.3, which focuses on the development of ML methods to monitor, evaluate, and improve the performance of safety barriers. Specifically, this contribution explores the use of ML to develop proactive tools for improving the performance of industrial alarm systems.

As depicted in Figure 2, industrial alarm systems are one of the first layers of protection in preventing process deviations from escalating into hazardous events. Ideally, alarms should inform control room operators about dangerous deviations from normal operating conditions. However, improper alarm design may severely impact the efficacy of the alarm system (Izadi et al., 2009), leading to periods of intense alarm activity, also called alarm floods. During a flood episode, a large number of alarms are triggered in a short time span, impeding control room operators from identifying the causes of the abnormality and providing adequate response. Many of the alarms during such flood events are known as “chattering alarms” – i.e., alarms that rapidly transition between active and not active state in a short period of time (ANSI/ISA, 2016). Therefore, detecting and removing alarm chatter is paramount to decrease the number and severity of flood episodes.

Chattering alarms may be defined as alarms that produce three or more records in one minute (Kondaveeti et al., 2010). They pose a significant nuisance to control room operators by significantly inflating the alarm count. Currently, chattering alarms are only addressed and removed retrospectively (e.g., during periodic audits). Hence, a proactive method that predicts future chattering based on past and current process conditions could greatly enhance alarm system performance. This proactive approach would empower control room operators with predictive insights, facilitating preventive measures and enabling real-time monitoring of the alarm system, as opposed to the conventional reactive approach that addresses alarm chatter only after it occurs.

In this context, Articles VI and VII investigate how classification models can be used to predict future alarm chatter. While Article VI outlines a preliminary methodology, Article VII offers a comprehensive analysis by exploring diverse algorithms and delving deeper into the results. The approach described in these studies involves training ML classification models on historical alarm data in order to identify whether an alarm will exhibit chattering within the next hour. The algorithms take the characteristic of an alarm as an input (e.g., the instrument that triggered the alarm, the time of activation, the type of alarm) and output a binary label (Y_i in Eq. (2)) assigning a value of 1 if the alarm is going to show chattering within the next hour, and 0 otherwise. Historical alarm data are used to train the model, learning how present and past process conditions affect alarm behavior. The methodology has been demonstrated in a real case-study, taking

advantage of historical alarm data from an ammonia production plant. Three classification algorithms have been tested and compared: logistic regression, Feed Forward Neural Network, and a hybrid model incorporating both linear and Neural Network components. The results indicate that while all models exhibit remarkable performance, the Logistic Regression algorithm outperforms the others, offering accuracy, precision, and recall larger than 0.93.

7.6. Article VIII: support control room operators' response to alarms

Article VIII explores the use of ML methods to support control room operators and improve the performance of industrial alarm systems. Specifically, it focuses on ML methods to predict the effectiveness of operators' actions following process alarms. Similar to Articles VI and VII, this contribution addresses Objective 2.3, as it proposes ML-based approaches to improve the performance of industrial alarm systems. Contrary to Articles VI and VII, which focus on technical solutions to identify future alarm chatter, Article VIII explores how ML may be used to support and guide the operators by providing live feedback on the efficacy of their actions. Therefore, Article VIII also addresses Objective 3, which concerns the interplay between ML and human activities, exploring potential and limitations of utilizing ML techniques to support risk-based decision-making.

In industrial alarm system, alarms are often configured with different levels of criticality. For instance, a process variable may be associated with a low-level alarm (LL), a very-low-level alarm (LTRP), a high-level alarm (HH), and a very-high-level alarm (HTRP). When a low criticality alarm is triggered (i.e., LL and HH) the operators should acknowledge the alarm, diagnose the situation, and eventually take corrective actions in order to restore normal operations. However, if corrective measures are not adequate to tackle the issue, the deviation may worsen, triggering critical alarms (i.e., LTRP or HTRP), leading to a plant shutdown or more serious consequences. In this context, Article VIII proposes an ML-based approach to predict whether a critical alarm will reoccur within 30 minutes after a low-criticality alarm is acknowledged the control room operators. Specifically, when an operator acknowledges an alarm, the ML algorithm takes as an input the features of the alarm being acknowledged, such as the instrumentation that triggered the alarm, the value of the process variable, and the timestamp, and returns a binary label that takes the value 1 if the algorithm predicts that a critical alarm will occur within 30 minutes, or 0 otherwise. Therefore, the ML model provides live guidance on the effectiveness of the operation actions, informing that different (\more effective) measures are required to resolve the situation. The model, specifically a Wide&Deep classification algorithm, was trained on a real industrial database. Threshold tuning was performed to increase the recall, which is the most meaningful metric considering the criticality of producing false negatives. Obtained recall and precision are 0.9 and 0.34, respectively. The results are promising, but several limitations need to be addressed. For instance, the dataset is heavily imbalanced, meaning that most alarms in the dataset have label equal to 0 (i.e., they did not lead to a more critical alarm after the acknowledgment), while only a few events have label

equal to 1. This significantly impacts the model's ability to learn from the minority class, leading to overfitting and high bias toward the majority class.

7.7. Article IX: online classification of alarm floods

Article IX aligns with Objective 2 and specifically addresses Objective 2.3, which focuses on the development of ML methods to monitor, evaluate, and improve the performance of safety barriers. Similar to Articles VI and VII, this contribution explores the use of ML for improving the performance of industrial alarm systems. Specifically, Article IX focuses on ML methods to identify the most likely causes of alarm floods.

As previously mentioned, alarm floods (AFs) are periods of intense alarm activity characterized 10 or more annunciated alarms per 10 minutes per operator (ANSI/ISA, 2016). AFs are often cited in accident reports as contributing factors in major accidents, raising the need for better strategies to design, maintain, and operate industrial alarm systems. Unfortunately, identifying the causes of AFs is extremely challenging due to the large volume of triggered alarms, which often bury critical alarms under a plethora of uninformative ones, preventing the operators from diagnosing the issue.

Online AF classification deals with the development of methodologies and algorithms to identify and categorize ongoing alarm floods. This allows for the early detection of potential root causes of abnormal conditions, thereby enabling plant operators to initiate corrective actions before situations worsen. Such online AF classification methods integrate well into a Dynamic Risk Management (DRM) framework, as they provide diagnostic tools that enable continuous monitoring of industrial alarm systems.

Article IX introduces a novel approach to online AF classification through the use of Natural Language Processing (NLP) algorithms. The problem is reframed as an 'authorship identification' task, based on the assumption that different fault categories produce unique 'fingerprints' in the form of AF patterns. This is akin to how a text written by a particular author has distinctive characteristics, such as word usage, sentence structure, and length. In this view, a fault category (e.g., a valve stuck in an open position, or a malfunction of a specific control loop) can be considered an author describing a story related to the plant. Similar to how human authors use words to communicate, a fault communicates through alarms. Specifically, it produces episodes of AF, which can be regarded as a sentence of words. In other terms, we assume that a fault category uses alarms as words that form sentences (AFs), and these sentences are unique to that particular author (fault category). Consequently, the flood classification problem can be treated as an authorship identification problem, where an NLP model is trained on the writings of a specific author (fault) and then used to determine if a new flood belongs to that same author (fault category). Specifically, an ensemble of $\mathcal{N}$ word2vec models (see Section 6.3.3) was employed to learn contextual similarities between alarms in historical AFs, where $\mathcal{N} \in \mathbb{N}^{+}$ represents the number of fault categories considered. A scoring system was also proposed to evaluate model predictions based on their ability to identify the correct contextual similarities, ultimately allowing the identification of the most probable fault category.

The approach has been tested on simulated alarm data and yielded encouraging results, achieving an accuracy rate higher than 0.77 for four out of the six fault categories examined. Despite these positive outcomes, the models were less effective in identifying AFs from two specific fault categories, indicating areas for future research. Potential improvements could include hyperparameter tuning, alternative NLP models, or a more sophisticated scoring system using metrics like Term Frequency-Inverse Document Frequency (TF-IDF) to assign greater importance to crucial alarms.

7.8. X and XI: safety barrier assessment in environmental-critical systems

Articles X and XI align with Objective 2, specifically addressing Objective 2.3, which focuses on the development of ML methods to monitor, evaluate, and improve the performance of safety barriers. These articles outline a robust, digitally-based framework that dynamically assesses safety barriers in environmentally critical industrial facilities. The proposed approach leverages traditional hazard-identification techniques, data-driven simulation models, and resilience analysis to assess the efficacy of safety barriers, enabling the comparison between design alternatives and removing the need for on-site testing or simulation through first-principles models.

Facilities with a considerable potential to harm the environment – such as Waste-to-Energy (WtE) plants – operate under strict pollution control guidelines, often deploying Flue Gas Treatment (FGT) systems to mitigate emissions of hazardous pollutants like nitrogen oxides (NOX), hydrogen chloride (HCl), and sulfur dioxide (SO2). In this context, it is critical to ensure that FGT systems perform as intended and that safety barriers installed to prevent or mitigate excessive emissions are correctly designed and operated. However, there is no established methodology to evaluate and optimize the performance of safety barriers in environmental critical systems. Current industry practices heavily depend on empirical analyses and extensive on-site testing, which not only are resource-intensive but also present challenges for maintaining regulatory compliance. Addressing this gap, Articles X and XI introduce a novel method that fuses conventional hazard analysis, a digital model of the FGT system, and resilience analysis. This hybrid approach enables the identification, simulation, and evaluation of safety barriers that may prevent or mitigate excessive emissions in case of process deviations.

While Article X lays the foundational methodology and explores a basic case study, Article XI improves the framework and deepens the analysis, critically discussing the merits and constraints of this approach. The methodology employs a dual-tool strategy: it integrates classic risk assessment tools –used for identifying a set of critical scenarios that may lead to exceeding the emission limits and proposing additional safety measures– with modern, data-driven modeling techniques –used for the simulation of hazardous scenarios and assessment of the safety barriers, thereby sidestepping the need for unpractical field tests or first-principles models. The digital model effectively acts as a 'digital twin' of the actual facility, allowing for simulations that mimic real-world disturbances and assess the effectiveness of safety barriers. Resilience

analysis supports the evaluation by providing quantitative metrics to compare different barrier configurations, thus aiding risk-informed decision-making.

7.9. Article XII: investigate the role of ML and humans in the future of risk management.

Article XII addresses Objective 3, which delves into the synergy between ML and human interventions, examining the advantages and challenges of leveraging ML methods to support risk-based decision-making. Specifically, the article delves into the intricate relationship between human expertise and AI, critically examining their respective contributions. It raises important questions about whether ML can autonomously control risks and to what extent human involvement will continue to be crucial in the future of Risk Management.

The article discusses the evolving landscape of risk management, particularly as it enters its "4.0 phase" (Pasman and Fabiano, 2021) characterized by the increasing integration of cyber-technological systems and enhanced computational capabilities. While these advancements offer promising benefits like early warnings and proactive strategies, the article raises questions about whether digital risk management can fully live up to its promises. This contribution questions if recent advancements in AI and ML will eventually lead to a so-called “no-brainer” Risk Management, indicating a condition in which the responsibility for human and system safety is entirely moved to the machine, relegating humans to the role of observators or mere executors of the decisions taken by the machines.

To investigate these issues, the article describes an interesting case study involving the use of unsupervised learning –specifically, clustering algorithms– to categorize countries based on their similarity toward natural disaster exposure. The multifaceted nature of the associated RIFs, including environmental, social, and geopolitical elements, makes this a complex issue, even for experts in the field. This study aims to explore the effectiveness of clustering algorithms in extracting meaningful insights from disaster databases. The goal is to identify countries with similar exposure to natural hazards, thereby creating opportunities for knowledge sharing between nations. Concurrently, the research seeks to evaluate the level of autonomy achievable by ML algorithms, questioning the need for human involvement. This dual focus provides a forward-looking perspective on the evolving role of human and machine collaboration in risk management.

Using data from the EM-DAT database, the k-Means clustering algorithm was applied (see Section 6.3.2). The results indicate that clustering algorithms can discern meaningful patterns, yielding informative clusters of countries with shared risk profiles, thus highlighting the potential for inter-country knowledge exchange. However, the most intriguing findings arising from the analysis of the case study concern the interplay between ML and human involvement. Despite its analytical power, ML does not eliminate the uncertainties intrinsic to risk assessment. Uncertainty persists in data adequacy, modeling choices, and prediction interpretations, necessitating ongoing human oversight throughout the entire ML lifecycle. In essence, ML

introduces new layers of uncertainty that demand human expertise for effective management. Hence, the study concludes that ML technologies should complement, not replace, human involvement in the Risk Management framework. This aligns with the European Commission's principles of 'trustworthy AI' (European Commission, 2021), which advocates for explainable AI that prioritizes interpretability, avoids information overload, and ensures transparency. The role of humans in employing these tools is thus more pivotal than ever, reinforcing the idea that ML serves as an aide, not a replacement, to human judgment.

8. Discussion

In this section, the contributions previously described are discussed. The discussion is divided according to the three main objectives of this thesis: evaluate the current state of the art of ML techniques in the domain of RM (Section 8.1), develop ML-based methods to support and promote DRM (Section 8.2), examine the interplay between Machine Learning and human actors in the future of Risk Management (Section 8.3).

8.1. ML techniques in the domain of safety and reliability

Article I focused on investigating the current state of the art of ML applications for safety and reliability. Results suggest that since 2010, and particularly after 2017, the interest in ML has witnessed a remarkable surge. Several catalysts are behind this phenomenon, including the emergence of more sophisticated ML algorithms, advancements in computational power, greater accessibility to knowledge resources, and the availability of user-friendly tools that facilitate the development of advanced algorithms. Increased availability of labeled data due to advances in digitalization and simulation technologies has paved the dominance of supervised learning. However, it is pivotal to underline that this abundance primarily involves fault detection and diagnosis, which make large use of supervised techniques. In contrast, areas like risk assessment, given the infrequency of catastrophic events, face a notable data scarcity challenge, pushing the research toward unsupervised methods.

The analysis of the relevant literature reveals a pronounced tilt toward fault detection, diagnosis, and anomaly detection. In contrast, domains like risk assessment, system prognosis, and reliability analysis remain relatively unexplored. This bias may be explained by considering that industrial systems frequently encounter faults and anomalies, leading to a richer dataset on these occurrences. Risk assessment, conversely, focuses on less frequent, large-scale mishaps. Such events, spreading across systems and deeply connected to human actions, showcase a diverse spectrum of initiating factors and consequences. Harnessing data-driven strategies to address risk assessment is, thus, inherently intricate, given the multifaceted dynamics and the diverse nature of the involved phenomena.

In general, the findings indicate that ML is an emergent, influential trend in safety and reliability. With its ability to make predictions under uncertainty, ML can improve a wide range of domains, including predictive maintenance, early warning systems, and process monitoring. Yet, it is also worth acknowledging limitations and research gaps. Primarily, the challenge of training supervised models on infrequent events, hindered by the scarcity of labeled data. Also, a notable observation is that the use of ML for risk assessment in the chemical and process industry is largely unexplored, indicating a potential knowledge gap and underscoring the need for further exploration. Lastly, a detailed analysis of the research dynamics reveals potential bottlenecks. While several research groups are visibly active, inter-group collaborations seem sparse. Such a

segmented landscape can hinder the free flow of ideas and best practices. Amplifying collaborative initiatives is essential, as it can pave the way for interdisciplinary breakthroughs and accelerate the impact of ML on safety and reliability.

8.2. ML-based methods to support and promote Dynamic Risk Management

In this investigation, several methods have been developed to support different phases of DRM, including consequence evaluation (Articles II to IV), frequency estimation (Article V), and monitoring of safety barriers (Articles V to IX).

Overall, the findings indicate that ML possesses the ability to derive safety-critical insights from data, subsequently leveraging this knowledge to enhance essential DRM tasks. These methods can capture the influence of subtle and uncertain RIFs, which are difficult or impossible to model through canonical risk management techniques. Most of the computational burden associated with the development of ML algorithms is required to train and optimize the models, while deployment and inference are relatively inexpensive both in terms of computational resources and time. In addition to continuous monitoring, ML also facilitates ongoing training. New observations can be seamlessly integrated into the models, consistently updating and refining their knowledge base, and enabling them to capture the dynamics of evolving and degrading processes. This makes these techniques particularly suitable for DRM applications thanks to their fast predictions and their inherent characteristic of upgradability.

In addition, ML models, especially deep learning variants, have a unique capability to distill complexity. They can decipher and represent high-dimensional data spaces, extracting patterns and relationships from vast amounts of data that might be unintuitive to humans or that may be challenging to model through first principles. By modeling complex systems with fewer inputs, ML can help practitioners in the field of DRM. For example, instead of measuring many parameters (some of which might be hard to measure directly), they can focus on a select few, making the process more efficient and actionable. This is particularly useful in real-world industrial contexts where certain measurements can be costly, time-consuming, or technically challenging.

Diving deeper into the merits of each contribution, Articles II to IV demonstrate that ML has the capability to glean insights from historical accident data, subsequently harnessing the acquired knowledge to forecast the severity of new scenarios. These contributions address a critical gap in process safety, standing as a pioneering endeavor in learning from the past and applying Transfer Learning to Chemical Process Safety, unlocking safety-related insights embedded in accident data. These methods are particularly relevant for DRM as they ensure that lessons from the past are not forgotten. In addition, they hold great potential to support dynamic methods for risk analysis. For example, they integrate well into Dynamic Probabilistic Risk Assessment (DPRA) frameworks, where the use of Dynamic Event Trees and Monte Carlo simulations often requires the evaluation of many accidental scenarios. The resulting computational burden is one of the main

challenges of DPRA. In this context, the methods proposed in this study can guide the analysis toward the most critical scenarios, significantly reducing the need for rigorous simulations. Nevertheless, while the results are promising, various challenges arose during the investigations. For example, the performance of the models decreased as more severe consequences were considered, revealing the challenges linked with learning from rare events, such as those causing a large number of fatalities or involving specific substances, like Hydrogen. In fact, dataset imbalance is confirmed as one of the main obstacles to supervised classification, hindering the model's ability to learn infrequent events. This limitation may be partially mitigated by tuning the model parameters, as shown in Articles II and III. However, more advanced techniques, such as oversampling or class weighting, are needed to address the issue and allow the model to learn rare occurrences. In addition, classification schemes with unequal misclassification costs present unique challenges, such as the need for employing a multi-metric evaluation. In fact, reliance on metrics like accuracy can be misleading, as it might not reflect the true model performance. Thus, a comprehensive assessment, taking into account multiple metrics, becomes imperative to ensure the reliability and suitability of the classification scheme. Furthermore, the contributions on consequence predictions highlight the crucial role of data quality. The results demonstrate that incomplete or uncertain accident data markedly influence model performance. This underscores the limitations of many existing accident databases, which frequently contain inconsistent or incomplete information, stressing the importance of the data preprocessing activities, as discussed in Section 6.2.

Article V showcases the capabilities of Neural Networks in predicting the TTF of atmospheric tanks exposed to external fires. The model's ability to forecast TTFs, while accounting for the impact of mitigation measures, paves the way for improved risk assessments and evaluation of cascading effects. For example, the approach can significantly benefit dynamic frameworks for the calculation of escalation probabilities in domino scenarios. Currently, these frameworks rely on simplified correlations that link the characteristics of fire scenarios to the TTF. However, such correlations provide an incomplete overview of the phenomena, not accounting for safety barriers, and producing inaccurate results. The approach proposed in this study addresses these limitations by providing a fast, accurate, and inherently dynamic tool. In addition, the calculation of confidence intervals permits a better understanding of the robustness of the predictions, allowing more informed judgments based on the level of confidence. Predictions are computationally inexpensive, and the required input is intuitive and easy to obtain, making the approach well-suited for the dynamic assessment of domino scenarios. The results highlight how the synergy between digital simulations and ML models can mitigate computational demands, ultimately expediting and improving predictions for complex accident scenarios. However, in spite of the promising results, it is also worth acknowledging some limitations of the proposed approach. For instance, simulated data used to train the model come from a lumped parameter model, which may introduce errors in estimating the TTF values. Incorporating more rigorous TTF data, such as those obtained from large-scale experimental set-ups and/or validated CFD and

FEM models, could enhance the performance of the models. Also, it has been observed that the evaluation of mitigated scenarios is more challenging, requiring a more complicated model, and leading to relatively larger errors than unmitigated scenarios. In this view, more efforts should be directed toward optimizing the network or testing more advanced ML algorithms.

Articles VI to IX focus on the development of data-driven models for monitoring, evaluating, and improving safety barriers. In spite of the diverse approaches and applications, the studies show great potential, proving that ML models can learn from heterogeneous data and enable dynamic evaluation of safety barriers. Specifically, Articles VI to IX focus on industrial alarm systems, demonstrating how to harness ML techniques to build predictive methods that can link past and present process conditions to future alarm behavior. These versatile and dynamic tools can provide real-time guidance to control room operators and support the decision-making process under various scenarios. For instance, the approach detailed in Articles VI and VII offers early warnings for alarm chatter, allowing operators to investigate the issue in advance and eventually silence alarms before they become a nuisance. These models also forecast the end of a chattering sequence, letting operators promptly reactivate alarms and reducing the frequency of manual checks. Also, these techniques can consider the interaction between technical systems and operators' actions, as discussed in Article VIII, factoring in the complex interplay between these two actors in their predictions. Article IX significantly contributes to the field of online flood identification by proposing a novel, efficient, and accurate method. The findings suggest that ML, especially NLP algorithms, holds considerable promise in addressing alarm floods by offering fast and adaptive diagnostic tools that can improve safety in daily operations. The major limitations of these approaches are, again, issues related to data imbalance, especially considering Article VIII, which deals with infrequent events (i.e., the reoccurrence of a critical alarm after an operator acknowledgment), and the inherent challenges linked with unbalanced misclassification costs.

Articles XI and XII, focusing on data-driven process simulation models for the evaluation of safety barriers in environmental-critical facilities, showcase the potential of ML techniques to learn the dynamics of industrial processes directly from plant data, eliminating the need for unpractical and dangerous field tests. This approach goes beyond the conventional, static perspective on safety barriers (i.e., effective – not effective, with a context-independent Probability of Failure) towards a dynamic vision of the risk, where the effectiveness of safety barriers is closely linked to the dynamics of the underlying phenomena. This underscores the potential of data-driven models to complement and enrich traditional risk management methods, thereby enhancing environmental safety and achieving outcomes that are otherwise unfeasible through traditional approaches. Moreover, the innovative use of resilience analysis for the dynamic evaluation of safety barriers, along with the adoption of inherently updatable digital models, constitutes a substantial contribution to the domain of DRM.

8.3. Potential and limitations of Machine Learning in supporting risk-based decision-making.

This Ph.D. study highlights the potential of ML in achieving DRM, forecasting a future enriched by synergistic human-ML interactions. Article IV, for instance, shows how ML algorithms can learn from past data to guide control room operators, providing timely and informative feedback on the effectiveness of their actions. In addition, Article VIII demonstrates the potential of ML in supporting risk-based inspection and maintenance, which is a safety-critical and traditionally human-centric activity. Thanks to its inherent ability to model the complexity of unsafe interactions, taking into account heterogeneous data sources and diverse risk-influencing factors, ML offers significant potential to support risk-based decision-making. By analyzing vast and diverse datasets, ML techniques can extract safety-relevant knowledge, providing decision-makers with intuitive and contextually relevant insights on the present and future status of the system, improving risk communication, and thereby paving the way towards better-informed decisions.

As the role of ML in CPS expands, the interplay between human decisions and machine feedback will intensify, with ML predictions becoming increasingly integrated in the decision-making process. This evolving landscape prompts inquiries into the future roles of both humans and machines, questioning whether the decision process could be entirely moved to the machine, and investigating the limitations employing ML algorithms in safety-critical applications. In this context, Article XII investigates whether advancements in ML might render human judgment obsolete. In this regard, the findings reveal that while ML addresses uncertainty, it does not eradicate it. On the contrary, uncertainty persists in other forms, such as the uncertainty in data used to train the model, in the reliability of the model itself, and in the significance and interpretability of the model's predictions. In other words, notwithstanding the undisputed potential of ML techniques to manage the uncertainty related to unwanted events, it must be acknowledged that these techniques introduce new sources of uncertainty that must be understood and managed.

For example, there is uncertainty concerning data used to train the model, specifically regarding the significance and completeness of data. Ensuring that data encompasses all the necessary information and accurately represents the phenomena being assessed often hinges on human expertise. While automation can streamline certain aspects, human design remains integral to the data pipeline, particularly in selecting pertinent features and determining the most appropriate preprocessing techniques. Furthermore, despite rigorous efforts to ensure data completeness and relevance, there may be occasions when the process conditions deviate significantly from those seen during training. While the model will continue to generate outputs, their reliability may be in question. In these instances, human expertise is crucial to interpret the situation, contextualize the predictions, recognize the model's limitations, and, if appropriate, ignore the model's recommendations.

Another layer of uncertainty stems from the ML model itself. Considering the vast array of ML algorithms at our disposal, model selection becomes crucial. While automated model selection tools are available, depending solely on them can lead to skewed conclusions. For example, during the preliminary investigation of the approach demonstrated in Article V, several regression models were tested, including Random Forest. Notably, this model consistently overperformed the Neural Networks described in the article. However, a more in-depth analysis, supported by human expertise and domain knowledge, uncovered that Random Forest regressors struggle to accurately predict the TTF of events when tested with features beyond the bounds of their training data. In simpler terms, while Random Forest performed admirably when the scenarios matched the training distribution, its efficacy degraded when presented with new, unfamiliar scenarios. In this situation, an automated model selection algorithm might have concluded that RF was the best model, overlooking its severe limitations in novel situations. This underscores the indispensable role of human insight in tailoring modeling choices to the unique challenges of a problem.

Lastly, uncertainty is intrinsically tied to the interpretation of the model outputs. Specifically, human expertise is required to contextualize the results, taking into account modeling choices and the characteristics of the algorithm. Trust in the results becomes a concern due to the prevalence of black-box algorithms among the best-performing techniques. Methods to estimate prediction confidence are needed to address this issue. In this context, safety practitioners must gain a solid understanding of ML methodologies. Similarly, data scientists venturing into tools for high-risk sectors need to be thoroughly grounded in the safety aspects of their applications.

In light of these considerations, this study indicates that we are not close to a “no-brainer” condition, where the responsibility for safety-related decisions will be entirely moved to the machine. Conversely, while advanced ML algorithms offer valuable insights, human judgment remains at the heart of the decision-making process. These observations align with the EU proposal for a regulatory framework governing Artificial Intelligence (European Commission, 2021), which underscores the crucial role of human supervision, particularly for AI systems that hold a high-risk profile, like those overseeing the management and functioning of critical infrastructure. The European Union's recommendations support a 'human-in-the-loop' approach, emphasizing that AI platforms must be crafted to ensure that individuals entrusted with supervisory duties have the capability to execute the following tasks:

a) fully understand the capacities and limitations of the high-risk AI system and be able to duly monitor its operation;
b) remain aware of the possible tendency of automatically relying or over-relying on the output produced by a high-risk AI system (automation bias), in particular for high-risk AI systems used to provide information or recommendations for decisions to be taken by natural persons;
c) be able to correctly interpret the high-risk AI system’s output, taking into account in particular the characteristics of the system and the interpretation tools and methods available;

d) be able to decide, in any particular situation, not to use the high-risk AI system or otherwise disregard, override or reverse the output of the high-risk AI system;
e) be able to intervene on the operation of the high-risk AI system or interrupt the system through a stop button or a similar procedure.

This idea of human-machine cooperation integrates well into the concept of sustainable, human-centric, and resilient industry embedded into Industry 5.0 (European Commission Directorate-General for Research and Innovation et al., 2021). Furthermore, the emphasis on human involvement underscores the importance of interdisciplinarity, blending safety science with digital technologies. This demands professionals who are well-versed in safety-related concerns and also have a clear understanding of the intricacies and limitations of ML models.

9. Conclusions

The primary goal of this thesis is to advance the knowledge of ML techniques to support DRM in the chemical and process industry. A preliminary analysis of the relevant literature reveals that although data-driven techniques are gaining traction among safety researchers, the landscape appears relatively scattered, with many challenges and obstacles that need to be addressed, including difficulties linked with learning from rare events, trust issues arising from black-box models, and sparse inter-group collaborations limiting research interdisciplinarity. In this vein, this study aimed at providing a practical contribution to the research on ML methods in DRM, specifically addressing three critical tasks: consequence prediction, frequency evaluation, and monitoring of safety barriers. The investigation concretized into a set of dynamic and inherently updatable methods and tools that have great potential to advance the research on DRM and proceed toward a more proactive and dynamic approach to process safety. The contributions show how ML methods can be used to take advantage of the wealth of data made available by the widespread digitalization of industrial sectors in order to extract and retain safety-relevant knowledge. ML has demonstrated its effectiveness in capturing complex process dynamics and the intricate interplay between RIFs, such as the interactions between technical systems and human operators, or the multifaceted relationship between accident features and its consequences. Given the escalating complexity and interconnectedness of industrial facilities, ML emerges as a feasible alternative to first-principle modeling, which often proves cost-prohibitive or not applicable in such scenarios.

A broad spectrum of ML methodologies, spanning from classification and regression to clustering and NLP algorithms, have been applied and discussed. This underlines the rich diversity within the ML domain, offering a comprehensive array of tools adaptable to the needs of various challenges in DRM. Also, a variety of data types have been utilized, including numerical, categorical, mixed, and event sequences, proving the flexibility of ML methods in analyzing heterogeneous data sources. The flexibility of these tools showcases the immense potential of ML, underlining its distinct capability to develop predictive solutions that seamlessly correlate past and current process conditions to anticipated risk levels. In addition to the advantages of ML techniques, limitations of data-driven approaches have also been thoroughly discussed, including challenges linked to imbalanced and cost-sensitive classification, the importance of good quality data and sound data preprocessing, model interpretability, quantification of prediction uncertainty, and the challenges related to model selection and hyperparameters tuning.

In addition to proposing novel technical solutions, this research also provided insights into the potential and the limitations of adopting ML to support risk-based decision-making. While the trajectory of progress suggests an increasing adoption of intelligent systems, human expertise will remain pivotal in addressing the uncertainties introduced by ML algorithms. This envisions a collaborative paradigm of human-machine synergy rather than rivalry. Expertise and domain knowledge will be paramount, ensuring that humans

exercise appropriate oversight and grasp the model limitations. This understanding will enable safety practitioners to effectively contextualize predictions and, when necessary, disregard machine-driven suggestions.

Nonetheless, while this study has made significant advances in addressing knowledge gaps, the research on ML in the realm of Risk Management remains nascent. The methodologies outlined in this investigation represent a preliminary—yet promising—integration between these two domains. Future efforts should be directed at refining the presented approaches and venturing into alternative solutions. Additionally, it is imperative to recognize that there are other potential avenues for improvement not explored in this research, but which will likely be relevant in the future of risk management. These include the application of ML techniques to Human Reliability Analysis (HRA) – for instance, fatigue detection aided by computer vision, dynamic estimation of human error probability, and monitoring of maintenance activities aided by smart devices. Another promising area is automated inspection and maintenance of industrial equipment through, e.g., remotely operated or autonomous drones. Cybersecurity concerns, encompassing both the potential of ML in preventing malicious intrusions and the vulnerabilities introduced by incorporating ML algorithms into industrial IT systems, also require attention.

Notwithstanding the limitations and scope for further improvements, this Ph.D. research has effectively addressed all the established objectives. It has broadened the understanding of ML methods within the context of CPS, provided practical tools and strategies to assist a variety of DRM tasks, offered valuable perspectives on the challenges associated with the integration of ML techniques, and provided insights into the evolving roles of humans and intelligent systems in the future of safety sciences.

10. References

Abdallah, Z.S., Du, L., Webb, G.I., 2017. Data Preparation, in: Encyclopedia of Machine Learning and Data Mining. Springer US, Boston, MA, pp. 318–327. https://doi.org/10.1007/978-1-4899-7687-1_62

Abdul Aziz, H., Mohd Shariff, A., 2017. A Journey of Process Safety Management Program for Process Industry. Int. J. Eng. Technol. Sci. 4, 119–127. https://doi.org/10.15282/ijets.8.2017.1.10.1085

Acuña, E., Rodriguez, C., 2004. The Treatment of Missing Values and its Effect on Classifier Accuracy, in: Classification, Clustering, and Data Mining Applications. Springer Berlin Heidelberg, Berlin, Heidelberg, pp. 639–647. https://doi.org/10.1007/978-3-642-17103-1_60

AEA Technology, 1999. MHIDAS (Major Hazard Incident Data Service).

American Petroleum Institute, 1991. API 750 - Management of Process Hazards. Washington.

ANSI/ISA, 2016. ANSI/ISA–18.2–2016 Management of Alarm Systems for the Process Industries. ANSI/ISA.

ARAMIS project team, 2004. Deliverable D.1.C.

Aven, T., 2016. Risk assessment and risk management: Review of recent advances on their foundation. Eur. J. Oper. Res. 253, 1–13. https://doi.org/10.1016/j.ejor.2015.12.023

Aven, T., Renn, O., 2009. On risk defined as an event where the outcome is uncertain. J. Risk Res. 12, 1–11. https://doi.org/10.1080/13669870802488883

Barr, A., Feigenbaum, E.A., 1981. Introduction, in: The Handbook of Artificial Intelligence. Elsevier, pp. 1–17. https://doi.org/10.1016/B978-0-86576-089-9.50006-5

Bengio, Y., Ducharme, R., Vincent, P., Janvin, C., 2003. A Neural Probabilistic Language Model. J. Mach. Learn. Res. 3, 1137–1155.

Brink, H., Richards, J., Fetherolf, M., 2016. Real-World Machine Learning, first. ed. Manning Publications, Shelter Island.

Caldarini, G., Jaf, S., McGarry, K., 2022. A Literature Survey of Recent Advances in Chatbots. Information 13, 41. https://doi.org/10.3390/info13010041

Canadian Society for Chemical Engineering, 2012. Process Safety Management Standard. Ottawa, Ontario K1P 6E2.

Center for Chemical Process Safety, 2010. Layer of Protection Analysis: Simplified Process Risk Assessment. Wiley. https://doi.org/10.1002/9780470935590

Chen, M.X., Lee, B.N., Bansal, G., Cao, Y., Zhang, S., Lu, J., Tsay, J., Wang, Y., Dai, A.M., Chen, Z., Sohn, T., Wu, Y., 2019. Gmail Smart Compose: Real-Time Assisted Writing.

Chen, Y.-J., Wu, C.-H., Chen, Y.-M., Li, H.-Y., Chen, H.-K., 2017. Enhancement of fraud detection for narratives in annual reports. Int. J. Account. Inf. Syst. 26, 32–45. https://doi.org/10.1016/j.accinf.2017.06.004

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., Anil, R., Haque, Z., Hong, L., Jain, V., Liu, X., Shah, H., 2016. Wide & Deep Learning for

Recommender Systems.

Chu, X., 2018. Data Cleaning, in: Encyclopedia of Big Data Technologies. Springer International Publishing, Cham, pp. 1–7. https://doi.org/10.1007/978-3-319-63962-8_3-1

Chung, P.W.H., Jefferson, M., 1998. The integration of accident databases with computer tools in the chemical industry. Comput. Chem. Eng. 22, S729–S732. https://doi.org/10.1016/S0098-1354(98)00135-5

Clarivate, 2022. Web of Science [WWW Document].

Cozzani, V., Gubinelli, G., Antonioni, G., Spadoni, G., Zanelli, S., 2005. The assessment of risk caused by domino effect in quantitative area risk analysis. J. Hazard. Mater. 127, 14–30. https://doi.org/10.1016/j.jhazmat.2005.07.003

Creswell, J.W., 2014. Research design: qualitative, quantitative, and mixed methods approaches, 4th ed. SAGE Publications, Thousand Oaks, California.

de Oliveira, N.R., Pisa, P.S., Lopez, M.A., de Medeiros, D.S. V., Mattos, D.M.F., 2021. Identifying Fake News on Social Networks Based on Natural Language Processing: Trends and Challenges. Information 12, 38. https://doi.org/10.3390/info12010038

Diekmann, E.J., 1992. Risk analysis: lessons from artificial intelligence. Int. J. Proj. Manag. 10, 75–80. https://doi.org/10.1016/0263-7863(92)90059-I

Dimaio, F., Scapinello, O., Zio, E., Ciarapica, C., Cincotta, S., Crivellari, A., Decarli, L., Larosa, L., 2021. Accounting for Safety Barriers Degradation in the Risk Assessment of Oil and Gas Systems by Multistate Bayesian Networks. Reliab. Eng. Syst. Saf. 216, 107943. https://doi.org/10.1016/j.ress.2021.107943

Donders, A.R.T., van der Heijden, G.J.M.G., Stijnen, T., Moons, K.G.M., 2006. Review: A gentle introduction to imputation of missing values. J. Clin. Epidemiol. 59, 1087–1091. https://doi.org/10.1016/j.jclinepi.2006.01.014

El Zahran, T., Geha, M., Sakr, F., Bachir, R., El Sayed, M., 2022. The Beirut Port Blast: spectrum of injuries and clinical outcomes at a large tertiary care center in Beirut, Lebanon. Eur. J. Trauma Emerg. Surg. 48, 4919–4926. https://doi.org/10.1007/s00068-022-02023-9

European Commission, 2021. Proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on Artificial Intelligence (Artificial Intelligence act) and amending certain Union Legislative Acts.

European Commission Directorate-General for Research and Innovation, Breque, M., De Nul, L., Petridis, A., 2021. Industry 5.0 – Towards a sustainable, human-centric and resilient European industry. https://doi.org/10.2777/308407

European Parliament Council of the European Union, 2012. Directive 2012/18/EU of the European Parliament and of the Council of 4 July 2012 on the control of major-accident hazards involving dangerous substances.

Fernández, A., García, S., Galar, M., Prati, R.C., Krawczyk, B., Herrera, F., 2018. Cost-Sensitive Learning, in: Learning from Imbalanced Data Sets. Springer International Publishing, Cham, pp. 63–78. https://doi.org/10.1007/978-3-319-98074-4_4

Ferrari, R., 2015. Writing narrative style literature reviews. Med. Writ. 24, 230–235. https://doi.org/10.1179/2047480615Z.000000000329

Finlay, J., Dix, A., 2020. An Introduction to Artificial Intelligence. CRC Press. https://doi.org/10.1201/9781003072485

García, S., Luengo, J., Herrera, F., 2015. Data Preparation Basic Models. pp. 39–57. https://doi.org/10.1007/978-3-319-10247-4_3

Gascard, E., Simeu-Abazi, Z., 2018. Quantitative Analysis of Dynamic Fault Trees by means of Monte Carlo Simulations: Event-Driven Simulation Approach. Reliab. Eng. Syst. Saf. 180, 487–504. https://doi.org/10.1016/j.ress.2018.07.011

Goodfellow, I., Bengio, Y., Courville, A., 2016. 5.2 Capacity, Overfitting and Underfitting, in: Adaptive Computation and Machine Learning Series. MIT Press.

Gritten, D., 2022. Toxic gas leak at Jordan's Aqaba port kills 13, injures hundreds. BBC news.

Han, J., Kamber, M., Pei, J., 2012. 3 - Data Preprocessing, in: Han, J., Kamber, M., Pei, J.B.T.-D.M. (Third E. (Eds.), The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Boston, pp. 83–124. https://doi.org/https://doi.org/10.1016/B978-0-12-381479-1.00003-4

Han, J., Kamber, M., Pei, J., 2011. Chapter 11. Advanced Cluster Analysis, in: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA.

Hancock, J.T., Khoshgoftaar, T.M., 2020. Survey on categorical data for neural networks. J. Big Data 7, 28. https://doi.org/10.1186/s40537-020-00305-w

Hashemi, S.J., Ahmed, S., Khan, F.I., 2014. Risk-based operational performance analysis using loss functions. Chem. Eng. Sci. 116, 99–108. https://doi.org/10.1016/j.ces.2014.04.042

Hasib, K.M., Iqbal, M.S., Shah, F.M., Mahmud, J. Al, Popel, M.H., Showrov, M.I.H., Ahmed, S., Rahman, O., 2020. A Survey of Methods for Managing the Classification and Solution of Data Imbalance Problem. https://doi.org/10.3844/jcssp.2020.1546.1557

Hastie, T., Tibshirani, R., Friedman, J., 2009. Model Assessment and Selection. pp. 219–259. https://doi.org/10.1007/978-0-387-84858-7_7

Hauge, S., Okstad, E., Paltrinieri, N., Edwin, N., Vatn, J., Bodsberg, L., 2015. Handbook for monitoring of barrier status and associated risk in the operational phase, the risk barometer approach. SINTEF F27045. Trondheim, Norw.

Hegde, J., Rokseth, B., 2020. Applications of machine learning methods for engineering risk assessment – A review. Saf. Sci. 122, 104492. https://doi.org/10.1016/j.ssci.2019.09.015

Hirschberg, J., Manning, C.D., 2015. Advances in natural language processing. Science (80-.). 349, 261–266.

https://doi.org/10.1126/science.aaa8685

International Organization for Standardization (ISO), 2018. ISO 31000 Risk Management.

Izadi, I., Shah, S.L., Shook, D.S., Chen, T., 2009. An Introduction to Alarm Analysis and Design. IFAC Proc. Vol. 42, 645–650. https://doi.org/10.3182/20090630-4-ES-2003.00107

Jain, A., Patel, H., Nagalapatti, L., Gupta, N., Mehta, S., Guttula, S., Mujumdar, S., Afzal, S., Sharma Mittal, R., Munigala, V., 2020. Overview and Importance of Data Quality for Machine Learning Tasks, in: Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. ACM, New York, NY, USA, pp. 3561–3562. https://doi.org/10.1145/3394486.3406477

James, G., Witten, D., Hastie, T., Tibshirani, R., 2021. 10.3.1 K-means clustering, in: An Introduction to Statistical Learning. pp. 390–393. https://doi.org/10.1007/978-1-0716-1418-1_12

Johnson, R.B., Christensen, L., 2015. Educational Research: Quantitative, Qualitative, and Mixed, 5th ed. SAGE Publications, Inc, Thousand Oaks, California.

Kabir, S., Papadopoulos, Y., 2019. Applications of Bayesian networks and Petri nets in safety, reliability, and risk assessments: A review. Saf. Sci. 115, 154–175. https://doi.org/10.1016/j.ssci.2019.02.009

Kalantarnia, M., Khan, F., Hawboldt, K., 2009. Dynamic risk assessment using failure assessment and Bayesian theory. J. Loss Prev. Process Ind. 22, 600–606. https://doi.org/10.1016/j.jlp.2009.04.006

Khakzad, N., Landucci, G., Cozzani, V., Reniers, G., Pasman, H., 2018. Cost-effective fire protection of chemical plants against domino effects. Reliab. Eng. Syst. Saf. 169, 412–421. https://doi.org/10.1016/j.ress.2017.09.007

Khan, F., Hashemi, S.J., Paltrinieri, N., Amyotte, P., Cozzani, V., Reniers, G., 2016. Dynamic risk management: a contemporary approach to process safety management. Curr. Opin. Chem. Eng. 14, 9–17. https://doi.org/10.1016/j.coche.2016.07.006

Kletz, T., 2012. The history of process safety. J. Loss Prev. Process Ind. 25, 763–765. https://doi.org/10.1016/j.jlp.2012.03.011

Kondaveeti, S.R., Izadi, I., Shah, S.L., Black, T., 2010. Graphical Representation of Industrial Alarm Data. IFAC Proc. Vol. 43, 181–186. https://doi.org/10.3182/20100831-4-FR-2021.00033

Kothari, C.R., 2004. Research methodology: methods and techniques, 2nd ed. New Age International (P) Ltd, New Delhi.

Landucci, G., Gubinelli, G., Antonioni, G., Cozzani, V., 2009. The assessment of the damage probability of storage tanks in domino events triggered by fire. Accid. Anal. Prev. 41, 1206–1215. https://doi.org/10.1016/j.aap.2008.05.006

Le Coze, J.C., 2013. What have we learned about learning from accidents? Post-disasters reflections. Saf. Sci. 51, 441–453. https://doi.org/10.1016/j.ssci.2012.07.007

Lee, J., Cameron, I., Hassall, M., 2019. Improving process safety: What roles for Digitalization and Industry 4.0? Process Saf. Environ. Prot. 132, 325–339. https://doi.org/10.1016/j.psep.2019.10.021

Lees, F., 2012. Hazard Assessment, in: Mannan, S. (Ed.), Lees' Loss Prevention in the Process Industries. Elsevier, pp. 284–404. https://doi.org/10.1016/B978-0-12-397189-0.00009-4

Li, J., Cheng, K., Wang, S., Morstatter, F., Trevino, R.P., Tang, J., Liu, H., 2016. Feature Selection: A Data Perspective. https://doi.org/10.1145/3136625

Li, Y., 2017. Deep Reinforcement Learning: An Overview.

Liberati, A., Altman, D.G., Tetzlaff, J., Mulrow, C., Gøtzsche, P.C., Ioannidis, J.P.A., Clarke, M., Devereaux, P.J., Kleijnen, J., Moher, D., 2009. The PRISMA statement for reporting systematic reviews and meta-analyses of studies that evaluate health care interventions: explanation and elaboration, Journal of clinical epidemiology. https://doi.org/10.1016/j.jclinepi.2009.06.006

Mannor, S., Jin, X., Han, J., Jin, X., Han, J., Jin, X., Han, J., Zhang, X., 2011. K-Means Clustering, in: Encyclopedia of Machine Learning. Springer US, Boston, MA, pp. 563–564. https://doi.org/10.1007/978-0-387-30164-8_425

Mikolov, T., Chen, K., Corrado, G., Dean, J., 2013. Efficient Estimation of Word Representations in Vector Space.

Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective, Adaptive Computation and Machine Learning. MIT Press, Cambridge, Massachusetts, United States.

OECD, 2015. Frascati Manual 2015. https://doi.org/https://doi.org/https://doi.org/10.1787/9789264239012-en

OpenAI, 2023. GPT-4 Technical Report.

Otter, D.W., Medina, J.R., Kalita, J.K., 2021. A Survey of the Usages of Deep Learning for Natural Language Processing. IEEE Trans. Neural Networks Learn. Syst. 32, 604–624. https://doi.org/10.1109/TNNLS.2020.2979670

Palacio-Niño, J.-O., Berzal, F., 2019. Evaluation Metrics for Unsupervised Learning Algorithms.

Paltrinieri, N., Comfort, L., Reniers, G., 2019. Learning about risk: Machine learning for risk assessment. Saf. Sci. 118, 475–486. https://doi.org/10.1016/j.ssci.2019.06.001

Paltrinieri, N., Khan, F., Amyotte, P., Cozzani, V., 2014a. Dynamic approach to risk management: Application to the Hoeganaes metal dust accidents. Process Saf. Environ. Prot. 92, 669–679. https://doi.org/10.1016/j.psep.2013.11.008

Paltrinieri, N., Reniers, G., 2017. Dynamic risk analysis for Seveso sites. J. Loss Prev. Process Ind. 49, 111–119. https://doi.org/10.1016/j.jlp.2017.03.023

Paltrinieri, N., Scarponi, G.E., Khan, F., Hauge, S., 2014b. Addressing Dynamic Risk in the Petroleum Industry by Means of Innovative Analysis Solutions. Chem. Eng. Trans. 36. https://doi.org/10.3303/CET1436076

Paltrinieri, N., Tugnoli, A., Buston, J., Wardman, M., Cozzani, V., 2013. Dynamic Procedure for Atypical Scenarios Identification (DyPASI): A new systematic HAZID tool. J. Loss Prev. Process Ind. 26, 683–695. https://doi.org/10.1016/j.jlp.2013.01.006

Pan, S.J., Yang, Q., 2010. A Survey on Transfer Learning. IEEE Trans. Knowl. Data Eng. 22, 1345–1359. https://doi.org/10.1109/TKDE.2009.191

Pasman, H.J., 2009. Learning from the past and knowledge management: Are we making progress? J. Loss Prev. Process Ind. 22, 672–679. https://doi.org/10.1016/j.jlp.2008.07.010

Pasman, H.J., Duxbury, H.A., Bjordal, E.N., 1992. Major hazards in the process industries: Achievements and challenges in loss prevention. J. Hazard. Mater. 30, 1–38. https://doi.org/10.1016/0304-3894(92)87072-N

Pasman, H.J., Fabiano, B., 2021. The Delft 1974 and 2019 European Loss Prevention Symposia: Highlights and an impression of process safety evolutionary changes from the 1st to the 16th LPS. Process Saf. Environ. Prot. 147, 80–91. https://doi.org/10.1016/j.psep.2020.09.024

Pasman, H.J., Fouchier, C., Park, S., Quddus, N., Laboureur, D., 2020. Beirut ammonium nitrate explosion: Are not we really learning anything? Process Saf. Prog. 39. https://doi.org/10.1002/prs.12203

Peres, R.S., Jia, X., Lee, J., Sun, K., Colombo, A.W., Barata, J., 2020. Industrial Artificial Intelligence in Industry 4.0 - Systematic Review, Challenges and Outlook. IEEE Access 8, 220121–220139. https://doi.org/10.1109/ACCESS.2020.3042874

Petroleum Safety Authority, 2013. Principles for barrier management in the petroleum industry.

Pruzan, P., 2016. Research Methodology. Springer International Publishing, Cham. https://doi.org/10.1007/978-3-319-27167-5

Qin, Y., Ding, S., Wang, L., Wang, Y., 2019. Research Progress on Semi-Supervised Clustering. Cognit. Comput. 11, 599–612. https://doi.org/10.1007/s12559-019-09664-w

Rabiti, C., Mandelli, D., Cogliati, J., Kinoshita, R., 2013. Mathematical Framework For The Analysis Of Dynamc Stochastic Systems With The Raven Code. United States.

Ramírez-Gallego, S., García, S., Mouriño-Talín, H., Martínez-Rego, D., Bolón-Canedo, V., Alonso-Betanzos, A., Benítez, J.M., Herrera, F., 2016. Data discretization: taxonomy and big data challenge. WIREs Data Min. Knowl. Discov. 6, 5–21. https://doi.org/10.1002/widm.1173

Raschka, S., 2018. Model Evaluation, Model Selection, and Algorithm Selection in Machine Learning. https://doi.org/1811.12808

Rousseeuw, P.J., 1987. Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. J. Comput. Appl. Math. 20, 53–65. https://doi.org/10.1016/0377-0427(87)90125-7

Sammut, C., Webb, G.I., 2017a. Supervised Learning, in: Sammut, Claude, Webb, Geoffrey I (Eds.), Encyclopedia of Machine Learning and Data Mining. Springer US, Boston, MA, pp. 1213–1214. https://doi.org/10.1007/978-1-4899-7687-1_803

Sammut, C., Webb, G.I., 2017b. Unsupervised Learning, in: Encyclopedia of Machine Learning and Data Mining. Springer US, Boston, MA, pp. 1304–1304. https://doi.org/10.1007/978-1-4899-7687-1_976

Shi, C., Wei, B., Wei, S., Wang, W., Liu, H., Liu, J., 2021. A quantitative discriminant method of elbow point for

the optimal number of clusters in clustering algorithm. EURASIP J. Wirel. Commun. Netw. 2021, 31. https://doi.org/10.1186/s13638-021-01910-w

Shultz, T.R., Fahlman, S.E., Craw, S., Andritsos, P., Tsaparas, P., Silva, R., Drummond, C., Ling, C.X., Sheng, V.S., Drummond, C., Lanzi, P.L., Gama, J., Wiegand, R.P., Sen, P., Namata, G., Bilgic, M., Getoor, L., He, J., Jain, S., Stephan, F., Jain, S., Stephan, F., Sammut, C., Harries, M., Sammut, C., Ting, K.M., Pfahringer, B., Case, J., Jain, S., Wagstaff, K.L., Nijssen, S., Wirth, A., Ling, C.X., Sheng, V.S., Zhang, X., Sammut, C., Cancedda, N., Renders, J.-M., Michelucci, P., Oblinger, D., Keogh, E., Mueen, A., 2011. Clustering, in: Encyclopedia of Machine Learning. Springer US, Boston, MA, pp. 180–180. https://doi.org/10.1007/978-0-387-30164-8_124

Sievers, M., Madni, A.M., 2022. Dynamic Causal Hidden Markov Model Risk Assessment, in: Recent Trends and Advances in Model Based Systems Engineering. Springer International Publishing, Cham, pp. 141–150. https://doi.org/10.1007/978-3-030-82083-1_13

Sklet, S., 2006. Safety barriers: Definition, classification, and performance. J. Loss Prev. Process Ind. 19, 494–506. https://doi.org/10.1016/j.jlp.2005.12.004

Sola, J., Sevilla, J., 1997. Importance of input data normalization for the application of neural networks to complex industrial problems. IEEE Trans. Nucl. Sci. 44, 1464–1468. https://doi.org/10.1109/23.589532

Solangi, Y.A., Solangi, Z.A., Aarain, S., Abro, A., Mallah, G.A., Shah, A., 2018. Review on Natural Language Processing (NLP) and Its Toolkits for Opinion Mining and Sentiment Analysis, in: 2018 IEEE 5th International Conference on Engineering Technologies and Applied Sciences (ICETAS). IEEE, pp. 1–4. https://doi.org/10.1109/ICETAS.2018.8629198

Stanula, P., Ziegenbein, A., Metternich, J., 2018. Machine learning algorithms in production: A guideline for efficient data source selection. Procedia CIRP 78, 261–266. https://doi.org/10.1016/j.procir.2018.08.177

Stoop, J., de Kroes, J., Hale, A., 2017. Safety science, a founding fathers' retrospection. Saf. Sci. 94, 103–115. https://doi.org/10.1016/j.ssci.2017.01.006

Strike, K., El Emam, K., Madhavji, N., 2001. Software cost estimation with incomplete data. IEEE Trans. Softw. Eng. 27, 890–908. https://doi.org/10.1109/32.962560

Taleb-Berrouane, M., Pasman, H., 2022. Integrated dynamic risk management in process plants. pp. 525–560. https://doi.org/10.1016/bs.mcps.2022.05.006

Tamascelli, N., Rao, H.R.M., Cozzani, V., Paltrinieri, N., Chen, T., 2023. Online Classification of Alarm Floods Using a Word2vec Algorithm, in: IECON 2023 – 49th Annual Conference of the IEEE Industrial Electronics Society.

Van Der Maaten, L., Postma, E., den Herik, J., 2009. Dimensionality reduction: a comparative review. J Mach Learn Res 10, 66–71.

van Engelen, J.E., Hoos, H.H., 2020. A survey on semi-supervised learning. Mach. Learn. 109, 373–440.

https://doi.org/10.1007/s10994-019-05855-6

Vanschoren, J., 2018. Meta-Learning: A Survey.

Vatn, J., 2012. Can we understand complex systems in terms of risk analysis? Proc. Inst. Mech. Eng. Part O J. Risk Reliab. 226, 346–358. https://doi.org/10.1177/1748006X11405944

Villa, V., Paltrinieri, N., Khan, F., Cozzani, V., 2016. Towards dynamic risk analysis: A review of the risk assessment approach and its limitations in the chemical process industry. Saf. Sci. 89, 77–93. https://doi.org/10.1016/j.ssci.2016.06.002

Vinnem, J.E., Bye, R., Gran, B.A., Kongsvik, T., Nyheim, O.M., Okstad, E.H., Seljelid, J., Vatn, J., 2012. Risk modelling of maintenance work on major process equipment on offshore petroleum installations. J. Loss Prev. Process Ind. 25, 274–292. https://doi.org/10.1016/j.jlp.2011.11.001

Vlachos, M., 2017. Dimensionality Reduction, in: Encyclopedia of Machine Learning and Data Mining. Springer US, Boston, MA, pp. 354–361. https://doi.org/10.1007/978-1-4899-7687-1_71

Wang, H., Wu, H., He, Z., Huang, L., Church, K.W., 2022. Progress in Machine Translation. Engineering 18, 143–153. https://doi.org/10.1016/j.eng.2021.03.023

Wang, S., Tang, J., Liu, H., 2016. Feature Selection, in: Encyclopedia of Machine Learning and Data Mining. Springer US, Boston, MA, pp. 1–9. https://doi.org/10.1007/978-1-4899-7502-7_101-1

Wen, J.X., Marono, M., Moretto, P., Reinecke, E.-A., Sathiah, P., Studer, E., Vyazmina, E., Melideo, D., 2022. Statistics, lessons learned and recommendations from analysis of HIAD 2.0 database. Int. J. Hydrogen Energy 47, 17082–17096. https://doi.org/10.1016/j.ijhydene.2022.03.170

Witten, I.H., Frank, E., Hall, M.A., 2011. Data Transformations, in: Witten, I.H., Frank, E., Hall, M.A.B.T.-D.M.P.M.L.T. and T. (Third E. (Eds.), Data Mining: Practical Machine Learning Tools and Techniques. Elsevier, Boston, pp. 305–349. https://doi.org/10.1016/B978-0-12-374856-0.00007-9

Xiong, H., Gaurav Pandey, Steinbach, M., Vipin Kumar, 2006. Enhancing data analysis with noise removal. IEEE Trans. Knowl. Data Eng. 18, 304–319. https://doi.org/10.1109/TKDE.2006.46

Xu, Z., Saleh, J.H., 2021. Machine learning for reliability engineering and safety applications: Review of current status and future opportunities. Reliab. Eng. Syst. Saf. 211, 107530. https://doi.org/10.1016/j.ress.2021.107530

Yang, J., Zhang, M., Zuo, Y., Cui, X., Liang, C., 2023. Improved models of failure time for atmospheric tanks under the coupling effect of multiple pool fires. J. Loss Prev. Process Ind. 81, 104957. https://doi.org/10.1016/j.jlp.2022.104957

Zeng, T., Chen, G., Yang, Y., Chen, P., Reniers, G., 2020. Developing an advanced dynamic risk analysis method for fire-related domino effects. Process Saf. Environ. Prot. 134, 149–160. https://doi.org/10.1016/j.psep.2019.11.029

Zeng, Z., Zio, E., 2018. Dynamic Risk Assessment Based on Statistical Failure Data and Condition-Monitoring Degradation Data. IEEE Trans. Reliab. 67, 609–622. https://doi.org/10.1109/TR.2017.2778804

Zhu, X., 2017. Semi-supervised Learning, in: Encyclopedia of Machine Learning and Data Mining. Springer US, Boston, MA, pp. 1142–1147. https://doi.org/10.1007/978-1-4899-7687-1_749

Zio, E., 2018. The future of risk assessment. Reliab. Eng. Syst. Saf. 177, 176–190. https://doi.org/10.1016/j.ress.2018.04.020

Annex – Report of the academic activities carried out during the doctoral path and list of publications

In fulfillment of the requirements of the University of Bologna concerning the content of the Ph.D. thesis, the present annex reports an outline of the activities and the list of the publications. These were extensively discussed in the main body of this thesis.

Activity 1. A literature review on Artificial Intelligence and Machine Learning methods was conducted and submitted for publication in a peer-reviewed journal (Article I). The review focuses on Machine Learning techniques in fault and anomaly detection, reliability analysis, system diagnosis and prognosis, risk analysis, and risk assessment of engineering-related systems. Systematic and narrative approaches have been utilized to provide a quantitative and qualitative analysis of the state of the art.

Activity 2. Machine Learning classification models were used to learn from accident databases and use the acquired knowledge to predict the consequences of major accidents involving dangerous substances (Article II). These algorithms take a range of easily accessible and informative accident characteristics as input and estimate the number of fatalities and injuries caused by the accident. The models' capability to transfer learning across databases has also been investigated (Article III). Finally, the potential application of this methodology for Risk-Based Inspection (RBI) in pure hydrogen environments has been examined (Article IV).

Activity 3. A Neural Network-based method for estimating the Time-To-Failure (TTF) of atmospheric tanks exposed to external fires was developed (Article V). The algorithm is trained using data from a lumped-parameter model and can be used to predict TTF under both unmitigated and mitigated conditions. Predictions are paired with confidence intervals to reflect the level of uncertainty involved.

Activity 4. Machine Learning (ML) techniques for monitoring and improving industrial alarm systems were investigated. Specifically, the research focused on developing ML models for several key areas, such as (i) the prediction of nuisance alarms (Articles VI and VII), (ii) the evaluation of control room operator actions (article VIII), and (iii) the utilization of Natural Language Processing (NLP) algorithms to detect the causes of alarm floods (article IX).

Activity 5. A methodology for the evaluation of safety barriers in environmentally critical facilities was developed (Articles X and XI). The proposed method integrates conventional risk assessment tools with data-driven process simulation models, showcasing the potential of digital technologies to enable a dynamic and proactive approach to process safety.

Activity 6. The candidate explored the present and future role of humans in the context of digital safety. An example of unsupervised clustering of natural events has been described and discussed (article XII). The findings underscore the ability of Machine Learning to complement and support traditional risk management approaches. Nonetheless, human expertise remains indispensable for managing the added uncertainty arising from ML algorithms and interpreting the data to transform recommendations into concrete actions.

The publications produced in this Ph.D. study are listed below:

I. Tamascelli, N., Campari, A., Parhizkar, T., Paltrinieri, N. (2023). Artificial Intelligence for Safety and Reliability: A Systematic Review Focusing on Machine Learning. Journal of Loss Prevention in the Process Industries. *Submitted for publication*.

II. Tamascelli, N., Solini, R., Paltrinieri, N., Cozzani, V. (2022). Learning from major accidents: A machine learning approach. Computers & Chemical Engineering, 162, 107786. https://doi.org/10.1016/j.compchemeng.2022.107786.

III. Tamascelli, N., Paltrinieri, N., & Cozzani, V. (2023). Learning From Major Accidents: A Meta-Learning Perspective. Safety Science, 158, 105984. https://doi.org/10.1016/j.ssci.2022.105984.

IV. Giannini, L., Tamascelli, N., Salzano, E., Paltrinieri, N. (2023). Predicting the Consequences of Hydrogen Releases: how a Machine Learning Approach May Improve Risk-Based Inspection Planning. Proceedings of the PSAM 2023 Topical Conference on AI & Risk Analysis for Probabilistic Safety/Security Assessment & Management. *Accepted for publication*.

V. Tamascelli, N., Scarponi, G.E., Amin, M. T., Sajid, Z., Paltrinieri, N., Khan, F., Cozzani, V., (2023). A Neural Network Approach to Predict the Time-to-Failure of Atmospheric Tanks Exposed to External Fire. Reliability Engineering & System Safety. *Revised version submitted for publication*.

VI. Tamascelli, N., Arslan, T., Shah, S.L., Paltrinieri, N., Cozzani, V. (2020). A Machine Learning Approach to Predict Chattering Alarms. Chem. Eng. Trans. 82. https://doi.org/10.3303/CET2082032.

VII. Tamascelli, N., Paltrinieri, N., & Cozzani, V. (2020). Predicting Chattering Alarms: A Machine Learning Approach. Computers & Chemical Engineering, 143, 107122. https://doi.org/10.1016/j.compchemeng.2020.107122.

VIII. Tamascelli, N., Scarponi, G.E., Paltrinieri, N., Cozzani, V., (2021). A data-driven approach to improve control room operators' response. Chem. Eng. Trans. 86, 757–762. https://doi.org/10.3303/CET2186127.

IX. Tamascelli, N., Rao, H. R. M., Cozzani, V., Paltrinieri, N., Chen, T. (2023). Online Classification of Alarm Floods Using a Word2vec Algorithm. IECON 2023 – 49th Annual Conference of the IEEE Industrial Electronics Society, Singapore, 2023. https://doi.org/10.1109/IECON51785.2023.10312435.

X. Tamascelli, N., Dal Pozzo, A., Liu, Y., Cozzani, V., Paltrinieri, N. (2022). Integration between data-driven process simulation models and resilience analysis to improve environmental risk management in the Waste-to-Energy industry. Proceedings of the 32nd European Safety and Reliability Conference (ESREL 2022), Dublin, Ireland, 2022. 1409–1416. https://doi.org/10.3850/978-981-18-5183-4_R23-03-206-cd.

XI. Tamascelli, N., Dal Pozzo, A., Scarponi, G.E., Paltrinieri, N., Cozzani, V. (2023) Assessment of Safety Barrier Performance in Environmentally Critical Facilities: Bridging Conventional Risk Assessment Techniques with Data-Driven Modelling. Process Safety and Environmental Protection. https://doi.org/10.1016/j.psep.2023.11.021.

XII. Tamascelli, N., Nakhal Akel, A.J., Patriarca, R., Paltrinieri, N., Cruz, A.M. (2022). Are we going towards "no-brainer" risk management? A case study on climate hazards. Proceedings of the 16th Probabilistic Safety Assessment & Management Conference (PSAM16), Honolulu, Hawaii, 2022. ISBN: 9781713863755.

Part II

Articles

Article I.

Tamascelli, N., Campari, A., Parhizkar, T., Paltrinieri, N. (2023). **Artificial Intelligence for Safety and Reliability: A Systematic Review Focusing on Machine Learning**. Journal of Loss Prevention in the Process Industries. *Submitted for publication*.

This paper is submitted for publication and is therefore not included.

Article II.

Tamascelli, N., Solini, R., Paltrinieri, N., Cozzani, V. (2022). **Learning from major accidents: A machine learning approach**. Computers & Chemical Engineering, 162, 107786. https://doi.org/10.1016/j.compchemeng.2022.107786.

Computers and Chemical Engineering 162 (2022) 107786

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Learning from major accidents: A machine learning approach

Nicola Tamascelli [a,b,*], Riccardo Solini [b], Nicola Paltrinieri [a,b], Valerio Cozzani [b]

[a] Department of Mechanical and Industrial Engineering, NTNU, Trondheim, Norway
[b] Department of Civil, Chemical, Environmental and Materials Engineering, University of Bologna, Bologna, Italy

ARTICLE INFO

Article history:
Received 29 October 2021
Revised 17 March 2022
Accepted 25 March 2022
Available online 27 March 2022

Keywords:
Chemical process safety
Learning from past accidents
Machine learning
Classification
Severity prediction

ABSTRACT

Learning from past mistakes is crucial to prevent the reoccurrence of accidents involving dangerous substances. Nevertheless, historical accident data are rarely used by the industry, and their full potential is largely unexpressed. In this setting, this study set out to take advantage of improvements in data science and Machine Learning to exploit accident data and build a predictive model for severity prediction. The proposed method makes use of classification algorithms to map the features of an accident to the corresponding severity category (i.e., the number of people that are killed and injured). Data extracted from existing databases is used to train the model. The method has been applied to a case study, where three classification models – i.e., Wide, Deep Neural Network, and Wide&Deep – have been trained and evaluated on the Major Hazard Incident Data Service database (MHIDAS). The results indicate that the Wide&Deep model offers the best performance.

1. Introduction

1.1. Background

Learning from the past has always played a significant role in driving innovation and promoting advancements. Undoubtedly, mistakes are a part of human nature, but we all have inherent abilities to learn from them. Though, deriving a lesson and applying the acquired knowledge to avoid recurring errors is not as trivial as it may appear. History tends to repeat itself, and lessons may be ignored or forgotten (Paltrinieri et al., 2013).

Different human activities have different tolerance for errors. Within the chemical industry, significant efforts have been put in avoiding mistakes and ensuring safe operations. However, before the second half of the sixties, the words *"process safety"* and *"loss prevention"* did not exist (Kletz, 2012; Pasman et al., 1992); handling and storing dangerous substances were regulated by traditional occupational safety and good engineering practice (Hanida and Azmi, 2017). Later, a series of terrible accidents – including Woodbine (1971), Seveso (1976), Bhopal (1984), and Pasadena (1989) – highlighted the need to go beyond the existing standard and develop a different approach to prevent major accidents and their consequences (Hanida and Azmi, 2017; Pasman et al., 1992). Those unfortunate events were the driving force for the formulation and development of modern safety management programs (Hanida and Azmi, 2017).

In the ever-changing field of process safety, it has always been clear that lessons derived from past accidents would have been crucial to ensure safer design and operations (Pasman, 2009). After the investigations on the Piper Alpha disaster in 1988, Lord Cullen (1990) stated the following: "I am convinced that learning from accidents and incidents is an important way of improving safety performance". Also, the European Parliament and Council Directive 2012/18/EU (European Union, 2012) stresses the need to learn from past accidents or near misses. Still, learning, applying, and retaining the acquired knowledge is not an easy task (Jefferson et al., 1997; Pasman, 2009).

Chung and Jefferson (1998) stated that "it is widely recognized that the chemical industry as a whole does not learn from past accidents". More than ten years later, the situation has not changed much (Mannan and Waldram, 2014). Process safety has certainly improved over the last 40 years, but progress has been slow (Pasman and Fabiano, 2020). Automation, production technologies, IT, and computer simulations have witnessed extraordinary growth over the last decade. The tide of digitalization and the advent of Industry 4.0 are re-shaping the manufacturing process. Likewise, process safety is moving toward the so-called Safety 4.0 (Pasman and Fabiano, 2020). However, loss prevention and risk management struggle to keep pace, especially when it comes to learn and apply the lesson from past accidents. Accidents still happen, as evidenced by the explosion and fires that occurred at the Ming Dih Chemical factory on the 7th of July 2021 in Bangkok,

* Corresponding author.
E-mail address: nicola.tamascelli@ntnu.no (N. Tamascelli).

https://doi.org/10.1016/j.compchemeng.2022.107786

where one person was killed, more than 60 were injured, and thousands evacuated (Al Jazeera, 2021).

Undoubtedly, digitalization has brought new and effective means of information storage and transfer. The creation of digital accident databases, such as MHIDAS (AEA Technology, 1999), eMARS (European Commission, 2022), and NRC (United States Environmental Protection Agency, 2020), has made information retrieval quick and easy. However, these are hardly used by the industry (Pasman, 2009) because they are often not detailed enough or because efforts must be invested into translating case-specific information into a lesson. So, even if information has been made largely available, its potential remains unexploited. Pasman (2009) argued that the problem with learning from past accidents is not knowledge availability. Instead, the problem is that knowledge is not absorbed by individuals, nor is retained by companies. Humans do not absorb information as machines do. If a person is not interested in learning, he/she will ignore the message (Pasman, 2009). Furthermore, even if the lesson is learned, it may be forgotten in few years because "organizations do not learn from the past or, rather, individuals learn but they leave the organization, taking their knowledge with them, and the organization as a whole forgets" (Kletz, 1993).

The abundance of accident data offers a great opportunity to learn from past errors. However, the current learning process has significant limitations and appears incapable of seizing this opportunity. Therefore, there is a strong need for new tools and techniques to extract and retain knowledge from accident data. In this context, advancements in computer science and artificial intelligence have led to the construction of algorithms capable of extracting knowledge from data (Brink et al., 2016). On top of that, research has been focused on Machine Learning (ML) techniques. Currently, in the field of safety and risk assessment, Machine Learning algorithms have been proposed for fault detection and diagnosis (Xu and Saleh, 2021; Zope et al., 2019), system prognosis (Carvalho et al., 2019; Paolanti et al., 2018), diagnosis and prognosis of industrial alarm systems (Langstrand et al., 2021; Tamascelli et al., 2021; N. 2020b), and Dynamic Risk Assessment (Paltrinieri et al., 2020, 2019). Although the topic is still young and fragmented (Xu and Saleh, 2021), several authors have argued that AI and Machine Learning will play an increasingly important role in the future of process safety (Alcides et al., 2018; Lee et al., 2019; Pasman and Fabiano, 2020).

Since learning from major accidents is deeply affected by human factors, one may argue that an artificial learner would be a good support to enhance learning opportunities. Machine Learning algorithms could be trained to link accident characteristics (e.g., substances and equipment involved, release magnitude, population density) to accident consequences – e.g., the number of people involved. Such predictive models would be a quick, effective, and inexpensive means of supporting risk-based decision-making and process safety. Nonetheless, the analysis of process accident data through ML algorithms is still a largely unexplored topic. In this context, this investigation aims to contribute to this area of research by exploring the use of Machine Learning methods to analyze and extract knowledge from historical accident data. This study responds to specific and compelling needs for tools to extract knowledge from past accidents, retain and easily recall such knowledge for future use. The authors believe that the approach described in this study may provide safety managers and practitioners with advanced predictive models that may significantly improve decision making, accident prevention, and accident mitigation, representing an essential step toward Safety 4.0. What users can learn from the approach described herein is to (i) evaluate the criticality of different accident scenarios based on a set of simple and readily available features, (ii) discriminate between different criticality levels and direct efforts to prevent/mitigate high criticality scenarios, (iii) estimate the consequences of new accident scenarios without resorting to computation-intensive techniques (e.g., CFD models) and detailed modeling.

1.2. Objectives

The purpose of this study is to determine whether Machine Learning methods might be used to exploit the knowledge embedded in accident databases and predict the outcomes of new accidents and incidents. Specifically, the research focuses on classification algorithms and their ability to capture the relationship between accident features and consequences to humans in terms of people injured or killed.

There are three primary aims of this study:

- to propose and describe a methodology for the analysis of accident databases through Machine Learning classification models;
- to describe how these models might be used to predict the severity category of process accidents;
- to test and compare different models, highlighting the advantages and limitations and discussing optimization strategies.

In order to achieve objectives 1 and 2, a generic framework has been developed, which might be promptly adapted for use on different accident databases and ML models. The methodology has been applied to a test case in order to reach the third objective. Specifically, three classification models (i.e., Wide, DNN, and Wide&Deep) have been trained and tested on a generic accident database – i.e., the Major Hazard Incident Data Service (MHIDAS).

1.3. Related works

Several studies have proposed Machine Learning methods to extract safety-critical information from historical data and predict the outcomes of accidental events. For instance, Sarkar et al. (2020) used six different classification algorithms to predict injury severity of accidents that occurred in a steel manufacturing plant; investigation reports and inspection reports collected in a time period of 3 years are used to train and evaluate the models. Phark et al. (2018) discussed the application of naïve Bayes classifiers and Multi-Layer Perceptron for predicting the issuance of emergency evacuation orders after the release of toxic substances. A method for the semiautomatic retrieval of Natech scenarios from the National Response Center database has been proposed by Luo et al. (X. 2020), which employed Long Short-Term Memory and Convolutional Neural Network as classification models.

Also, several studies focused on Natural Language Processing (NPL) and Machine Learning methods to analyze accident narratives and extract useful information. For example, Kurian et al. (2020) proposed a Machine Learning approach to classify unstructured accident reports into basic accident types (e.g., "health/safety", "leak/spill", "operation"). Also, they proposed NPL algorithms to derive a more informative and helpful set of keywords from raw accident reports. Jing et al. (2022) used Word2Vec (Mikolov et al., 2013) and bidirectional Long Short Term Memory neural network (Bi-LSTM) with an attention mechanism to (i) analyze the correlation between accidents and extract accident precursors, causes, and high-frequency types of chemical accidents, and (ii) automatically classify accident reports into their respective accident type (i.e., "fire", "explosion", "poisoning", and "other"). A proprietary dictionary was developed to improve word segmentation and classification performance. Bi-LSTM was also used by Wang and Whao (2022) to extract and estimate the frequency of contributory factors from confined space accident reports. The authors used BERT algorithm to build word embedding and a BiLSTM with a conditional random field (CRF) to classify accidents

based on their contributory factors (e.g., improper tool, gas detection, inadequate supervision). Since the approach is fully supervised, manual intervention by experts is needed to extract fundamental characteristic of accidents and their contributory factors. Instead, a semi-supervised approach was proposed by Ahadh et al. (2021) to automatically classify accident reports from different domains based on user-defined topics. The approach is domain-independent and requires minimal human intervention. The authors proposed to extract domain-relevant keywords from a domain corpus (e.g., guidelines, standard manuals, scholarly articles, and Wikipedia pages) and identify the accident cause (e.g., "External force", "Equipment Failure", "Incorrect Operation") or other user-defined accident characteristics from accident narratives. A guided version of the Latent Dirichlet Allocation (Jelodar et al., 2019) algorithm was used to extract the accident features.

Although the investigations described above represent a valuable attempt to extract information from accident reports, their intent and methodology differ significantly from the approach described in this study. For instance, unstructured accident narratives are analyzed, while this study focuses on structured accident databases. In addition, the primary aim of those studies is to automate the extraction of key pieces of information from unstructured text and, therefore, to reduce the need for manual intervention by experts, which is time-consuming and expensive. Instead, the algorithms proposed in our study are not designed to extract generic accidents characteristics (e.g., the substance involved, the cause, the amount of substance released) since this information is already available in the structured database used for the analysis. Instead, this study seeks to extract higher-level knowledge, which experts cannot extract by simply reading accident reports. Specifically, the proposed algorithms aim to capture and quantify the relationship between accident features and consequences in terms of people killed and injured. In other words, the objective is to extract knowledge from historical accident reports to build a mapping between accident features and accident consequences. The method presented in this study can be used to perform predictions; given a short list of accident features, the model returns the number of people involved in the accident. Instead, the studies described above take a large text (i.e., accident narratives) as an input and extract key information. In other words, their aim is not knowledge extraction to predict the outcome of accidental events; they just mimic the knowledge discovery process of a human reader.

Similar to this study, Chebila (2021) proposed a Machine Learning-based method to predict whether accidents involving dangerous substances will cause damage to humans, the environment, and material assets. Specifically on the consequences on people, a set of binary classifications was performed using six different models in order to predict the occurrence of at least one injured or killed. The study concluded that Random Forest ensures the best performance. Also, Neural Networks provided good results, but they proved to be less effective than Random Forest in dealing with unbalanced datasets. The investigation by Chebila (2021) shares some features with this study, such as the overall intent and the approach; however, there are also significant differences. For instance, the approach proposed by Chebila (2021) did not distinguish between injuries and killed, while the present study considers these outcomes separately. Furthermore, the present study uses a set of multiple discrete outcome variables to differentiate accidents according to their severity (i.e., from 1 to 10 killed, from 11 to 100 killed, etc.). On the other hand, a greater number of classification models were used and tested by Chebila (2021), which also considers more targets (i.e., the environment and material assets). Finally, different databases are used; eMARS was used by Chebila (2021), while this study focuses on MHIDAS.

The chemical and process industry is not the only industrial sector that has been involved in this line of research. For example, Gerassis et al. (2020) proposed the use of a Multiple Correspondence Analysis in conjunction with Bayesian Networks to classify mining accidents as fatal or non-fatal. The approach was tested on an occupational accident database and allowed the identification of the factor contributing most to the accident severity. A different approach has been developed by Yedla et al. (2020) to predict the number of days away from work after a mining accident. The method makes use of regression and classification models – such as Logistic Regression, Decision Trees, Random Forests, and Artificial Neural Networks – to predict the number of days away from work and the degree of injury. Similarly, Choi et al. (2020) demonstrated that accident data could be used to build classification models for the prediction of the likelihood of mortality in the event of an accident in a construction site.

Several studies have also focused on the transportation industry. In the analysis proposed by Zhang et al. (2018), four different Machine Learning algorithms were compared based on the ability to predict the severity of crashes that occurred in freeway segments. The study concluded that Machine Learning models produce better performance than traditional statistical methods in this specific task. Also, the results suggested that Random Forest and K-Nearest Neighbors were the best models. Assi et al. (2020) investigated the use of Feed Forward Neural Networks and Support Vector Machine to predict the severity level of traffic crashes. In addition, the study investigates the use of fuzzy c-means clustering to enhance the model prediction capabilities. A similar approach was proposed by Wahab and Jiang (2019), which focused on the prediction of motorcycle crash severity using Decision Trees, Random Forest (RF), and Instance-Based Learning. Also, Burnett and Si (2017) demonstrated the use of Machine learning classification techniques to predict the levels of injuries and fatalities in aviation accidents. The analysis concluded that Artificial Neural Networks performed better than the other models.

Overall, a search of the literature revealed that the attention of the scientific community has only recently focused on the application of Machine Learning methods for accident severity prediction. The idea of utilizing process data to update the risk picture has already been proposed in past works – e.g., (Landucci and Paltrinieri, 2016). However, the growing body of research on Machine Learning methods indicates that the approach may play a significant role in the future of safety assessment and management in several areas. Also, the search revealed that there is a notable paucity of studies investigating the application of such methods to accidents involving dangerous substances. In this context, this is the first study to propose a Machine Learning-based method to predict the consequences of accidents involving dangerous substances in terms of people killed and injured. Only one similar study was found in the literature (Chebila, 2021), which only considered whether or not the accident damaged people, therefore lacking the level of detail provided in this investigation. In addition, this study makes use of a set of multiple discrete outcome variables to estimate the number of people involved, therefore providing a much more detailed and valuable output.

1.4. Outline

The paper is organized into 7 Sections. Section 2 presents the methodology, including the pre-processing of accident data and the Machine Learning simulations. The test case is described in section 3, which also includes a description of the database used for the simulations. Section 4 presents a selection of the most representative findings, while the full results are provided separately in the supplementary material. Results are discussed in section 5, which

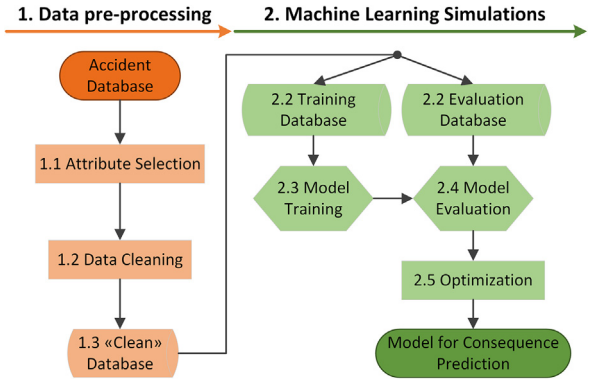


Fig. 1. Methodology workflow. Colors represent two main stages: Data pre-processing (orange), and Machine Learning Simulations (green).

also highlights the limitations of the study and provides suggestions for future works. Finally, conclusions are drawn in section 6.

2. Method and data

The overall workflow of the methodology developed to analyze and extract knowledge from accident databases through Machine Learning techniques is outlined in Fig. 1. The method involves two main steps: data pre-processing and Machine Learning simulations. In the first step, raw accident data are converted in a suitable format for Machine Learning analyses. Next, part of accident data is used to train the Machine Learning classification algorithm. Finally, the trained model is used to predict the severity of new events. Predictions are compared with expectations in order to assess the model performance and discuss optimization strategies. A detailed description of the methodology is provided in the following 2 sections.

The method has been demonstrated on a test case study using three classification models, namely Wide, Deep Neural Network, and a hybrid Wide&Deep model. The algorithms were trained and evaluated separately on the same datasets, and their performance was compared to highlight their strength and limitations. It is worth mentioning that this is the first study that takes advantage of these algorithms to predict the consequences of process accidents with a high level of detail. Also, this is the first study that investigates the use of a hybrid Wide&Deep model for the analysis of accident data.

2.1. Accident database and features selection

Accident data are extracted from the data source and stored in a convenient format, such as a CSV file. The database has a matrix-like shape where each row represents an event and each column an attribute of the event (e.g., the date, the substance involved, the incident type).

Some of the attributes included in the database may not be meaningful or useful for the analyses; these attributes must be removed (step 1.1 in Fig. 1). In general, the database should contain only attributes that link event characteristics to event consequences. After removing unnecessary attributes, the database must be prepared for the Machine Learning simulations (step 1.2 in Fig. 1). This task requires three steps:

- Missing data must be imputed or removed because most Machine Learning models cannot process null values. Different techniques have been developed to impute or remove missing values based on the type and characteristics of the data (i.e., numerical or categorical, random or not random). An overview of the most used methods can be found in Brink et al. (2016), Bruha (2017), and Makaba and Dogo (2019). In this study, missing values have been substituted by the user-defined string "Na". This should allow the model to deal with uncertainty and learn the impact of missing values on the outcome measure.
- Attributes that may contain more than one entry must be split so that each column in the database contains only one entry.
- The attributes indicating the Number of People that are Injured (NPI) and Killed (NPK) must be converted into their respective severity categories. To this end, a set of consequence categories are considered to reflect severity categories used by risk matrices and other risk analysis methods (ARAMIS project team, 2004) (Table 1).

Table 1
Accident consequence categories.

Category	Description
NO	no killed/injured
1 - 10	from 1 to 10 killed/injured
10 - 100	from 10 to 100 killed/injured
- 1000	from 100 to 1000 killed/injured
> 1000	more than 1000 killed/injured

After these steps, a clean version of the original database is obtained, which is eventually used for the simulations. The Machine Learning algorithms are trained to classify accidents into one of the categories described in Table 1, therefore predicting the severity of accidental events with a high level of detail.

2.2. Machine learning simulations

Machine Learning (ML) refers to a class of computer algorithms designed to gain experience from data and leverage the acquired knowledge to perform accurate predictions, reveal correlations between variables, and identify hidden patterns and trends (Brink et al., 2016; Hastie et al., 2009). In other words, Machine Learning concerns training a machine to learn from past understanding (Schottenfels, 2019).

There are three macro-categories of Machine Learning algorithms: Supervised Learning, Unsupervised Learning, and Reinforcement Learning (Murphy, 2012). Supervised Learning is used when the problem involves the prediction of an outcome measure based on one or more input variables (Hastie et al., 2009). Instead, if no output measure is applicable, Unsupervised Learning algorithms may be used to analyze input data and reveal relationships and patterns with little or no human intervention (IBM Cloud Education, 2020; Jukes, 2018). In Reinforcement Learning, the learner (e.g., an industrial robot) is not passively analyzing input data; instead, it collects data from the environment through a set of actions, and a reward system is used to guide the learning process (Stone, 2017).

In this study, both the input (i.e., the features of an event) and the outcome measure (i.e., the event severity) are available and reported in the data source. Therefore, Supervised Learning algorithms are a natural choice. Further, the objective of this study is to categorize (i.e., classify) accidents based on their severity of consequences, which may be expressed in terms of the number of people that are killed or injured in the event - for this reason, two distinct sets of simulations are performed. Therefore, the problem is a classification task. However, a regression approach may also be possible and should be investigated by further research.

2.3. Classification: training and evaluation

The aim of a classification algorithm is to classify *objects* into two or more categories (Drummond, 2017). An *object* is described by a set of features (i.e., meaningful attributes of the object, say

X) and one label (i.e., its category, say Y); in this study, releases of dangerous substances are the objects.

At first, the clean database is divided into two parts: the training database and the evaluation database (step 2.2 in Fig. 1). The former comprises 80% of the events, and the remaining part (20%) forms the latter. Next, the training database is fed to the algorithm, which tunes the internal parameters of a function f in order to find the optimal mapping between features and corresponding labels (James et al., 2013). The function f is also called the *model* of the Machine Learning algorithm (TensorFlow.org, 2020a).

$$Y \approx f(X) \tag{1}$$

Where:

- $X = N \times$ M matrix of the features. N is the number of objects, and M is the number of features;
- $Y = N \times 1$ vector of the labels;
- $f =$ function with tunable parameters.

This phase is the so-called *training* phase (2.3 in Fig. 1). Next, unlabeled objects are fed to the trained model, which predicts the corresponding labels according to the following equation.

$$f(X_i) = \hat{Y} \tag{2}$$

Where:

- $X_i = 1 \times$ M vector of the features of the unlabeled object i;
- $\hat{Y} =$ label probabilities produced by the model for the object i.

Finally, predicted labels are compared with the true labels to evaluate the performance of the model. This phase is the so-called *evaluation* phase (steps 2.4 in Fig. 1). The batch of objects used to evaluate the algorithm is the evaluation database.

It is worth noting that the output of the model (i.e., $\hat{Y}_i$) is not a single label but a vector that contains the label probabilities (James et al., 2013). In other words, if K different categories are possible, $\hat{Y}_i$ is a $K \times 1$ vector whose elements represent the probability of each category. In order to convert label probabilities into one predicted label, a probability decision threshold is used (Google, 2020a), which is often 0.5 by default.

3. Models

Different models are available to perform a classification task. In this study, a Linear model, a Deep Neural Network, and a hybrid Wide&Deep model are used to demonstrate the approach.

3.1. Linear model

The Linear model represents the labels as a linear combination of features (James et al., 2013). Therefore, Eq. (1) can be written as:

$$Y \approx \beta_0 + \sum_{j=1}^{M} \beta_j X_j \tag{3}$$

Where:

- $Y =$ label;
- $\beta_0 =$ bias;
- $X_j =$ a feature;
- $\beta_j =$ weight of the j-th feature.

Linear models are robust, fast, easy to interpret, and suitable for analyzing large datasets (Brink et al., 2016; Hastie et al., 2009; James et al., 2013). On the other hand, they cannot capture nonlinear relationships between features. Also, linear models cannot infer the impact of combinations of features that have not occurred in the past (Cheng et al., 2016).

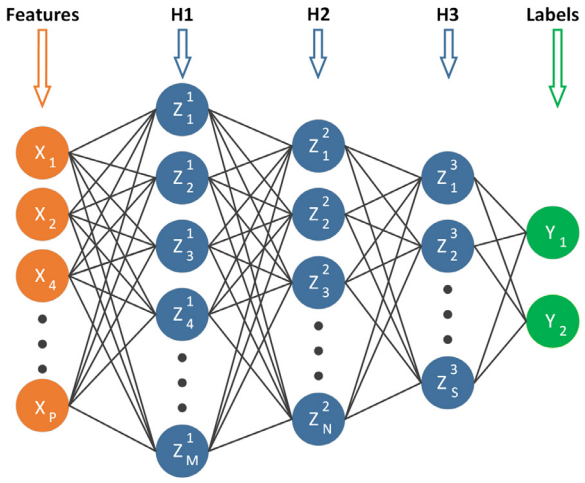


Fig. 2. Schematic representation of a Deep Neural Network. Orange, blue, and green circles represent input features (X_i), hidden units (Z_i^j), and labelsY_k. Adapted from Tamascelli et al. (2020a).

3.2. Deep neural network

Deep Neural Networks (DNNs) are directed acyclic graphical models consisting of densely interconnected units (Goodfellow et al., 2016). A visual representation of a DNN is shown in Fig. 2.

In these models, the features of an object (orange circles in Fig. 2) are converted into label probabilities (green circles in Fig. 2) through a series of linear combinations and nonlinear transformations (Hastie et al., 2009). In between the Input and Output layers, a series of interconnected *hidden units* (blue circles in Fig. 2) is arranged into one or more *hidden layers* (e.g., H1, H2, and H3 in Fig. 2). The unit of a generic hidden layer H_i is obtained by a nonlinear transformation of the linearly combined units of the previous layer. In this study, the Rectified Linear Unit (TensorFlow.org, 2020b) is used to perform the nonlinear transformation. Further details and formulas behind Neural Networks may be found in Goodfellow et al. (2016) and Hastie et al. (2009).

DNNs have good generalization capabilities and can capture nonlinear relationships between features (Goodfellow et al., 2016). As a drawback, they are sensitive to poor quality input data and are prone to overfitting and overgeneralization (Brink et al., 2016; Goodfellow et al., 2016; Hastie et al., 2009). In addition, the computational cost required for training a DNN is larger if compared to simpler models (Goodfellow et al., 2016).

3.3. Wide&Deep

In an attempt to combine the advantages of the Linear and Deep models, Cheng et al. (2016) developed the Wide&Deep model, whose structure is displayed in Fig. 3.

The model comprises a Linear part (top of Fig. 3) and a Deep part (bottom of Fig. 3). During the training phase, the Linear and Deep models are *jointly trained* –i.e., predicted labels (green circles in Fig. 3) are obtained by combining the outputs of both models, and the weights of the models are optimized simultaneously (Cheng et al., 2016). Usually, the linear part of the model takes as in input a small set of crossed-features (Cheng et al., 2016), which are synthetic features obtained by taking the cartesian product of two or more features (Google, 2020b). On the contrary, the Deep part uses all available features (X_D in Fig. 3). Hence, the Deep part is a full-size DNN model, while the Linear part integrates and "complements the weaknesses of the deep part with a small number of cross-product" (Cheng et al., 2016). As an example, the fea-

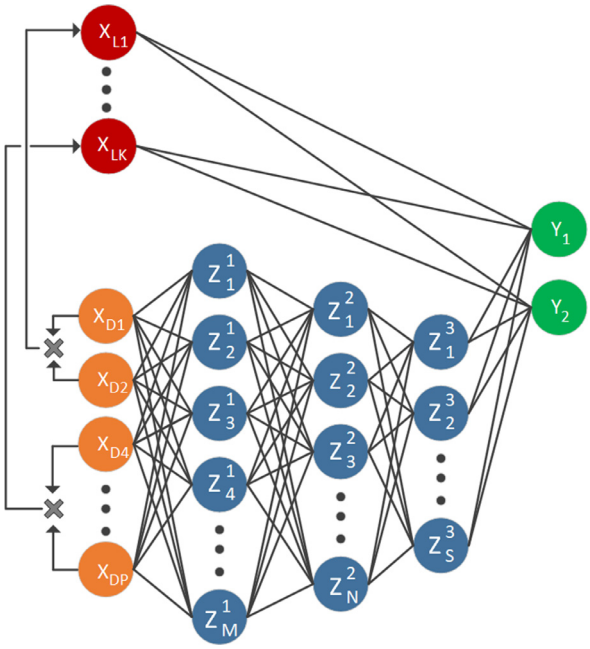


Fig. 3. Schematic representation of a Wide&Deep model, which consists of a deep part (bottom) and a wide part (top). The deep part is a DNN and takes as an input a full set of features (X_{Di}). The wide part is a Linear model and takes as an input a small set of crossed-features (X_{Li}).

ture X_{L1} in Fig. 3 is obtained by crossing X_{D1} and X_{D2}. In general, the hybrid nature of the Wide&Deep model ensures good memorization (Linear part) and generalization (Deep part) capabilities.

3.4. Performance metrics

The performance of a Classification algorithm is assessed during the evaluation phase. For instance, the classification may consider classes "Y" and "N", respectively positive and negative. Whenever the model predicts the class of an object, there are four possible outcomes:

- TP = True Positive –i.e., predicted label = Y, true label = Y;
- TN = True Negative –i.e., predicted label = N, true label = N;
- FP = False Positive –i.e., predicted label = Y, true label = N;
- FN = False Negative –i.e., predicted label = N, true label = Y.

The sum of True Positives and True Negatives represents the number of correct predictions, while the sum of False Positives and False Negatives indicates the number of wrong predictions.

True Positives, True Negatives, False Positives, and False Negatives are used to obtain three performance indicators:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4}$$

$$Precision = \frac{TP}{TP + FP} \tag{5}$$

$$Recall = \frac{TP}{TP + FN} \tag{6}$$

Accuracy represents the fraction of objects that have been correctly classified. Precision indicates the success rate of a positive prediction. Recall denotes the fraction of actual positives that have been correctly identified.

Accuracy alone is not informative if the problem involves the identification of rare classes –i.e., when the dataset is class imbalanced (Google, 2020c); in these situations, Precision and Recall are more representative of the model performance (Google, 2020d). In addition, if the cost for a False Negative is higher than the cost for a False Positive, the Recall is the most meaningful metric.

Rather than considering Precision and Recall individually, one may aggregate them into the so-called F-score (Chinchor, 1992).

$$F_\beta = \left(1 + \beta^2\right) \cdot \frac{Precision \cdot Recall}{\left(\beta^2 \cdot Precision\right) + Recall} \tag{7}$$

Where:

- β = non-negative real number.

If $\beta = 1$, the score represents the harmonic mean between Precision and Recall (Han et al., 2012). If $\beta > 1$, the score is Recall oriented (Sasaki, 2007), meaning that the Recall is considered to be β times more important than Precision.

Finally, it is worth mentioning that the metrics and indicators presented depend on the probability decision threshold (section 2.2.1). In fact, the decision threshold might be tuned in order to optimize the model (step 2.5 in Fig. 1) (Google, 2020a). For example, if the decision threshold is lowered, the model may produce more positive predictions. As a result, the Recall might increase, but the Precision might decrease (Scikit-learn.org, 2020). In fact, actions aimed at increasing Recall often lower the Precision, and vice-versa (Google, 2020d).

A convenient means of displaying the effect of the decision threshold is the Precision-Recall curve –i.e., a plot where each point represents the couple Precision vs. Recall at a specific decision threshold (Murphy, 2012). A convenient means of summarizing the information in the Precision-Recall curve is the area under the curve (AUC P-R) (Murphy, 2012), which takes values between 0 and 1. Being independent on the decision threshold, the AUC PR is considered a more comprehensive indicator of the model performance if compared with Accuracy, Precision, and Recall. In general, a large AUC P-R value indicates good performance (Scikit-learn.org, 2020).

3.5. Test case analysis

An accident database was used to validate the proposed methodology and compare the performance of the models. A brief description of the database and Machine Learning simulations are provided in the following sections.

4. MHIDAS

Founded in 1986 by the UK Safety and Reliability Directorate (SRD) and the Health and Safety Executive (HSE), the Major Hazard Incident Data Service (MHIDAS) is an accident database that contains records of more than 8900 incidents involving hazardous materials (AEA Technology, 1999). Initially, the database included only events that involved the ignition of flammable substances. Later, the scope was widened to include toxic gas dispersion and those incidents that "have the potential to produce an off-site impact" (AEA Technology, 1999). The database had been managed and updated by AEA Technology until the early 2000s, when it was eventually decommissioned. Incident data are entirely drawn from public domain sources, such as accident reports, newspapers, and journals (Harding, 1997); this ensures the widest dissemination but, as a drawback, it raises issues of missing, incomplete, or biased information and inconsistencies (Harding, 1997).

4.1. Attributes distribution

Accidents in MHIDAS are described by a list of 22 different attributes. Some attributes have a strong link, such as the type of substance released and its quantity. Other attributes may have

Table 2
Accident attributes used in the Machine Learning simulations. * marked attributes are Multiple entry fields (e.g., "Release" AND "Pool Fire" for IT, "Flammable" AND "Toxic" for MH).

Attribute	Description	
DA	Date	Date of the incident.
LO	Location	Town, region, and country of the incident
GC	General Cause	The general cause - or causes - which triggered the event (e.g., Mechanical failure, Human Error)
SC	Specific Cause	The specific cause - or causes - which triggered the event (e.g., Brittle fracture, Overpressure, Fire)
GOG	General Origin	Area of the plant where the incident originated from (e.g., Process, Storage, Warehouse)
SOG	Specific Origin	Equipment that originated the incident (e.g., Pump, Vessel, Pipeline)
MN	Material Name*	Names of dangerous substances involved in the incident
MH	Material Hazard*	The hazard class of the substances involved (e.g., Toxic, Explosive, Corrosive, Oxidizing)
MC	Material Code*	Four-digit code of the substance involved
QY	Quantity	The amount of substances released (tons)
IS	Ignition Source	Type of ignition source (e.g., hot surface, flares, boilers)
IT	Incident Type*	Incident typology (e.g., Release, BLEVE, Physical Explosion)
NPE	Evacuated	Number of people that are evacuated
PD	Population Density	Population density in the Area (i.e., "Rural" for low - sparse population, "Urban" for highly populated Area)
NPI	Injured	Number of people that are injured in the incident
NPK	Fatalities	Number of fatalities in the accident

a weaker link, such as the date and the location of the accident. However, date and location may be an indirect measure of the socioeconomic status of the area. As is known, industrializing and impoverished countries are more exposed to industrial risk due to intense urbanization, disordered industrialization, and less elaborate safety measures (Souza et al., 1996). For example, the Bhopal disaster (Kalelkar, 1988) and the recent Beirut explosion (Pasman et al., 2020) are infamous events where unsatisfactory safety measures and uncertain emergency planning had contributed to the accident. Therefore, the date and location have not been removed from the database.

In this study, six attributes were discarded during the Feature Selection phase. As a result, only the attributes listed in Table 2 have been used for the analyses. The reason for this choice is availability and completeness; that is, these attributes are reported natively in the accident database used to perform the analysis, and they provide a synthetic but exhaustive description of the accident, from its causes to consequences.

The first 14 attributes in Table 2 represent the input of the Machine Learning models (i.e., the features). Instead, the last two attributes are the outputs of the models.

It is worth examining the frequency distribution of some of these attributes more in detail because the performance of the Machine Learning models is deeply affected by the characteristics of the dataset. The frequency distribution of attributes General Origin, Incident Type, General Cause, Specific Cause, Material Name, and the number of people affected (i.e., NPI and NPK) is shown in Fig. 4.

The figure indicates that most of the incidents involved releases or explosions and subsequent fires (Fig. 4b), which often occurred during the transportation of the substance (Fig. 4a). Also, a significant part of the incidents originated in the process and storage areas of chemical plants (Fig. 4a). The most frequent incident causes are "Impact", "Mechanical", and "Human" failures Fig. 4e. Also, it is worth noting that the missing value frequency ("Na") is high for the attributes General Cause and Specific Cause. This may be due to the public domain nature of the database because such technical and sector-specific information is rarely reported in newspapers and journals. Finally, Fig. 4f indicates that most of the incidents in the database did not cause any injured or killed. Also, the number of records in the database decreases as a larger number of people involved is considered; that is, the rarity of events increases with the severity of the consequences. Furthermore, incidents that resulted in injuries are more frequent than those that caused fatalities. It is also worth mentioning that the consequence category "> 1000" is not shown in Fig. 4f because there are only 5 and 13 accidents with more than 1000 killed or injured, respectively; therefore, the box would not have been visible.

4.2. Simulations

The Machine Learning models have been trained and tested on MHIDAS as described in section 2.2. Specifically, the database has been split into a training dataset containing 7100 events and an evaluation dataset containing 1872 events. Next, two sets of binary classifications have been performed. The first set focuses on predicting the number of people that are killed in the accident (i.e., NPK), while the second focuses on the number of people that are injured (i.e., NPI). Within each set of simulations, distinct binary classifications were performed for each consequence category and model using different iteration steps, which represent the number of times the training dataset is fed to the model during the training phase (TensorFlow.org, 2020c). A large number of iteration steps simulate a more extensive database, and therefore may improve the learning phase. However, the model may overfit the training data if a large number of iteration steps are used (TensorFlow.org, 2021). In this study, a number of iteration steps equal to 200, 2000, 20,000, and 200,000 were used in order to assess the effect of different iteration steps on the model performance.

5. Results

The full results of the study are provided in the supplementary material. A selection of the most representative findings is displayed in Fig. 5 and Fig. 6, which show the AUC P-R, Recall, Accuracy, and Precision for the category NPI and NPK, respectively. A decision threshold equal to 0.5 is used to obtain the Accuracy, Precision, and Recall values.

The results shown in Fig. 5 and Fig. 6 have been obtained using the iteration steps displayed in Table 3. The simulations have been selected based on the AUC PR value – i.e., the number of steps that led to the highest AUC PR has been selected and shown in this section. If two simulations had comparable AUC PR values, the one with the highest Recall has been chosen.

6. Discussion

This paragraph is divided into two sections. In the first section, the feature selection phase will be described more in detail; specifically, the choice of the attributes listed in Table 2 will be discussed, the limitations of the approach will be highlighted, and

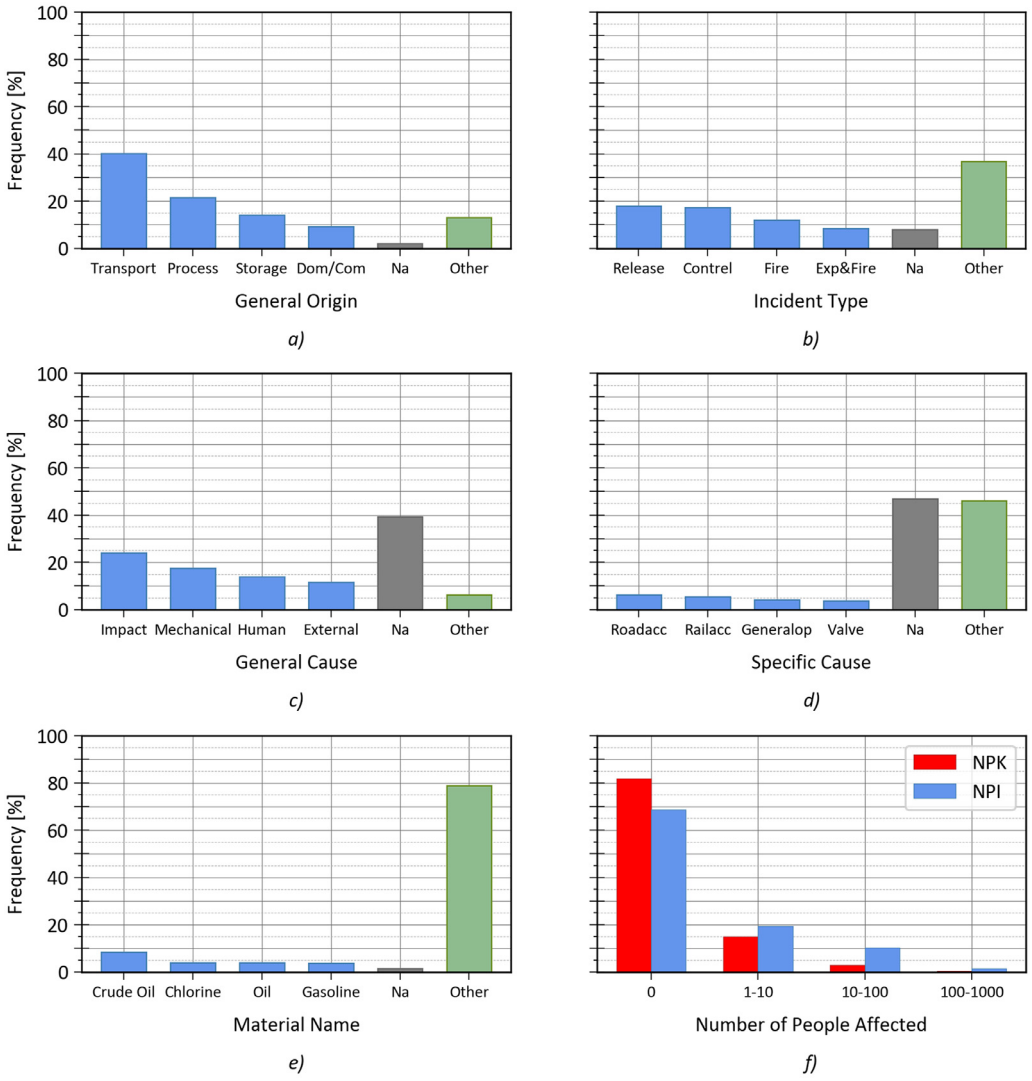


Fig. 4. Frequency distribution of the attributes GOG (a), IT (b), GC (c), SC (d), MN (e), NPK and NPI (f). "Na" refers to missing values, "Other" refers to attribute codes that have not been represented in the figure.

Table 3
Number of iteration steps used to obtain the metrics in Fig. 5 (i.e., Number of People that are Injured "NPI") and Fig. 6 (i.e., Number of People that are Killed "NPK").

Category	Models	NO	1 – 10	10 – 100	100 – 1000	>1000
NPI (Fig. 5)	Wide	200	20,000	2000	2000	200,000
	Deep	2000	2000	20,000	200	200
	Wide&Deep	200	20,000	200	2000	20,000
NPK (Fig. 6)	Wide	20,000	20,000	200	200,000	200,000
	Deep	2000	2000	20,000	200	200
	Wide&Deep	2000	200,000	2000	200	200,000

recommendations will be drawn. In the second part, the discussion of the results will be specifically addressed.

6.1. Attributes selection and the need for a standardized taxonomy

As previously stated, the reasons behind the selection of the attributes described in section 2.1 are convenience and completeness. Regarding the last motivation, it is worth analyzing the role of each attribute in more detail. To this end, a graphical representation – such as a bow-tie diagram – can be a helpful support. Bow-ties are clear and direct means of indicating the causal relationships between *Undesirable Events* (i.e., the causes of an incident), *Critical Events* (i.e., Top Events), and *Major Events* (i.e., Thermal radiation, Overpressure, Toxic effects, Missiles). Taking the generic Bow-Tie structure proposed by the ARAMIS project as a reference (ARAMIS project team, 2004), it might be argued that the attributes described in Table 2 can be mapped into the diagram so that each intermediate event is described by one or more attributes. Fig. 7 clarifies this insight. The Bow-Tie in the figure is divided into nine different intermediate events, as suggested by the ARAMIS framework. The codes describing the names of these events are shown at the top of Fig. 7. The attributes used in the Machine Learning simulations (Table 2) may be used to describe each event of the Bow-Tie, as shown at the bottom of Fig. 7.

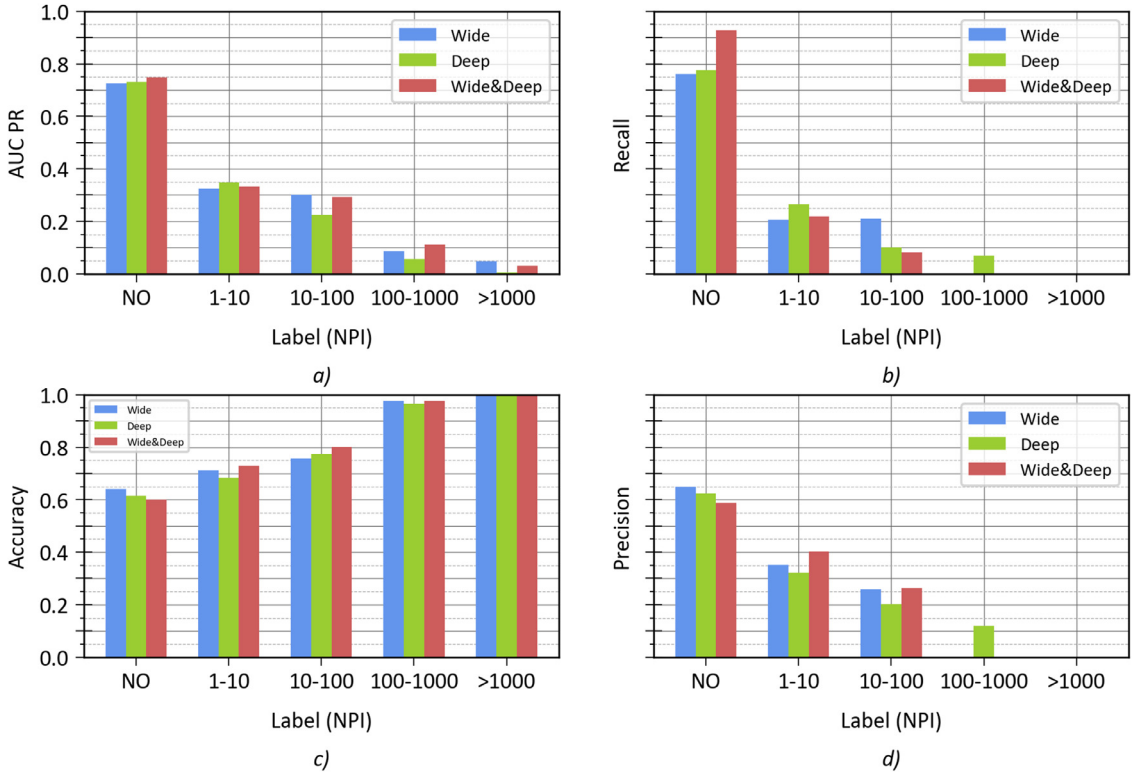


Fig. 5. Area Under the Curve Precision-Recall (AUC PR) (a), Recall (b), Accuracy (c), and Precision (d) obtained from a small selection of simulations for the category "Number of People that are Injured" (NPI). Labels are represented on the x-axis. Recall, Precision, and Accuracy are obtained at threshold = 0.5.

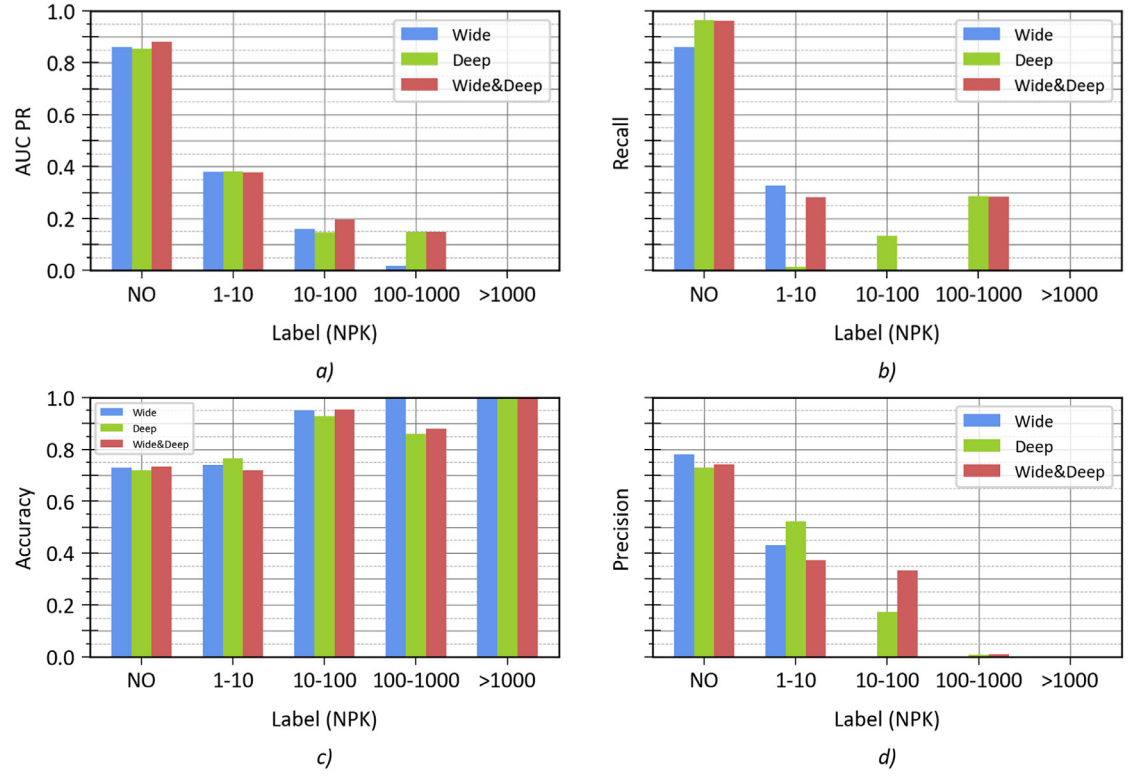


Fig. 6. Area Under the Curve Precision-Recall (AUC PR) (a), Recall (b), Accuracy (c), and Precision (d) obtained from a small selection of simulations for the category "Number of People that are Killed" (NPK). Labels are represented on the x-axis. Recall, Precision, and Accuracy are obtained at threshold = 0.5.

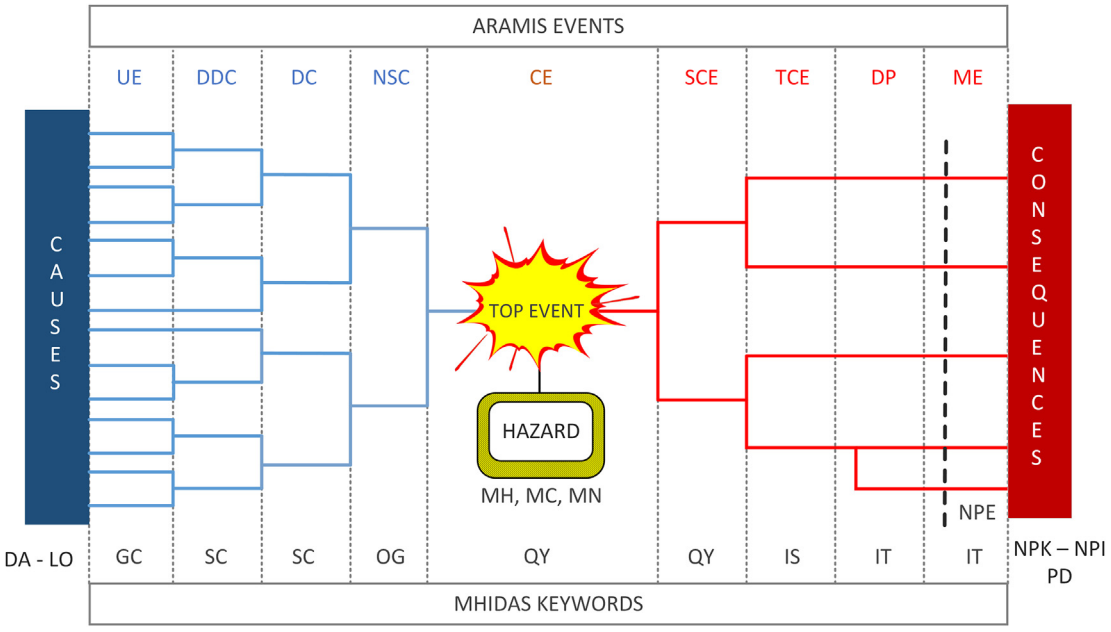


Fig. 7. Schematic representation of a Bow-Tie diagram. Names of intermediate events (top side) are defined according to MIMAH methodology (ARAMIS project team, 2004). Database attributes listed in Table 2 are associated with each intermediate event (bottom side). The bold dashed line indicates that the Number of People that are Evacuated (NPE) may act as a safety barrier between the Major Event (ME) and the accident consequences.

The attributes Date (DA) and Location (LO) may provide background information for the accident causes; therefore, they are represented at the bottom-left side of the diagram. Proceeding to the right, the attribute General Cause (GC) may describe the Undesirable Event (UE) that started the incident. The Specific Cause (SC) may be associated with both Detailed Direct Causes (DDC) and Direct Causes (DC). General and Specific Origin (GOG and SOG) may describe the Necessary and Sufficient Cause (NSC). The Critical Event (CE) may be defined by the type of substances involved (i.e., Material Hazard, Material Code, and Material Name) and by the quantity of substance released (QY). On the event tree side, the attributes Quantity (QY), Ignition Source (IS), and Incident Type (IT) may be used to describe the events Secondary and Tertiary Critical Events (SCE and TCE), Dangerous Phenomena (DP), and Major Event (ME). The effect of Major Events on humans are described by the attributes Population Density (PO), Number of People that are Injured (NPI), and Number of People that are Killed (NPK), which are on the rightmost side of the diagram. Finally, the Number of People that are Evacuated (NPE) may indicate the effectiveness of the Emergency Response Plan. For this reason, NPE is represented in Fig. 7 as a safety barrier that mitigates the harmful effects of a Major Event.

In conclusion, the attributes provide a synthetic but rather exhaustive description of the incident, from its causes to consequences on humans. Therefore, it appears reasonable to use this set of attributes for the Machine Learning simulations. However, there is not a globally accepted standard methodology for recording accidents into digital databases. That is, different databases use different sets of attributes and taxonomies; this implies that prior to applying the method described in this work to other accident databases, one must convert attributes and taxonomies to match those described in Table 2, which is a difficult and time-consuming task. Instead, one may decide to use a different set of attributes and taxonomy, but the issue will not be solved because the model will still be limited to one of many taxonomies. For these reasons, it would be advisable that institutions and academics discuss and propose a standardized system to record accidents, incidents, and near misses into digital databases. Such a harmonized recording system would terribly improve the use of advanced analysis methods whose potential is not fully exploited due to the differences between existing databases.

In this work, MHIDAS has been used despite being decommissioned and no longer updated. The authors believe that this choice does not affect the validity of the analysis since the database has a well-organized and rational structure and contains records of a large number of incidents and accidents that occurred worldwide in more than a decade. Indeed, there are more recent and updated databases that it may be beneficial to analyze, such as eMARS (European Commission, 2022), ARIA (Bureau for Analysis of Industrial Risks and Pollutions, 2022), ZEMA (Bundesministerium für Umwelt Naturschutz Bau und Reaktorsicherheit, 2022), and FACST (Unified Industrial and Harbour Fire Department, 2022). However, their use would not guarantee more reliable and accurate results. The exhaustive and informative set of attributes used in MHIDAS simplifies the analyses and avoids time-consuming and expensive data pre-processing. Instead, different datasets may require extra efforts to extract the most relevant features from limited native accident representation.

7. Discussion of results

The results reported in Fig. 5 and Fig. 6 suggest that each performance metric follows a particular trend. Specifically, the AUC PR appears to decrease as the task involves the identification of accidents with an increasing number of people involved, as shown in Fig. 5a and Fig. 6a. The trend might be explained by considering the rarity of events with a large number of people involved. In fact, the frequency distribution of the attributes NPI and NPK (section 3.1.1) highlights that the number of events in the database decreases as the number of people that are injured or killed increases. As a result, the performance of the models may have degraded because there are fewer chances to learn from events that have never or rarely occurred.

A similar trend is observed for the metrics Precision and Recall. The only exception is the label “100 – 1000” of the category NPK (Fig. 6b), for which the Deep and Wide&Deep models

Table 4
Example of two similar accidents that led to different classification results. Only the most relevant features are displayed.

ID	MN1	IT1	IT2	GOG1	SOG2	GC1	GC2	IS1	Result
1	Crude Oil	Contrel	Fire	Transport	Pipeline	Mechanical	Human	Electric	TP
2	Crude Oil	Contrel	Na	Transport	Pipeline	Mechanical	Na	Nonignite	FN

produced a Recall higher than the label "10 – 100". The trend might be explained with the same considerations made for the AUC PR; that is, the performance of the models degrades as rarer events are considered because there are fewer chances to learn from the data. The relatively high Recall value shown by the Deep and Wide&Deep for the label "100 – 1000" in Fig. 6b may be explained by considering that the evaluation database contains only 7 events labeled as "100 – 1000"; therefore, detecting a few of them would make a significant difference in terms of Recall. In fact, the Deep and Wide&Deep models could identify 2 of the 7 target events, which explains the Recall value of 0.28. The reason for this unexpected behavior may lie in the advanced abstraction capabilities of these models, which might be able to capture the correct feature combinations leading to these rare events. The characteristics of the datasets may also have played a role. Specifically, considering the label "100 – 1000", the ratio of events in the training dataset/events in the evaluation dataset is 2.86; instead, the ratio is 1.9 for the label "10 – 100". This means that the models have more chances to learn and fewer chances to be tested on the label "100 – 1000" than on the label "10 – 100". Further tests must be performed to verify this insight and assess whether a different label distribution in the training and evaluation databases will change the performance of the models.

The results shown in Fig. 5c and Fig. 6c suggest that the model accuracy increases as a larger number of people involved is considered. However, it is worth recalling that high accuracy does not imply good performance when the task involves the identification of rare events. For instance, if there are only a few examples of a specific label in the training dataset, the model could achieve a high Accuracy by predicting that no event in the dataset has that specific label. That is, ignoring extremely rare labels would produce better results in terms of accuracy. Therefore, one possible explanation for Accuracy behavior is that the model "confidence" in performing positive predictions decreases when it deals with rare events; as a result, the model may conclude that ignoring the label and not performing any positive prediction may be more efficient, as the accuracy would not be affected.

In order to investigate the above-mentioned hypotheses and provide more insights into how the models performed their predictions, examples of correct and incorrect classification have been studied more in detail. The analysis has focused on the results obtained by the Wide&Deep model on the category "NPK" and label "1 – 10" at 200,000 iteration steps. The results have been screened in order to identify groups of similar events (i.e., with similar features) that contain examples of True Positives (i.e., critical events correctly identified) and False Negatives (i.e., undetected critical events). In order to reduce the number of events to screen, only those involving crude oil have been analyzed. This substance has been selected because it is well represented in both the training and evaluation dataset. In fact, crude oil is the most frequent substance in the training dataset (639 events) and the third most frequent in the evaluation dataset (99 events). The analysis of the evaluation dataset reveals that two events that caused from 1 to 10 fatalities share most of their features. However, the model correctly classified only one of them, while the other generated a False Negative. These events have been examined more in detail to find a possible reason for this error. Table 4 displays the most relevant features of these accidents.

The events involved a continuous release (i.e., "Contrel" in IT1, Table 4) caused by a mechanical failure of a pipeline. The most notable difference is that the first event involved a fire while the second release did not ignite (i.e., "Nonignite" in IS1, Table 4). Concerning the second event, one may argue that a release of Crude Oil from a pipeline without ignition is unlikely to cause killed. In fact, six other events in the evaluation dataset involved the release of crude oil from pipelines without ignition, and none caused any fatalities. All of these events have been correctly labeled by the model (i.e., True Negatives). A search for similar events in the training database reveals that 112 events involved the continuous release of crude oil without ignition, and all but two did not cause any fatalities. The two events that resulted in fatalities were caused by sabotage, which may justify a high death toll. Also, the analysis of the results produced by the Wide model for the same category and label shows that the algorithm performed the same kind of predictions for these events. This evidence suggests that the misclassification of event 2 in Table 4 may be explained by at least two factors: (i) the event is extremely rare since there is no other record of a similar event in the dataset, and (ii) the event description in MHIDAS may not be accurate enough to clarify the circumstances surrounding the fatalities. This indicates that the combination of features that rarely or never occurred in the training dataset may seriously affect the model performance. The development of models with better generalization capabilities may partially overcome this limitation. In addition, a better-balanced and more comprehensive database may considerably improve the prediction capabilities of data-driven models. The model inability to classify the second event in Table 4 indicates the possibility to further improve the taxonomy used in MHIDAS. In fact, despite being rational and informative, it cannot fully explain those incidents where fatalities are not caused by physical effects, such as exposure to thermal radiation, toxic levels in ambient air, and overpressure.

In order to further investigate the role of class distribution among training and test datasets, additional analysis has been performed considering ammonia as a reference substance. In fact, ammonia is the most frequent substance in the evaluation database with 153 events. However, only 137 events involving ammonia are found in the training dataset, and only 14 caused 1 to 10 fatalities. Instead, 29 events in the evaluation dataset caused the same amount of deaths. The discussions made so far may suggest that the imbalance between train and test datasets could have significantly degraded the performance of the algorithm. In fact, the results confirm this insight; only 4 of the 29 critical events have been correctly classified by the Wide and Wide&Deep models. This result proves that label and feature balance among training and evaluation datasets is crucial for ensuring good prediction performance.

The number of missing features may also play a significant role in determining the performance of the models. Intuitively, events with more missing features may be more difficult to classify due to the uncertainty surrounding the accident characteristics. As a result, the models may lack essential information to learn from or predict the outcomes of these incomplete observations. To confirm this insight, the frequency distribution of missing values among the correct and incorrect predictions made by the Wide&Deep model on the same category and label discussed above has been assessed and represented in Fig. 8.

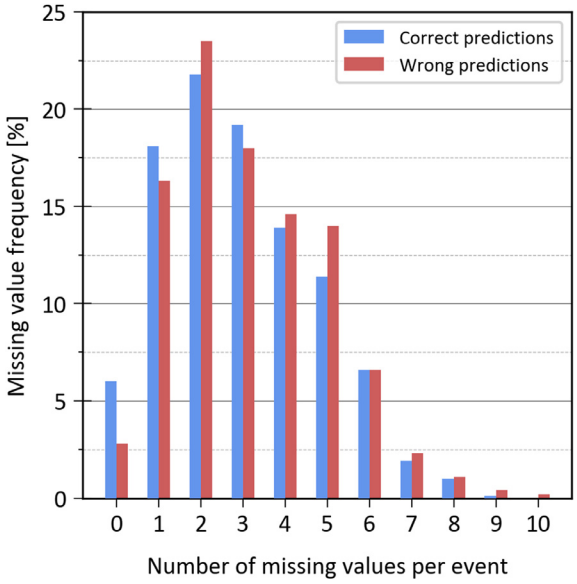


Fig. 8. Missing feature distribution among correct and wrong predictions made by the Wide&Deep model (category = "NPK", label = 1 – 10).

The x-axis in Fig. 8 represents the number of missing features, and the height of the bars shows the percentage of correctly (blue) or wrongly predicted events (red). The chart suggests that a correlation exists between missing values and classification performance. Specifically, events with a low number of missing features (i.e., 0, 1, and 3) are more likely to be correctly predicted. In contrast, events with a large number of missing features (i.e., ≥ 4) are more frequently misclassified. However, it is worth mentioning that the events with 2 missing features are more likely misclassified despite the low number of missing values. This abnormal behavior may be due to random effects in data distribution since most of the events in the training dataset have 2 missing features. Notwithstanding this anomaly, data appear to confirm that a high number of missing features has a negative impact on the model prediction capabilities.

As previously mentioned, one of the objectives of this study is to compare the performance of different models. The Wide model assumes a linear association between inputs and labels, while the Deep and Wide&Deep models can capture nonlinear relationships between features. The Bow-Tie representation shown in Fig. 7 suggests that the number of interactions between attributes increases as we consider an attribute that is far from the event to predict – the final outcome in this case. For this reason, the Deep and Wide&Deep may potentially provide better performance due to their ability to capture the effects of combinations of features. However, the results in Fig. 5, Fig. 6, and supplementary material indicate that there is not a single model that outperforms the others. In fact, the Deep model produces the best AUC PR and Recall for the label "NO" of the category "NPI" (Fig. 5a); however, the other models show larger Accuracy and Precision values for the same label of the category "NPK" (Fig. 6a). In addition, it may happen that a model produces the highest metric for the category NPI and the lowest metric for the category NPK; as an example, the deep model produces the largest Recall for the label "1 – 10" of the category NPI (Fig. 5b) and the smallest value for the same label of the category NPK (Fig. 6b). To further complicate the comparison, the number of iteration steps must be taken into account. Therefore, a scoring system was developed to rank and compare the models. The aim is to assign a score to each model according to its performance; two scores are obtained for each model: one for the category NPI and one for the category NPK. In order to simplify the method, the scoring system takes into account only the AUC PR, which is the most significant metric in this context. The process involves 8 steps:

Table 5
Scoring system multipliers.

Label	Multiplier
NO	1
1–10	2
10–100	3
100–1000	4
> 1000	5

- A category is selected (e.g., NPI).
- A number of iteration steps is selected (e.g., 200).
- The AUC PR values of the simulation performed for the pair category-number of iteration steps are selected and used in the following steps.
- For each label, the models are ranked based on the AUC PR values. Baseline scores are assigned to each model.
- 3 if the model produced the largest AUC PR,
- 2 if the model ranked second,
- 1 if the model produced the smallest AUC PR.
- Multipliers are assigned to each baseline score based on the severity category of the label (Table 5) - a model is "rewarded" when it outperforms the others on the identification of severe accidents.
- For each model, the scores obtained in step 5 are summed to obtain a partial score that indicates which model performs better on the pair category – number of iteration steps.
- Steps from 2 to 6 are repeated for each number of iteration steps. Partial scores of each model are summed to obtain a category score that indicates which model performs better on the category chosen in step 1.
- Steps from 1 to 7 are repeated for the other category.

The application of the procedure leads to the scores displayed in Table 6. The scoring system suggests that the best model in the category NPI and NPK is the Wide&Deep, followed by the Wide and Deep models. Obviously, the same ranking is obtained considering the overall score, which is the sum of the scores obtained in the categories NPI and NPK.

It is not surprising that the Wide&Deep model performed better than the others. In fact, the hybrid model combines the advantages of both the Linear and Deep models, as described in section 2.2.2.3. Nevertheless, a relatively unexpected result is that the Linear model performs better than the more sophisticated Deep model. This may suggest that the problem considered in this study requires stronger memorization capabilities rather than generalization. As already discussed, Deep models are prone to overfitting and overgeneralization. In addition, they need high-quality input data to perform as intended. The quality of MHIDAS database is sufficient, but certainly not excellent considering its public domain nature. Also, such advanced models may need more optimization and hyperparameters fine-tuning to perform adequately. On the contrary, the linear part of the Wide&Deep model may add stability and robustness to the algorithms, partially overcoming the issues related to the deep part. Apparently, the results indi-

Table 6
Scores assigned to the models.

Model	Score NPI	Score NPK	Overall score
Wide	134	108	242
Deep	80	99	179
Wide&Deep	146	153	299

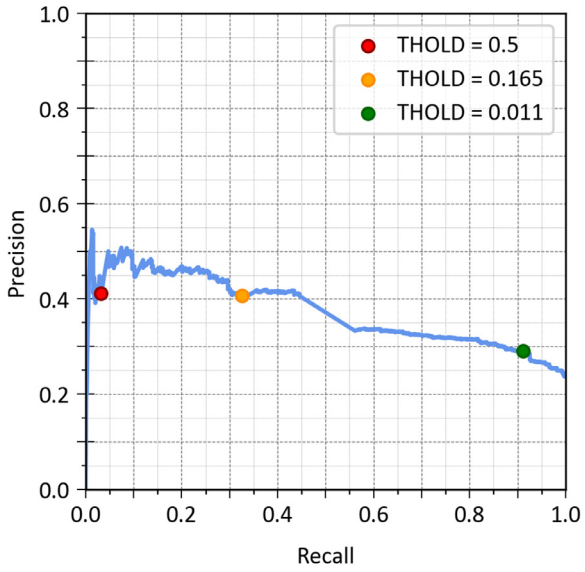


Fig. 9. Precision-Recall curve of the Deep model for the label 1 – 10 (NPK) at 2000 integration steps. THOLD represents the decision threshold.

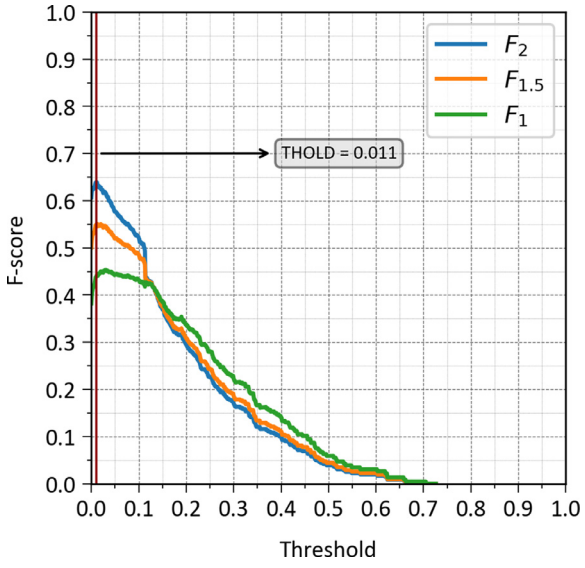


Fig. 10. F_1, $F_{1.5}$, and F_2 curves obtained by the Deep model for the label 1 – 10 (NPK) at 2000 integration steps. $F_{1.5}$, and F_2 show a global maximum for Threshold = 0.011. F_1 has a maximum at Threshold = 0.031.

cate that the approach benefits from a model capable of assessing the weights of each feature (or groups of features) independently rather than generalizing over all the features. A further study with more focus on the optimization of the model internal parameters (e.g., different number of hidden layers and units, activation function, learning decay) is suggested to test whether a different configuration of the Deep and Wide&Deep models would improve their performance.

In addition to these general considerations, it is worth discussing the role of the decision threshold in more detail. The Recall, Precision, and Accuracy values shown in Fig. 5 and Fig. 6 are obtained using a threshold equal to 0.5. One must bear in mind that low Recall and Precision values do not necessarily indicate poor performance; if the AUC PR is large enough, fine-tuning the decision threshold may improve the performance significantly. For example, consider the performance of the Deep model for the label "1 – 10" of the category NPK at 2000 integration steps (Fig. 6). The model produces a Recall close to 0 (Fig. 6b). But, the AUC PR value is in line with the other models (Fig. 6a). This suggests that a threshold of 0.5 may not be the best choice. In order to visualize the effect of this parameter on the performance metrics, the Precision-Recall curve is shown in Fig. 9.

Each point of the blue curve in Fig. 9 represents the Precision and Recall values at a specific threshold (THOLD). The red mark indicates Precision and Recall obtained using a threshold equal to 0.5 (i.e., the values shown in Fig. 6 for the Deep model and label "1 – 10"). The orange mark highlights that if the threshold is lowered to 0.165, the Deep model produces a Recall equal to 0.33 and a Precision of 0.41, which are in line with those obtained by the Wide and Wide&Deep models for the same label and category. This confirms that the Recall and Precision obtained using 0.5 as a threshold may not be representative of the model performance. In addition, it might be argued that misclassifying a "Deadly" accident as "Not Deadly" is more critical than misclassifying a "Not Deadly" event as "Deadly"; that is, False Negatives must be avoided, while False Positives may be tolerated. In this context, a good model must produce a high Recall, while a low precision might be considered acceptable and, to a certain extent, conservative. Therefore, the decision threshold may be further tuned in order to maximize a Recall oriented F-score (e.g., $F_{1.5}$ or F_2), as explained in section 2.2.3. The effect of the decision threshold on the F-measure is presented in Fig. 10, which describes three F-scores: F_1, $F_{1.5}$, and F_2.

From the data in Fig. 10, it is apparent that the recall-oriented $F_{1.5}$ and F_2 scores show a maximum for a decision threshold equal to 0.011. Instead, the F_1 score reaches its maximum at a threshold of 0.031. The green mark in Fig. 9 indicates that decreasing the decision threshold to 0.011 allows the Deep model to achieve a Recall equal to 0.91 and a Precision of 0.29, which means that the model can identify 9 out of 10 events that caused 1 – 10 killed with a precision of 29%. The performance is significantly improved considering that the same model can identify only 3 out of 100 events using 0.5 as a decision threshold (red mark in Fig. 9). As a drawback, the Precision has dropped from 0.41 to 0.29. However, Precision is not as crucial as Recall. In this study, a key requirement is that the model produces the fewest possible False Negatives (i.e., the Recall must be small) in order to prevent overlooking severe accidents. A small number of False Positives (i.e., a large Precision), although desirable, is not critical. Therefore, the significant improvement in Recall obtained through threshold tuning appear to compensate for the relatively small decrease in Precision.

In general, the results shown in Fig. 5 and Fig. 6 and the improvement obtained by an accurate threshold tuning suggest that the approach described in this study may be used to predict and discriminate the outcomes of accidents involving dangerous substances in terms of people injured and killed. The high level of detail, the ease of use, and the classification speed are some of the most significant benefits of this method. Furthermore, no earlier study prosed a Machine Learning approach for severity prediction that reached such a high level of detail. In addition to discriminating between injuries and fatalities, the algorithms proposed in this investigation provide additional information about the number of people involved. The detail level offered by these algorithms may permit the definition of more accurate preventive and mitigative actions and provides more practical and concrete support to safe design and operations.

8. Conclusions

The main goal of the current study was to demonstrate the use of Machine Learning techniques to (i) analyze and extract relevant knowledge from existing chemical accident databases and (ii)

use the acquired knowledge to predict the outcomes of new accidental events. A generic approach has been proposed, which relies on classification algorithms to predict the outcomes of chemical accidents in terms of people killed and injured. The method has been tested on a specific database, namely MHIDAS. To this end, three classification models have been used and compared, i.e., Wide, Deep, and Wide&Deep; the results indicate that the latter ensures the best performance.

The following conclusions can be drawn from the present study. Firstly, the results suggest that advanced analysis methods may be used to exploit existing accident data and perform predictions on the severity of new accidents. Secondly, the performance of the model largely depends on the quality of input data and the nature of the model itself. That is, if accident data are incomplete or uncertain, the choice of a model with advanced abstraction and generalization capabilities over a memorization-oriented model may not be advisable due to the risk of overgeneralization and overfitting. Thirdly, the performance of the model also depends on data availability. That is, the performance of the models degrades if extremely rare events are considered. Finally, the fine-tuning of the decision threshold to maximize a Recall-oriented F-measure may be an effective means of improving the performance of the algorithms, partially overcoming the issues of data scarcity and allowing the identification of more critical accidents.

However, although the results of the study appear promising, it is worth acknowledging some limitations. For instance, the approach has been tested on a specific database; further works should investigate whether the method might be applicable to different accident databases or industrial sectors. Also, it would be advisable to assess whether the knowledge extracted from a specific database might be used directly on different databases. A companion paper is proposed by Tamascelli et al. (2021) to investigate this topic. Another potential limitation is the choice of the attributes and taxonomy used to describe the accidents; the motivations behind this choice have been discussed in detail, but there is no guarantee that a different set of attributes would not improve the performance. In addition, the study reveals that the absence of an unambiguous and standardized system for recording accident data is a substantial obstacle to the spread of data-driven predictive methods. Therefore, the authors strongly encourage cooperation between institutions and academics to address this issue and exploit the potential of advanced analysis methods.

Notwithstanding the limitations, this is the first study that uses multiple discrete outcome variables and different ML models to predict the severity category of accidents involving dangerous substances. Therefore, this investigation makes a major contribution to research on Machine Learning methods for safety management and assessment in the chemical industry. In general, the approach may support the development of advanced predictive tools and represent an essential step toward Safety 4.0. More specifically, the techniques herein discussed may support hazard identification and consequence evaluation by providing a quick, practical, and easily understandable indication of the potential consequences of a release. Also, the approach may be used to identify the most important factors contributing to the accident severity. Finally, the method allows a reactive response to accidents by providing essential information to the emergency response team.

Declaration of Competing Interest

None.

References

AEA Technology, 1999. MHIDAS (Major Hazard Incident Data Service).

Ahadh, A., Binish, G.V., Srinivasan, R., 2021. Text mining of accident reports using semi-supervised keyword extraction and topic modeling. Process Saf. Environ. Prot. 155, 455–465. doi:10.1016/j.psep.2021.09.022.

Jazeera, Al, 2021. Thousands evacuated after Thai factory blast kills one rescue worker, wounds dozens - ABC News. aljazeera.

Alcides, J., Junior, G., Busso, C.M., Gobbo, S.C.O., Carreão, H., 2018. Making the links among environmental protection, process safety, and industry 4 . 0. Process Saf. Environ. Prot. 117, 372–382. doi:10.1016/j.psep.2018.05.017.

ARAMIS project team, 2004. Deliverable D.1.C.

Assi, K., Rahman, S.M., Mansoor, U., Ratrout, N., 2020. Predicting crash injury severity with machine learning algorithm synergized with clustering technique: a promising protocol. Int. J. Environ. Res. Public Health 17, 1–17. doi:10.3390/ijerph17155497.

Brink, H., Richards, J., Fetherolf, M., 2016. Real-World Machine Learning, First. Manning Publications, Shelter Island.

Bruha, I., 2017. Missing Attribute Values. In: Sammut, C., Webb, G.I. (Eds.), Encyclopedia of Machine Learning and Data Mining. Springer US, Boston, MA, pp. 834–841. doi:10.1007/978-1-4899-7687-1_954.

Bundesministerium für Umwelt Naturschutz Bau und Reaktorsicherheit, 2022. Central Reporting and Evaluation Office For Major Accidents and Incidents in Process Engineering Facilities - ZEMA [WWW Document]. URL https://www.infosis.uba.de/index.php/en/zema/index.html (accessed 8.28.20).

Bureau for Analysis of Industrial Risks and Pollutions, 2022. The ARIA Database - La référence du retour d'expérience sur accidents technologiques [WWW Document]. URL https://www.aria.developpement-durable.gouv.fr/the-barpi/the-aria-database/?lang=en (accessed 8.27.20).

Burnett, R.A., Si, D., 2017. Prediction of injuries and fatalities in aviation accidents through machine learning. In: ACM Int. Conf. Proceeding Ser. Part F1302, pp. 60–68. doi:10.1145/3093241.3093288.

Carvalho, T.P., Soares, F.A.A.M.N., Vita, R., Francisco, R., da, P., Basto, J.P., Alcalá, S.G.S., 2019. A systematic literature review of machine learning methods applied to predictive maintenance. Comput. Ind. Eng. 137, 106024. doi:10.1016/j.cie.2019.106024.

Chebila, M., 2021. Predicting the consequences of accidents involving dangerous substances using machine learning. Ecotoxicol. Environ. Saf. 208, 111470. doi:10.1016/j.ecoenv.2020.111470.

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., 2016. Wide & deep learning for recommender systems. In: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, pp. 7–10.

Chinchor, N., 1992. MUC-4 Evaluation Metrics. In: Proceedings of the 4th Conference on Message Understanding, MUC4 '92. Association for Computational Linguistics, USA, pp. 22–29. doi:10.3115/1072064.1072067.

Choi, J., Gu, B., Chin, S., Lee, J.S., 2020. Machine learning predictive model based on national data for fatal accidents of construction workers. Autom. Constr. 110, 102974. doi:10.1016/j.autcon.2019.102974.

Chung, P.W.H., Jefferson, M., 1998. The integration of accident databases with computer tools in the chemical industry. Comput. Chem. Eng. 22. doi:10.1016/s0098-1354(98)00135-5.

Cullen, W.D., 1990. The Public Inquiry Into the Piper Alpha Disaster. HMSO. London.

Drummond, C., 2017. Classification. In: Sammut, C., Webb, G.I. (Eds.), Encyclopedia of Machine Learning and Data Mining. Springer US, Boston, MA, pp. 205–208. doi:10.1007/978-1-4899-7687-1_111.

European Commission, 2022. eMARS Dashboard [WWW Document]. URL https://emars.jrc.ec.europa.eu/en/emars/content (accessed 8.27.20).

European Union, 2012. L 197. Off. J. Eur. Union 55, 38–71. doi:10.3000/19770677.L_2012.197.eng.

Gerassis, S., Saavedra, Á., Taboada, J., Alonso, E., Bastante, F.G., 2020. Differentiating between fatal and non-fatal mining accidents using artificial intelligence techniques. Int. J. Mining, Reclam. Environ. 34, 687–699. doi:10.1080/17480930.2019.1700008.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning, Adaptive Computation and Machine Learning Series. MIT Press, Cambridge, Massachusetts, United States.

Google, 2020a. Classification: thresholding [WWW Document]. URL https://developers.google.com/machine-learning/crash-course/classification/thresholding (accessed 6.15.20).

Google, 2020b. Feature Crosses: encoding Nonlinearity [WWW Document]. URL https://developers.google.com/machine-learning/crash-course/feature-crosses/encoding-nonlinearity (accessed 1.24.20).

Google, 2020c. Classification: accuracy | Machine Learning Crash Course [WWW Document]. URL https://developers.google.com/machine-learning/crash-course/classification/accuracy (accessed 1.24.20).

Google, 2020d Classification: precision and Recall | Machine Learning Crash Course [WWW Document]. URL https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall (accessed 1.24.20).

Han, J., Kamber, M., Pei, J., 2012. 8 - Classification: basic Concepts. In: Han, J, Kamber, M, Pei, J.B.T.D.M (Eds.), The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Boston, pp. 327–391. doi:10.1016/B978-0-12-381479-1.00008-3.

Hanida, A., Azmi, M., 2017. A Journey of Process Safety Management Program for Process Industry. Int. J. Eng. Technol. Sci. 8, 1–9. doi:10.15282/ijets.8.2017.1.10.1085.

Harding, A.B., 1997. MHIDAS: the first ten years. Inst. Chem. Eng. Symp. Ser. 39–50.

Hastie, T., Friedman, R., Tibshirani, J., 2009. The Elements of Statistical Learning. Springer-Verlag, New York doi:10.1007/978-0-387-84858-7.

IBM Cloud Education, 2020. What is Unsupervised Learning? | IBM [WWW Document]. URL https://www.ibm.com/cloud/learn/unsupervised-learning (accessed 5.27.21).
James, G., Hastie, T., Tibshirani, R., Witten, D., 2013. An Introduction to Statistical Learning: With Applications in R. Springer-Verlag, New York doi:10.1007/978-1-4614-7138-7.
Jefferson, M., Chung, P.W.H., Kletz, T.A., 1997. Learning the lessons from past accidents. Inst. Chem. Eng. Symp. Ser. 217–226.
Jelodar, H., Wang, Y., Yuan, C., Feng, X., Jiang, X., Li, Y., Zhao, L., 2019. Latent Dirichlet allocation (LDA) and topic modeling: models, applications, a survey. Multimed. Tools Appl. 78, 15169–15211. doi:10.1007/s11042-018-6894-4.
Jing, S., Liu, X., Gong, X., Tang, Y., Xiong, G., Liu, S., Xiang, S., Bi, R., 2022. Correlation analysis and text classification of chemical accident cases based on word embedding. Process Saf. Environ. Prot. 158, 698–710. doi:10.1016/j.psep.2021.12.038.
Jukes, E., 2018. Encyclopedia of Machine Learning and Data Mining, 2nd edition Reference Reviews doi:10.1108/rr-05-2018-0084.
Kalelkar, A.S., 1988. Investigation of large-magnitude incidents : bhopal as a case study. IChemE. Prev. Major Chem. Relat. Process Accid. 553–575.
Kletz, T., 2012. The history of process safety. J. Loss Prev. Process Ind. 25, 763–765. doi:10.1016/j.jlp.2012.03.011.
Kletz, T., 1993. Lessons from Disaster: How Organizations Have No Memory and Accidents Recur. Institution of Chemical Engineers, Rugby (UK.
Kurian, D., Sattari, F., Lefsrud, L., Ma, Y., 2020. Using machine learning and keyword analysis to analyze incidents and reduce risk in oil sands operations. Saf. Sci. 130, 104873. doi:10.1016/j.ssci.2020.104873.
Landucci, G., Paltrinieri, N., 2016. A methodology for frequency tailorization dedicated to the Oil & Gas sector. Process Saf. Environ. Prot. 104, 123–141. doi:10.1016/j.psep.2016.08.012.
Langstrand, J.-P., Nguyen, H.T., McDonald, R., 2021. Applying Deep Learning to Solve Alarm Flooding in Digital Nuclear Power Plant Control Rooms. In: Ahram, T. (Ed.), Advances in Artificial Intelligence, Software and Systems Engineering. Springer International Publishing, Cham, pp. 521–527.
Lee, J., Cameron, I., Hassall, M., 2019. Improving process safety : what roles for Digitalization and Industry. Process Saf. Environ. Prot. 132, 325–339. doi:10.1016/j.psep.2019.10.021.
Luo, X., Cruz, A.M., Tzioutzios, D., 2020. Extracting Natech Reports from Large Databases: development of a Semi-Intelligent Natech Identification Framework. Int. J. Disaster Risk Sci. 11, 735–750. doi:10.1007/s13753-020-00314-6.
Makaba, T., Dogo, E., 2019. A Comparison of Strategies for Missing Values in Data on Machine Learning Classification Algorithms. In: Proc. - 2019 Int. Multidiscip. Inf. Technol. Eng. Conf. IMITEC 2019 doi:10.1109/IMITEC45504.2019.9015889.
Mannan, M.S., Waldram, S.P., 2014. Learning lessons from incidents: a paradigm shift is overdue. Process Saf. Environ. Prot. 92, 760–765. doi:10.1016/j.psep.2014.02.001.
Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J., 2013. Distributed Representations of Words and Phrases and their Compositionality. In: Burges, C.J.C., Bottou, L., Welling, M., Ghahramani, Z., Weinberger, K.Q. (Eds.), Advances in Neural Information Processing Systems. Curran Associates, Inc.
Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective, Adaptive Computation and Machine Learning. MIT Press, Cambridge, Massachusetts, United States.
Paltrinieri, N., Comfort, L., Reniers, G., 2019. Learning about risk: machine learning for risk assessment. Saf. Sci. 118, 475–486. doi:10.1016/j.ssci.2019.06.001.
Paltrinieri, N., Dechy, N., Salzano, E., Wardman, M., Cozzani, V., 2013. Towards a new approach for the identification of atypical accident scenarios. J. Risk Res. 16, 337–354. doi:10.1080/13669877.2012.729518.
Paltrinieri, N., Patriarca, R., Stefana, E., Brocal, F., Reniers, G., 2020. Meta-learning for safety management. Chem. Eng. Trans. 82. doi:10.3303/CET2082029.
Paolanti, M., Romeo, L., Felicetti, A., Mancini, A., Frontoni, E., Loncarski, J., 2018. Machine Learning approach for Predictive Maintenance in Industry 4.0. 2018 14th IEEE/ASME Int. Conf. Mechatron. Embed. Syst. Appl. MESA doi:10.1109/MESA.2018.8449150, 2018.
Pasman, H.J., 2009. Learning from the past and knowledge management: are we making progress? J. Loss Prev. Process Ind. 22, 672–679. doi:10.1016/j.jlp.2008.07.010.
Pasman, H.J., Duxbury, H.A., Bjordal, E.N., 1992. Major hazards in the process industries: achievements and challenges in loss prevention. J. Hazard. Mater. 30, 1–38. doi:10.1016/0304-3894(92)87072-N.
Pasman, H.J., Fabiano, B., 2020. The Delft 1974 and 2019 European Loss Prevention Symposia: highlights and an impression of process safety evolutionary changes from the 1st to the 16th LPS. Process Saf. Environ. Prot. 147, 80–91. doi:10.1016/j.psep.2020.09.024.
Pasman, H.J., Fouchier, C., Park, S., Quddus, N., Laboureur, D., 2020. Beirut ammonium nitrate explosion: are not we really learning anything? Process Saf. Prog. 39. doi:10.1002/prs.12203.
Phark, C., Kim, W., Yoon, Y.S., Shin, G., Jung, S., 2018. Prediction of issuance of emergency evacuation orders for chemical accidents using machine learning algorithm. J. Loss Prev. Process Ind. 56, 162–169. doi:10.1016/j.jlp.2018.08.021.
Sarkar, S., Pramanik, A., Maiti, J., Reniers, G., 2020. Predicting and analyzing injury severity: a machine learning-based approach using class-imbalanced proactive and reactive data. Saf. Sci. 125, 104616. doi:10.1016/j.ssci.2020.104616.
Sasaki, Y., 2007. The truth of the F-measure. Teach Tutor mater 1–5.
Schottenfels, P., 2019. What is machine learning? A Google engineer explains [WWW Document]. URL https://www.blog.google/inside-google/googlers/ask-techspert-machine-learning/ (accessed 5.27.21).
Scikit-learn.org, 2020. Precision - Recall [WWW Document]. URL https://scikit-learn.org/stable/auto_examples/model_selection/plot_precision_recall.html
Souza, P., Freitas, M.F., Machado, C., 1996. Major Chemical Accidents in Industrializing Countries: the Socio-Political Amplification of Risk. Risk Anal 16, 19–29. doi:10.1111/j.1539-6924.1996.tb01433.x.
Stone, P., 2017. Reinforcement Learning BT - Encyclopedia of Machine Learning and Data Mining, in: Sammut, C., Webb, G.I. (Eds.), . Springer US, Boston, MA, pp. 1088–1090. doi:10.1007/978-1-4899-7687-1_720.
Tamascelli, N., Arslan, T., Shah, S.L., Paltrinieri, N., Cozzani, V., 2020a. A Machine Learning Approach to Predict Chattering Alarms. Chem. Eng. Trans. 82. doi:10.3303/CET2082032.
Tamascelli, N., Paltrinieri, N., Cozzani, V., 2020b. Predicting Chattering Alarms: a Machine Learning Approach. Comput. Chem. Eng. 107122. doi:10.1016/j.compchemeng.2020.107122.
Tamascelli, N., Scarponi, G., Paltrinieri, N., Cozzani, V., 2021. A data-driven approach to improve control room operators' response. Chem. Eng. Trans. 86, 757–762. doi:10.3303/CET2186127.
TensorFlow.org, 2021. Overfit and underfit | TensorFlow Core [WWW Document]. URL https://www.tensorflow.org/tutorials/keras/overfit_and_underfit (accessed 6.28.21).
TensorFlow.org, 2020a. Models and layers | TensorFlow.js [WWW Document]. URL https://www.tensorflow.org/js/guide/models_and_layers (accessed 1.24.20).
TensorFlow.org, 2020b. tf.nn.relu | TensorFlow Core v2.1.0 [WWW Document]. URL https://www.tensorflow.org/api_docs/python/tf/nn/relu (accessed 4.23.20).
TensorFlow.org, 2020c. tf.contrib.learn.Trainable | TensorFlow Core v1.15.0 [WWW Document]. URL https://www.tensorflow.org/versions/r1.15/api_docs/python/tf/contrib/learn/Trainable (accessed 12.17.20).
Unified Industrial & Harbour Fire Department, 2022. Failure and Accidents Technical information System (FACTS) [WWW Document]. URL http://www.factsonline.nl/.
United States Environmental Protection Agency, 2020. National Response System [WWW Document]. URL https://www.epa.gov/emergency-response/national-response-system (accessed 8.28.20).
Wahab, L., Jiang, H., 2019. A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. PLoS ONE 14, 1–17. doi:10.1371/journal.pone.0214966.
Wang, B., Zhao, J., 2022. Automatic frequency estimation of contributory factors for confined space accidents. Process Saf. Environ. Prot. 157, 193–207. doi:10.1016/j.psep.2021.11.004.
Xu, Z., Saleh, J.H., 2021. Machine learning for reliability engineering and safety applications: review of current status and future opportunities. Reliab. Eng. Syst. Saf. 211, 107530. doi:10.1016/j.ress.2021.107530.
Yedla, A., Kakhki, F.D., Jannesari, A., 2020. Predictive modeling for occupational safety outcomes and days away from work analysis in mining operations. Int. J. Environ. Res. Public Health 17, 1–17. doi:10.3390/ijerph17197054.
Zhang, J., Li, Z., Pu, Z., Xu, C., 2018. Comparing prediction performance for crash injury severity among various machine learning and statistical methods. IEEE Access 6, 60079–60087. doi:10.1109/ACCESS.2018.2874979.
Zope, K., Singh, K., Nistala, S.H., Basak, A., Rathore, P., Runkana, V., 2019. Anomaly detection and diagnosis in manufacturing systems: a comparative study of statistical, machine learning and deep learning techniques. Proc. Annu. Conf. Progn. Heal. Manag. Soc. PHM 11, 1–10. doi:10.36001/phmconf.2019.v11i1.815.

Article III.

Tamascelli, N., Paltrinieri, N., & Cozzani, V. (2023). **Learning From Major Accidents: A Meta-Learning Perspective**. Safety Science, 158, 105984. https://doi.org/10.1016/j.ssci.2022.105984.

Safety Science 158 (2023) 105984

Contents lists available at ScienceDirect

Safety Science

journal homepage: www.elsevier.com/locate/safety

Learning From Major Accidents: A Meta-Learning Perspective

Nicola Tamascelli *, Nicola Paltrinieri

Department of Mechanical and Industrial Engineering, NTNU, Trondheim, Norway
Department of Civil, Chemical, Environmental and Materials Engineering, University of Bologna, Bologna, Italy

Valerio Cozzani

Department of Civil, Chemical, Environmental and Materials Engineering, University of Bologna, Bologna, Italy

ARTICLE INFO

Keywords
Chemical Process Safety
Learning From Past Accidents
Machine Learning
MetaLearning
Transfer Learning

ABSTRACT

Learning from the past is essential to improve safety and reliability in the chemical industry. In the context of Industry 4.0 and Industry 5.0, where Artificial Intelligence and IoT are expanding throughout every industrial sector, it is essential to determine if an artificial learner may exploit historical accident data to support a more efficient and sustainable learning framework. One important limitation of Machine Learning algorithms is their difficulty in generalizing over multiple tasks. In this context, the present study aims to investigate the issue of meta-learning and transfer learning, evaluating whether the knowledge extracted from a generic accident database could be used to predict the consequence of new, technology-specific accidents. To this end, a classification algorithm is trained on a large and generic accident database to learn the relationship between accident features and consequence severity from a diverse pool of examples. Later, the acquired knowledge is transferred to another domain to predict the number of fatalities and injuries in new accidents. The methodology is evaluated on a test case, where two classification algorithms are trained on a generic accident database (i.e., the Major Hazard Incident Data Service) and evaluated on a technology-specific, lower-quality database. The results suggest that automated algorithms can learn from historical data and transfer knowledge to predict the severity of different types of accidents. The findings indicate that the knowledge gained from previous tasks might be used to address new tasks. Therefore, the proposed approach reduces the need for new data and the cost of the analyses.

1. Introduction

Learning from the past is essential for the advancement of every human activity, especially when mistakes may lead to disastrous consequences. In fact, lessons learned from past mistakes are vital to ensure safe operations in high-risk industries (Pasman, 2009). During the last decade, significant efforts have been made by regulators, academics, and industrials in order to avoid the re-occurrence of accidents involving dangerous substances. As an example, the Directive 2012/18/EU of the European Parliament and of the Council (European Union, 2012), also known as Seveso-III directive, stresses the importance of an effective learning strategy by introducing new requirements and providing guidelines for cross-organizational learning (Weibull et al., 2020). As an example, paragraph 4(c) of Annex II states that the safety report must include a "review of past accidents and incidents with the same substances and processes used, consideration of lessons learned from these, and explicit reference to specific measures taken to prevent such accidents" (European Union, 2012). Also, the directive requires Member States to investigate root causes of major accidents, and report their findings in the European Commission's eMARS database (European Commission, 2022).

Notwithstanding the undisputed importance of this topic, several authors highlighted that "the chemical industry as a whole does not learn from past accidents" (Chung and Jefferson, 1998). More than 10 years later, Pasman (2009) and Le Coze (2013) stated that little progress had been made; similar accidents reoccur, and organizations appear to struggle in deriving, retaining, and applying the lessons learned from the past. As an example, one might consider accidents related to ammonia production and utilization. In spite of its toxicity, ammonia is still an essential building block for the synthesis of nitrogen-based fertilizers,

* Corresponding author.
E-mail address: nicola.tamascelli@ntnu.no (N. Tamascelli).

https://doi.org/10.1016/j.ssci.2022.105984
Received 21 June 2022; Received in revised form 17 September 2022; Accepted 17 October 2022
Available online 2 November 2022

explosives, household cleaning solutions, and other chemicals (Pattabathula and Richardson, 2016). Globally, there is an ever-increasing demand for ammonia, which is mainly produced in large-scale plants, where significant quantities of dangerous substances (e.g., ammonia, methane, hydrogen, carbon monoxide) are handled and stored during daily activities. Ammonia was also proposed as a green fuel for maritime transportation (Chiong et al., 2021). For this reason, ammonia production may be considered a representative example of industrial activities that have a large potential to cause major accidents. Khan and Abbasi (1999) analyzed 1744 accidents that occurred between 1928 and 1997; the results indicate that ammonia was responsible for most events. For instance, the failure of a storage tank in Potchefstroom, South Africa, caused the release of 30 tons of anhydrous ammonia, which rapidly formed a gas cloud with a diameter of about 150 m. Eighteen people died during the accident, and 34 suffered serious injuries (Khan and Abbasi, 1999; Lonsdale, 1975). Since the late '90 s, the fundamentals of ammonia production have not changed much (Verma et al., 2019). Therefore, some may expect that the lesson learned from those accidents would have drastically lowered their occurrence. But unfortunately, this has not happened. Accidents still occur within the ammonia manufacturing industry: recent examples are those that took place in Phulpur, India (Pandya, 2020), and La Pobla de Mafumet, Spain (European Commission, 2019), causing two and one fatalities, respectively.

Learning from past accidents is still a new field (Le Coze, 2013), which lacks integration and standardization. An effective learning strategy relies on the interaction between organizations, institutions, and employees; several steps are needed to ensure the success of the process, and several obstacles must be faced. One may argue that human factors prevent an effective learning strategy (Pasman, 2009). In facts, humans have proven to have inherent generalization skills (Torrey and Shavlik, 2014) – i.e., the ability to transfer the knowledge gained in a specific task to a different domain – but there is a limit to the amount of data that can be processed and stored in our brain. Also, human learning may be biased and affected by emotions and interests (Weibull et al., 2020).

The idea of using data to update the risk picture has already been proposed in the past. For example, Landucci and Paltrinieri (2016) proposed a methodology to update the leak frequency based on technical, operational, human, and organizational factors. Recent advancements in IT, data science, and computational technology have led to the development of a new form of learning, named Machine Learning (ML), which relies on automated algorithms to extract knowledge from data. The growing interest in these algorithms has also affected the fields of safety and reliability. Several studies have proposed Machine Learning methods for predictive maintenance (Carvalho et al., 2019; Ge et al., 2017; Xu and Saleh, 2021), fault detection and diagnosis (He et al., 2005; Tian et al., 2015; Xu and Saleh, 2021; Zhong et al., 2014), diagnosis and prognosis of industrial alarm systems (Langstrand et al., 2021; Tamascelli et al., 2021, 2020), and Dynamic Risk Analysis (Paltrinieri et al., 2020, 2019).

On top of that, recent studies have focused on the application of ML methods to extract safety–critical knowledge from the abundance of accident data stored in the form of accident databases. Studies by Chebila (2021) and Tamascelli et al. (2022) suggest that classification algorithms might be used to acquire and retain knowledge about past accidents by analyzing existing databases. Specifically, these algorithms might be used to predict the consequences of an accident in terms of fatalities and injuries. In general, the approach suggests that artificial learners may partially overcome the limitations linked with the role of human factors in the learning framework. However, a major limitation of these studies is that they do not investigate whether the knowledge gained by these algorithms could be transferred to other domains. That is, the algorithms proposed in these studies have been trained and tested using data from a particular accident database, which is eMARS (Chebila, 2021) and MHIDAS in (Tamascelli et al., 2022). There is no guarantee that the knowledge extracted from these databases could be used to predict the outcomes of events from different data sources. In fact, humans have inherent transfer learning skills, but most Machine Learning algorithms cannot generalize over multiple tasks (Pan and Yang, 2010). A data-driven approach may overcome the issues related to the limited memorization and data processing skills of human beings, but whether these algorithms might be tailored to multiple tasks remains an open question.

In an attempt to address the challenges outlined above, a relatively new research line has focused on the so-called meta-learning (also known as learning to learn), which is a subfield of Machine Learning that focuses on "learning from prior experience in a systematic, data-driven way" (Vanschoren, 2018). The approach attempts to mimic the human ability to generalize and recall past experiences to increase the learning efficiency of new tasks (Griffiths et al., 2019). That is, meta-learning techniques aim to exploit the knowledge gained from previous tasks to improve and speed up the learning of new tasks (Lemke et al., 2015). These techniques may assist crucial and time-consuming stages of the Machine Learning Lifecycle (Ashmore et al., 2019), such as model selection (Stefana and Paltrinieri, 2021) and hyperparameters selection (Vanschoren, 2018). Several approaches have been developed to reach this goal; among them, there is the so-called Transfer Learning, which investigates methods to transfer knowledge from one task to another (Torrey and Shavlik, 2014). Depending on the problem under assessment, Transfer Learning may be divided into three categories: inductive, transductive, and unsupervised Transfer Learning (Pan and Yang, 2010).

This study set out to investigate the potential of inductive Transfer Learning to enhance and extend the scope of Machine Learning applications. To this end, a novel approach has been developed to leverage the knowledge extracted from nonspecific accident databases and predict the outcomes of technology-specific accidents. In particular, classification algorithms are trained on generic accident databases to learn the relationships between accident features and accident severity. Later, the pre-trained models are used to predict the outcomes of different, technology-specific accidents. Finally, performance metrics are produced to quantify and evaluate the success of the Transfer Learning procedure, and optimization strategies are proposed. The approach has been applied to a specific test case. A generic database named Major Hazard Incident Data Service (MHIDAS) (AEA Technology, 1999) has been used for the learning process. A specifically developed database reporting accidents involving ammonia releases has been used to test the generalization capabilities of the pre-trained model. The latter was developed in the present study by collecting data on accidents that involved ammonia or related substances.

The approach presented in this work will significantly accelerate the development of models for consequence prediction by reducing the need for new data and improving the generalization capabilities of Machine Learning algorithms. Furthermore, to the best of the authors' knowledge, there is no study in the field of process safety making use and investigating the potential role of Transfer Learning in the field of Chemical Process Safety. Therefore, this study makes a major contribution to research on the application of data-driven methods to extract safety-relevant knowledge hidden in accident data.

The overall structure of the study takes the form of six sections, including this introductory chapter. Section 2 provides the literature review. Section 3 describes the methodology, including data preprocessing, model training, transfer learning, performance evaluation, and optimization strategies. Section 4 describes the test case used to apply and evaluate the methodology. Results are presented and discussed in Section 5. Finally, conclusions are drawn in Section 6.

2. Related works

The use of Machine Learning to analyze accident data has gained traction in recent years (Sarkar and Maiti, 2020). Most of the studies on this topic pursue one of the following objectives: prediction of consequence severity, identification of influencing factors, or identification of

accident type. Also, based on the data source used for the analysis, we may distinguish between studies that analyze structured databases and those that focus on unstructured accident narratives. The research on this topic has focused on many industrial sectors, such as transportation (aviation, road, rail, and maritime), construction, mining, and petrochemical. However, a recent review on Machine Learning in occupational accident analysis (Sarkar and Maiti, 2020) concluded that most studies focus on road accidents (36.6 % of the analyzed articles), followed by construction sites (22 %), mining (6.9 %), aviation (5.2 %), manufacturing (5.2 %), and process industry (4.7 %).

In the context of chemical and process industries, some studies focused on the analysis of structured databases (e.g., MHIDAS). For example, Phark et al. (2018) demonstrated the use of classification algorithms to predict whether an emergency evacuation order would be issued after a release of toxic substances. In this study, the Hazardous Substances Emergency Events Surveillance dataset (HSEES) and the National Toxic Substance Incidents Program (NTSIP) were used to train and compare two classification algorithms: Naïve Bayes and Multi-Layer Perceptron (MLP). The results indicate that MLP achieves high accuracy on this specific task. Three years later, Chebila (2021) investigated the use of classification algorithms to predict the outcomes of major accidents involving dangerous substances in terms of consequences to humans, the environment, or material assets. To this end, the author analyzed the Major Accident Reporting System dataset (eMARS) (European Commission, 2022) with six different binary classification algorithms. The results indicate that the Random Forest (RF) offered the best performance in the prediction of damages to humans and the environment, while the Neural Network performed better in the "material damage" category. In spite of their remarkable performance, the models proposed by Chebila (2021) were not designed to discriminate between fatalities and injuries or to consider multiple severity levels. To overcome this limitation, Tamascelli et al. (2022) proposed a classification framework based on multiple discrete outcome variables to categorize accidents according to their severity (e.g., from 1 to 10 fatalities, from 11 to 100 fatalities, from 1 to 10 injuries). In this study, MHIDAS was used as a data source, and three classification algorithms were tested and compared: Linear, Deep Neural Network (DNN), and a hybrid Wide&Deep model. The study demonstrated the potential of ML algorithms to differentiate between different severity levels. However, the authors mentioned that data availability and poor data quality are significant obstacles to the diffusion of ML for consequence prediction. Similarly, Gangadhari et al. (2022) took advantage of rough set theory and classification algorithms to predict the outcome of accidents in the Oil&Gas industry. The authors considered four severity categories, namely "Near Miss", "Minor", "Major", and "Catastrophic". Accident reports were drawn from different sources and manually converted into a set of structured fields. Five classification algorithms were tested and compared. Hyperparameter tuning was performed to increase the model performance. The results indicate that the best model is XGbost (Chen and Guestrin, 2016), which returned an F1 score larger than 0.9 in every category. In spite of the good results, the authors mentioned that manual pre-processing of accident reports is extremely time-consuming; therefore, there is a need for techniques that can (i) automatically extract meaningful and accurate information from accident reports, or (ii) reduce the need for labeled data. A different approach was proposed by Nakhal A et al. (2021), who coupled ML and Business Intelligence (BI) to analyze MHIDAS and build a dynamic visualization tool that may greatly simplify information retrieval and facilitate the visualization of connections between accident characteristics. All of the articles mentioned so far take advantage of structured databases, such as eMARS, MHIDAS, HSEES, and NTSIP. Still, researchers have also focused on the analysis of unstructured accident narratives. Most of this research focuses on extracting accident features (e.g., the accident type, or the contributory factors) from textual accident reports in order to decrease the need for manual intervention. For example, Luo et al. (2020) proposed a semi-automatic algorithm to extract Natech events from the National Response Center (NRC). The method relies on a keyword extraction phase followed by a recurrent neural network for the classification of accident reports into different Natech categories (e.g., "Flood", "Hurricane", "Earthquake"). Kurian et al. (2020) investigated keyword extraction and classification algorithms to categorize unstructured accident reports based on the incident type (e.g., "Leak/Spill", "Operation", "Communication"). Jing et al. (2022) developed a method to identify the accident type (e.g., "Fires", "Explosions", "Poisoning") from unstructured chemical accident reports. They used a Natural Language Processing technique named word2vec to extract word embeddings and a Bidirectional Long Short Term Memory network (Bi-LSTM) with an attention mechanism to identify the accident category. Finally, Wang and Zhao (2022) focused on the extraction of contributory factors in confined space accidents. Accident reports were collected from websites such as safehoo.com and ichemsafe.com. In this study, Bidirectional Encoder Representations from Transformers (BERT) is used to extract word embeddings, which are eventually fed to a Bi-LSTM for the classification of contributory factors (e.g., "Improper tool", "Ventilation", "Inerting"). It is worth mentioning that most of the studies that focus on unstructured accident reports require manual intervention for labeling or converting unstructured reports into structured data. Certainly, these techniques have great potential to reduce the need for manual intervention in later stages (i.e., when new accident reports are analyzed). However, it is still unclear whether these models might be used to analyze reports that are different (in the content or in the format) from those used to train the models.

Apart from the chemical industry, it is also worth mentioning some contributions from other sectors, such as the transportation, mining, and construction industry. Significant efforts have been directed toward the analysis of road crashes (Assi et al., 2020; Kushwaha and Abirami, 2022; Wahab and Jiang, 2019; Zhang et al., 2018), aviation incidents (Andrei et al., 2022; Burnett and Si, 2017; Tanguy et al., 2016; Xu et al., 2020), and maritime incidents (Cakir et al., 2021; Lu et al., 2022; Rawson and Brito, 2022). In addition, many studies have focused on consequence prediction and influencing factors identification in construction sites (Choi et al., 2020; Goh and Chua, 2013; Poh et al., 2018; Tixier et al., 2016; Zhu et al., 2021), and mining operations (Gerassis et al., 2020; Kahraman, 2021; Palma et al., 2021; Yedla et al., 2020).

The literature analysis highlights several challenges that need to be addressed to advance the research on Machine Learning methods to predict the consequences of major accidents. Firstly, the research has mainly focused on the transportation, construction, and mining industries; few studies have analyzed accidents in the chemical industry. Secondly, data labeling and manual processing of unstructured reports are extremely time-consuming. Therefore, there is an urgent need for techniques that can automatically label accident reports or decrease the need for labeled data. In this context, Transfer Learning is particularly appealing because it may reduce the need for labeled data. To date, however, most studies do not investigate the model capability to transfer knowledge between different domains (e.g., different types of accidents). Therefore, a question remains unanswered; to what extent a model trained on a specific accident dataset can generalize the lesson to predict the outcome of accidents drawn from different sources? This study provides an exciting opportunity to address these challenges and advance our knowledge of Machine Learning models for consequence prediction. Firstly, this investigation is one of the few contributions that focus on the chemical industry. Secondly, only one other study used MHIDAS to develop predictive models (Tamascelli et al., 2022). Furthermore, a novel database on ammonia accidents is described in this study. Thirdly, this is one of the few contributions that investigates the potential of Transfer Learning in the analysis of accident databases with ML tools. To the best of the authors' knowledge, only Goldberg (2022) applied Transfer Learning in his recent work on Machine Learning techniques to automatically label accident narratives. However, there are significant differences between the approach presented in this study and the investigation described by Goldberg (2022). For example, this

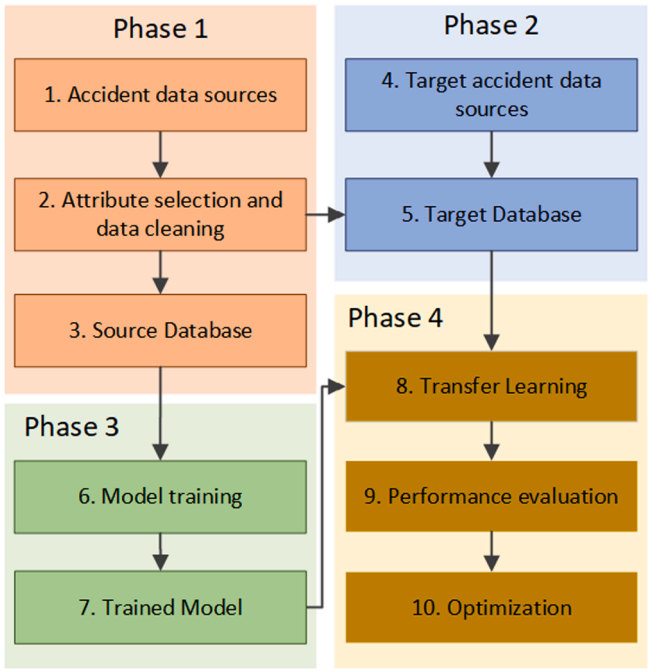


Fig. 1. Methodology workflow.

study analyzes structured accident data while unstructured accident narratives are used by Goldberg (2022). Also, occupational incidents are examined by Goldberg (2022), while this study focuses on major accidents involving dangerous substances.

3. Methodology

Fig. 1 reports the overall workflow of the methodology developed to extract information from an accident database (Source Database in Fig. 1) and use the acquired knowledge to predict the consequences of events included in a different database (Target Database in Fig. 1). The method involves four phases, each divided into several steps:

Phase 1. Source database creation (orange in Fig. 1);
Phase 2. Target database creation (blue in Fig. 1);
Phase 3. Model selection and training (green in Fig. 1);
Phase 4. Transfer Learning and optimization (ochre in Fig. 1).

In the first two phases, two databases are created: the first database (i.e., source) contains a large number of diverse accident data (i.e., non-technology-specific, non-substance-specific, and non-industry-specific), the second database (i.e., target) encloses accidents that occurred within a specific industry or involved a specific substance. In the third phase, a Machine Learning classification model is trained on the source database to learn the relationship between accident features and accident consequences in terms of fatalities and injuries. Finally, the pre-trained model predicts the outcomes of the events in the target database. Performance metrics are obtained, and optimization strategies are undertaken in order to fit the model to the new task.

3.1. Source database creation

Accident data from single or multiple data sources (step 1 in Fig. 1) are collected and used to populate the source database. Ideally, the source database should contain a large number of events that occurred in different industrial sectors (e.g., onshore and offshore), during different activities (e.g., processing, storage, transportation), and involving different substances. Data must be stored in tabular format, where rows represent accidental events and columns represent accident features (e.g., the date of the accident, the type of accident, and the substance involved).

Next, accident data must be pre-processed and cleaned (step 2 in Fig. 1). Accident features that are not considered important or informative must be removed. Also, accidents must be reported according to a uniform terminology. In addition, missing values must be removed or imputed because they are not recognized by Machine Learning algorithms. In this regard, one may refer to the extensive data-science literature, which offers many examples of missing values imputation techniques (Brink et al., 2016; Bruha, 2017; Makaba and Dogo, 2019).

Most data sources use integers to represent the number of fatalities and injuries. Since this study focuses on predicting the consequence category of the accident rather than the exact number of people involved, a set of categories are created in order to label the accidents according to their severity. As an example, one category might include incidents that caused no fatalities or injuries. Another category might contain accidents that caused from 1 to 10 fatalities or injuries, and so forth. The number and size of the categories can be adjusted to fit the user needs and the characteristics of the databases.

3.2. Target database creation

The target database should focus on specific accidents, such as those involving a particular substance. This is required in order to evaluate the capability of the model to generalize over different tasks. For the same reason, it would be preferable to use multiple data sources to populate the target database (step 4 in Fig. 1). However, if accidents are drawn from a single data source, it is critical to ensure that such data source was not used in the creation of the source database. Also, besides case-specific procedures, it is critical to ensure that source and target databases share the same structure. In other words, the databases must have the same number of attributes and same terminology; this requirement is represented by the connection between steps 2 and 5 in Fig. 1.

The procedure described above leads to the creation of two databases (i.e., source and target, steps 3 and 5 in Fig. 1) which are in the proper format for use in the Machine Learning simulations.

3.3. Model selection and training

According to Murphy (2012), Machine Learning is defined as "a set of methods that can automatically detect patterns in data, and then use the uncovered patterns to predict future data, or to perform other kind of decision making under uncertainty". That is, the term Machine Learning includes all the algorithms that can automatically extract knowledge from data and use that knowledge to make accurate predictions (Brink et al., 2016). The choice of the most appropriate algorithm depends on many factors, including the nature of the problem under assessment, data characteristics and availability, computational time requirements, and the expected output (Brink et al., 2016; Hastie et al., 2009; James et al., 2013; Khediri et al., 2012). In this study, the algorithm has to predict the severity of an accident given its main features, such as the amount of substance released and the equipment that originated the accident. Regression and classifications appear to be two feasible approaches to address this problem. Regression models could be used if the focus is on predicting the exact number of people involved in the accident. Instead, classification models should be used if the emphasis is on the prediction of a severity category (i.e., whether the accident has caused no fatalities/injuries, or whether the number of people involved is between 10 and 100). In this study, a classification approach has been adopted in order to reflect the implementation of severity categories in Risk Analysis techniques –e.g., the risk matrix proposed in (ARAMIS project team, 2004). Nevertheless, it would be advisable to investigate the use of Regression algorithms in further works.

3.3.1. Model training

Classification algorithms aim at categorizing objects into two or

more pre-defined categories. Briefly, the purpose of a classification algorithm is to learn the relationship between the features (i.e., meaningful attributes) of the object that must be classified, and its label (i.e., its category).

The development of the algorithm involves a training phase, where the algorithm "learns" the relationship between features and labels, and an evaluation phase, where the algorithm is tested against the ability to predict the labels of previously unseen objects. Often, the learning element of a classification algorithm is a function with tunable parameters (f). During the training phase, these internal weights are adjusted to find the optimal mapping between features (X) and labels (Y), as shown in the following equation (James et al., 2013).

$$Y \approx f(X) \tag{1}$$

In this study, the source database is used to train the Machine Learning model (step 6 in Fig. 1); that is, the entire database is fed to the model. During this phase, the user might decide to reiterate the training in order to simulate a more extensive database. In other words, the source database could be fed to the model multiple times. The number of reiterations over the source database is called the "number of iteration steps". A large number of iteration steps may improve the performance because the model has more chances to learn. In contrast, there is a risk of overfitting the model (TensorFlow.org, 2021).

3.3.2. Model description

The function f in Eq. (1) is the so-called model of the Machine Learning algorithm. In this study, two distinct models have been used to demonstrate the approach: a Linear model and a Deep Neural Network. Nevertheless, the methodology may be promptly adapted for use with different models.

3.3.2.1. Linear model. Linear models describe the labels as a linear combination of features (James et al., 2013). That is, Eq. (1) can be written as (Hastie et al., 2009):

$$Y = \alpha_0 + \sum_{i=1}^{N} x_i \alpha_i = X^T \alpha \tag{2}$$

Where:

Y = label;
α_0 = intercept (or bias);
α_i = coefficient (or weight);
x_i = feature;
X= (N + 1)-vector of features = $[1, x_1, x_2, \cdots, x_i, \cdots, x_N]$;
α= (N + 1)-vector of bias and weights =$[\alpha_0, \alpha_1, \alpha_2, \cdots, \alpha_i, \cdots, \alpha_N]$.

Linear models are one of the most simple and yet used methods (James et al., 2013). They are fast, robust, and suitable for analyzing large datasets (Hastie et al., 2009). The model coefficients can be easily accessed and compared to assess the relative importance of each feature (Brink et al., 2016).

As a drawback, linear models cannot capture nonlinear relationships between features and cannot interpret combinations of features that never occurred during the training phase (Cheng et al., 2016).

3.3.2.2. Deep model. The Deep model relies on Deep Neural Networks (DNNs) – i.e., multi-layer artificial networks whose creation had been loosely inspired by neuroscience (Goodfellow et al., 2016). The model consists of densely interconnected units that mimic the functioning of neurons in nervous tissues. These units – also called hidden units – are organized in hidden layers (Brink et al., 2016). These networks are also called Feedforward Neural Networks because information flows from features to labels through hidden units in a single direction (Goodfellow et al., 2016). Fig. 2 displays the structure of a Deep Neural Network.

The input layer of a DNN is the vector of the features (orange in Fig. 2, X in equation (1)). The output layer contains the labels (green in Fig. 2, Y in equation (1)). Between the input and output layers, there are one or more hidden layers (H1, H2, and H3 in Fig. 2), each comprising several hidden units (Z_i^k in Fig. 2). The mapping from features to labels involves both linear combinations and nonlinear transformations.

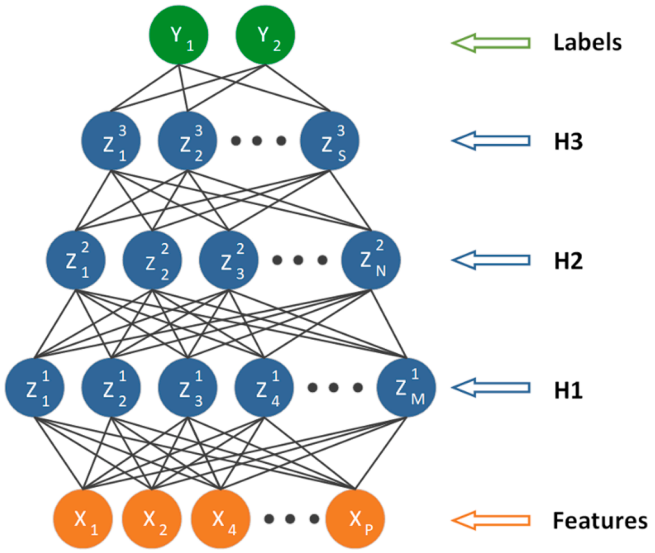


Fig. 2. Schematic representation of a DNN with three hidden layers (H1, H2, H3), P input features (X_i), and two output labels (Y_1 and Y_2).

The number of hidden units and hidden layers are design parameters. In general, deeper and wider networks perform better, but the computational effort required to train the model increases as more hidden units and layers are used (Hastie et al., 2009).

Deep Neural Networks have good generalization capabilities and can capture nonlinear relationships between features (Goodfellow et al., 2016). For these reasons, they are widely used in meta-learning approaches (Vanschoren, 2018). On the other hand, they are prone to overfitting and overgeneralization, and they are sensitive to poor-quality and missing input data (Goodfellow et al., 2016; Hastie et al., 2009).

3.4. Transfer Learning and optimization

3.4.1. Transfer learning

Torrey and Shavlik (2014) define Transfer Learning as "the improvement of learning in a new task through the transfer of knowledge from a related task that has already been learned". Pan and Yang (2010) classify Transfer Learning techniques into three categories: inductive, transductive, and unsupervised transfer learning. The categorization is based on label availability. Inductive transfer is used when the labels of source and target events are available. Instead, transductive learning is used if only source events are labeled. On the other hand, if source and target events are not labeled, unsupervised transfer is used. In this study, inductive transfer learning is used because both source and target datasets are labeled (i.e., the number of people involved in each event is known).

In inductive transfer learning, a model $\mathscr{L}$ is initially trained on one or more source tasks t (e.g., classification of accidental events of a broad and generic dataset). In this phase, the model configuration θ is tuned to perform well on t; as a result, an updated configuration θ^* is obtained. Finally, the pre-trained model $\mathscr{L}_{\theta^*}$ is optimized to fit a new task t_{new} (e.g., classification of substance-specific accidents). If the tasks t and t_{new} are relatively similar, the optimization of $\mathscr{L}_{\theta^*}$ will require less effort than starting from scratch (Torrey and Shavlik, 2014), especially in cases where t_{new} has a limited amount of data (Donahue et al., 2014).

In this study, the model trained on the source database (i.e., $\mathscr{L}_{\theta^*}$) is used to predict the labels of the events included in the target database

(step 8 in Fig. 1). The success of the operation depends on different aspects, including the quality of the source dataset and the similarity between the datasets (Vanschoren, 2018). Furthermore, the No-Free Lunch theorem states that "if an algorithm does particularly well on average for one class of problems then it must do worse on average over the remaining problems" (Wolpert and Macready, 1997), which means that there is no single algorithm that is universally best for different tasks (Yang, 2014). Thus, it is not guaranteed that the model that performs best on the source task will produce better results on the target task.

3.4.2. Performance evaluation

After the Transfer Learning procedure, the algorithm performance is assessed by comparing predicted and true labels (step 9 in Fig. 1). From now on, the letter "Y" will be used to identify a positive prediction (e.g., "Deadly") and the letter "N" will be used for a negative prediction (e.g., "Not Deadly"). Four metrics can be defined to take into account different outcomes:

- TP = True Positive –i.e., predicted label = Y, true label = Y;
- TN = True Negative –i.e., predicted label = N, true label = N;
- FP = False Positive –i.e., predicted label = Y, true label = N;
- FN = False Negative –i.e., predicted label = N, true label = Y.

In addition, these metrics are used to build three performance indicators:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{3}$$

$$Precision = \frac{TP}{TP + FP} \tag{4}$$

$$Recall = \frac{TP}{TP + FN} \tag{5}$$

Accuracy is the fraction of objects that have been correctly classified. Precision represents the "success rate" of a positive prediction. Recall indicates the fraction of real positives that have been correctly predicted.

In general, the performance of an algorithm cannot be evaluated by considering only one indicator (Brink et al., 2016). High accuracy does not ensure good performance because different problems have different requirements. For instance, if the problem involves the identification of classes that occur rarely, the indicator that must be optimized is the Recall (Brink et al., 2016).

3.4.2.1. Class probability and decision threshold. It is worth recalling that classification algorithms consider a certain grade of uncertainty when performing predictions. The model does not provide a single predicted label. Rather, the algorithm calculates the probabilities of each category (James et al., 2013). For example, if accident events are classified, *Y* in Eq. (1) is not a single label (i.e., "Deadly" or "Not Deadly") but a two-dimensional vector that contains the probability of each category (e.g., [P(Deadly) = 0.8, P(Not Deadly) = 0.2]). Therefore, a decision threshold is needed to convert probabilities into the predicted label. By default, a threshold value of 0.5 is used –i.e., if the probability of the class "Deadly" is greater than 0.5, the algorithm concludes that the accident resulted in fatalities. The decision threshold is a design parameter that may be tuned to optimize the algorithm based on the problem under assessment (Zhang et al., 2020).

3.4.3. Optimization

Further optimization of the pre-trained model is required to fit the target task (step 10 in Fig. 1). This need for optimization is common to most meta-learning approaches and arises from the intrinsic differences between tasks (Vanschoren, 2018). If the tasks are similar, fewer efforts will be required to offset the differences and learn the target task.

There are different methods to optimize and improve the performance of a Machine Learning algorithm, including hyperparameters tuning, thresholding, and optimizer tuning (Brink et al., 2016; Goodfellow et al., 2016; Hastie et al., 2009; James et al., 2013). In this study, attention has been directed toward thresholding because of its easy implementation. Other techniques, such as hyperparameters tuning or optimizer tuning, are beyond the scope of the work.

Thresholding (or threshold moving) consists in varying the decision threshold to optimize one of the metrics described in Section 3.4.2. Lowering the threshold causes the Recall to either increase or remain constant. Instead, Precision may fluctuate when the threshold is decreased. Usually, reducing the threshold causes the Precision to decrease because more False Positives may be generated. That is, Precision can be traded for Recall (Goodfellow et al., 2016) and vice-versa, but it is uncommon to improve both metrics by varying the threshold.

Precision-Recall (PR) curves are valuable means for evaluating how Precision and Recall change with the decision threshold. An example of a PR curve is shown in Fig. 3. The coordinates of points in the curve represent the values of Precision and Recall obtained using a specific decision threshold. The rightmost side of the curve (i.e., Recall = 1) is obtained at threshold = 0. In this case, every object in the evaluation database is labeled as "Y"; therefore, FN is equal to 0 in Eq. (5). The leftmost side of the curve (i.e., Recall = 0) is obtained at threshold = 1, which means that all the objects are labeled as "N"; therefore, TP is 0 in Eq. (5).

The Area Under the Curve Precision-Recall (AUC PR) is a comprehensive indicator of the model performance and, by extension, of the success of the transfer-learning procedure. The larger the area under the curve (i.e., closer to 1), the higher Precision and Recall values can be obtained. By default, the model returns Precision and Recall values obtained at threshold = 0.5. Nevertheless, the decision threshold may be changed to improve the Recall or/and the Precision, depending on the problem under assessment. For example, if the problem requires identifying rare or critical categories, the threshold might be lowered to increase the Recall.

One may decide to adjust the threshold to achieve a target value of Recall or Precision. Instead, Precision and Recall might be considered together in the so-called F-score (Chinchor, 1992):

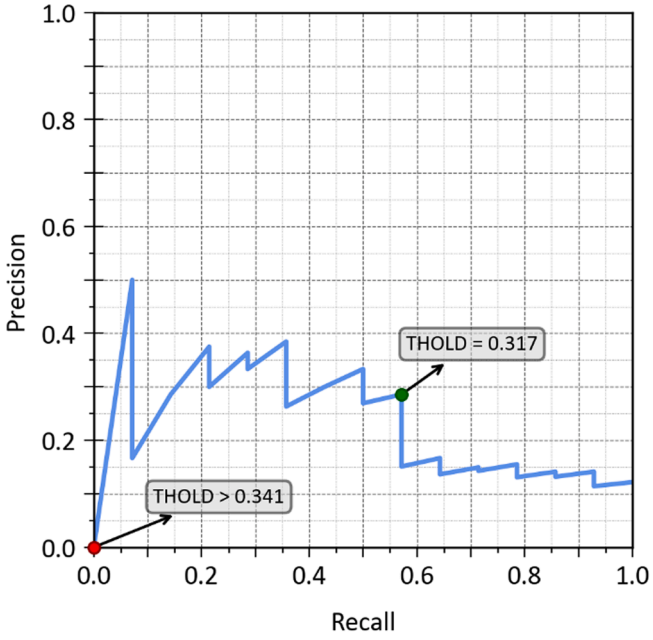


Fig. 3. Precision-Recall curve of the Deep model for the label 1 – 10 (NPK) at 2000 integration steps. THOLD represents the decision threshold.

$$F_\beta = (1+\beta^2) \bullet \frac{\text{Precision} \bullet \text{Recall}}{(\beta^2 \bullet \text{Precision}) + \text{Recall}} \quad (6)$$

Where:

β = non-negative real number.

The parameter β serves as a weight. If $\beta = 1$, the F-score represents the harmonic mean between Precision and Recall (Han et al., 2012). Therefore, F_1 is mostly used when Precision and Recall are equally important. If $\beta > 1$, the measure is Recall-oriented (Sasaki, 2007). For example, $\beta = 2$ means that Recall is twice as important as Precision (Chinchor, 1992). If $\beta < 1$, the measure is Precision-oriented.

The F-score assumes values between 0 and 1: the higher, the better the performance. F_β depends on the Precision and Recall values, which ultimately depend on the decision threshold. Thus, the best threshold might be identified as the one that maximizes the F-score.

4. Test case analysis

The following paragraphs describe the test case set out to demonstrate the approach. The first two paragraphs illustrate the datasets used as source and target databases. The last paragraph describes the Machine Learning simulations.

4.1. Source database: MHIDAS

Accident data extracted from MHIDAS has been used to build the source accident database, as described in Section 3.1. MHIDAS is an accident database founded in 1986 by the Safety and Reliability Directorate (SRD) and the Health and Safety Executive (HSE). It contains information about industrial incidents involving dangerous substances that "resulted in, or had the potential to produce, a significant impact on the public at large" (AEA Technology, 1999). Data are drawn from public domain sources (e.g., newspapers, journals, published reports) to grant the broadest dissemination. AEA Technology had been responsible for maintaining and updating the database from its foundation until the early 2000s, when the database was no more updated. The latest version of the database contains records of more than 8900 incidents from over 95 countries, covering a time span from the first years of the 20th century until the late nineties (AEA Technology, 1999).

Most of the events reported took place in the 1990s in the US and Europe, since it was easier to access incident information from these areas. The public domain nature of the database also affects its quality and completeness (Harding, 1997). For example, generic information –i.e., the date, the location, and the number of fatalities– are typically described in detail, while more specific ones –i.e., the incident type and the ignition source– may not be reported. In fact, the biggest limitations of MHIDAS are inaccuracy and missing information (Tauseef et al., 2011); for example, more than 40 % of the events in MHIDAS do not have any information on the causes of the accident (Tamascelli et al., 2022). However, the overall quality of the database is sufficient for the purposes of this study.

Incidents in MHIDAS are described by a list of attributes, each providing a piece of information about an incident (e.g., the location, the substances involved, the number of people involved). An attribute is described by one or more codes (i.e., standardized keywords). In total, 22 attributes are used in the database, which are not equally meaningful for the purpose of this study (e.g., the Accession Number, a unique identifier assigned to each record, and the number of hard copy references for the incident have not been considered since they do not convey any useful information from the safety perspective). In total, sixteen attributes have been selected for use in the source database (Table 1). A reduced version of the database was thus obtained, which contains 16 columns and 8972 rows. The first 14 columns represent accident features, and the last two columns (i.e., NPI and NPK in Table 1) represent the labels. Finally, the number of fatalities and injuries are converted into their respective consequence categories. To this end, the idea of "class of consequences" as used in risk matrices (ARAMIS project team, 2004) has inspired the creation of three consequence categories in order to label the accidents according to their severity, as shown in Table 2. For example, if an accident caused 5 fatalities and 70 injuries, NPK is "1 – 10", and NPI is "10 – 100". In addition, columns referring to multiple-features entries have been split so that each column includes one entry only. For instance, it has been found that the maximum number of entries for the feature "Incident Type" is three. Therefore, three columns have been used to represent this feature in the database (i.e., "IT1", "IT2", and "IT3"). Finally, missing values have been substituted by the string "NaN".

It is worth mentioning that the selection of attributes presented in Table 1 was manual and mainly guided by domain knowledge. In fact, each attribute represents a meaningful piece of information about an incident. Together, the keywords provide a synthetic but rather exhaustive description of the incident, from its causes to consequences on humans. For example, the attributes Date (DA) and Location (LO) may indicate something about the socio-economic status of the area affected by the incident. For example, Souza et al. (1996) highlighted that impoverished countries are more exposed to industrial risk. This insight is also confirmed by several accident reports, including the Bhopal disaster (Kalelkar, 1988) and the recent Beirut explosion (Pasman et al., 2020). Further, the ten attributes after "Location" in Table 1 focus on technical details, such as the origin, the source, the substance released, and the accident type. The effects on humans are described by the attributes Population Density (PO), Number of People Injured (NPI), and Number of People Killed (NPK), while the Number of People Evacuated (NPE) may indicate the effectiveness of the Emergency Response Plan.

Table 1
Selection of meaningful attributes used in this study. A brief description of each attribute is provided. * = Multiple entry fields (e.g., "Release" AND "Pool Fire" for IT, "Flammable" AND "Toxic" for MH).

Attribute		Description
DA	Date	Date of the incident.
LO	Location	Town, region, and country of the incident
GC	General Cause	The general cause - or causes - which triggered the event (e.g. Mechanical failure, Human Error)
SC	Specific Cause	The specific cause - or causes - which triggered the event (e.g. Brittle fracture, Overpressure, Fire)
GOG	General Origin	Area of the plant where the incident originated from (e.g. Process, Storage, Warehouse)
SOG	Specific Origin	Equipment that originated the incident (e.g. Pump, Vessel, Pipeline)
MN	Material Name*	Names of dangerous substances involved in the incident
MH	Material Hazard*	The hazard class of the substances involved (e.g. Toxic, Explosive, Corrosive, Oxidizing)
MC	Material Code*	Four-digit code of the substance involved
QY	Quantity	The amount of substances released (tons)
IS	Ignition Source	Type of ignition source (e.g. hot surface, flares, boilers)
IT	Incident Type*	Incident typology (e.g. Release, BLEVE, Physical Explosion)
NPE	Evacuated	Number of people evacuated
PD	Population Density	Population density in the Area (i.e. "Rural" for low - sparse population, "Urban" for highly populated Area)
NPI	Injured	Number of people injured in the incident
NPK	Fatalities	Number of people killed in the accident

Table 2
Accident consequence categories.

Category	Description
NO	no fatalities/injuries
1–10	from 1 to 10 fatalities/injuries
10–100	from 10 to 100 fatalities/injuries

Table 3
Number of events collected, per source, and source weight in the Ammonia Plant Accident Database. Source weight represents the contribution of each source to the database.

Source	Events	Weight [%]
NRC (United States Environmental Protection Agency, 2020)	39	27.9
Ammonia Plant Safety and Related Facilities (AIChE, 2001)	31	22.1
eMARS (European Commission, 2022)	21	15
Aria (Bureau for Analysis of Industrial Risks and Pollutions, 2022)	12	8.6
MHIDAS (AEA Technology, 1999)	11	7.9
JFKD (Japan Science and Technology Agency, 2005)	10	7.1
Lees' (Lees, 2004)	5	3.6
ZEMA (Bundesministerium für Umwelt Naturschutz Bau und Reaktorsicherheit, 2022)	5	3.6
OSHA (EU-OSHA, 1994)	4	2.8
Other	2	1.4

4.2. Target database: The Ammonia Plant Accident Database

To the best of the authors' knowledge, a specific database exclusively reporting accidents that affected ammonia production plants is not available. Thus, a new database was created – called the Ammonia Plant Accident Database – by collecting information about accidents and incidents from different sources. Only events that occurred in plants for ammonia production or in plants where similar technologies are used (e.g., Desulphurization, Reforming, Syngas Upgrading) were included in the database. The latter data were included to enrich the statistical significance of the database. Data on more than 140 relevant events were included in the database. The data were derived from nine main data sources, which are displayed in Table 3 together with the number of events found in each source. Specific checks were carried out to avoid the inclusion of duplicates and, in case, the entry was attributed to the database that provides the largest number of significant attributes.

The Ammonia Plant Accident Database (i.e., the target database) and the source database share the same structure, as suggested in Section 3.2. Specifically, each accident is described through a list of attributes (Table 1) and attribute codes, which have been entirely derived from the source database.

The frequency distribution of attribute codes in the target database (e.g., the fraction of incidents that lead to fire rather than explosion, or that involved syngas rather than ammonia) is a key piece of information to support the analysis, to interpret the results, and to highlight the limits of the database. As an example, the frequency distribution of the attributes General Origin, Incident Type, General Cause, Specific Cause, Material Name, and Number of People Affected are displayed in Fig. 4.

Most of the incidents in the target database involved the release of Ammonia or Syngas (Fig. 4.e and Fig. 4.b) within the Process area of the plant (Fig. 4.a). Mechanical failures and Human factors are the most frequent cause of accidents (Fig. 4.c). Fig. 4.f reveals that more than 80 % of the incidents had caused no injuries or fatalities. Considering the more severe accidents, the number of fatalities is always smaller than the number of injured, and the frequencies decrease as the number of people affected increases. No accident causing more than 100 injuries or fatalities is found in the database.

It is important to stress that most of the accidents in the target database are derived from a few different sources (Table 3). More than half of the events have been extracted from two sources only: the NRC database (United States Environmental Protection Agency, 2020) and the Ammonia Plant Safety and Related Facilities (AIChE, 2001). Thus, the overall features of the database are likely to be affected by the characteristics of these two sources. For instance, the NRC database does not always include the causes or the origin of the accident. Observing Fig. 4, it is clear how the characteristics of NRC affect the target database; the attributes that describe the cause and origin of the accident are not always registered (Fig. 4.c, and d).

Finally, it should be remarked that each source used to build the target database has its own way of describing an accident –e.g., different keywords, attributes, and codes. Thus, the detail level and the quality of information vary across different sources. In some instances, it has not been simple to find the most representative set of attribute codes because the original report uses different keywords or because the needed information is completely missing. Significant efforts are needed to gather and ensure consistency between data from different sources (Parmiggiani et al., 2022). It has been observed that accident reports were often not clear and incomplete, especially regarding detailed information. The database is affected by a significant incidence of missing values –i.e., "NaN" in Fig. 4. The attributes that describe the cause and the origin of the incident (e.g., GC, SC, GOG, SOG) show a high incidence of missing values. For instance, nearly 15 % of the incidents in the target database contain no information about the General Origin or the Incident Type (Fig. 4.a and Fig. 4.b). Additionally, General and Specific causes (Fig. 4.c and Fig. 4.d) show missing values frequency larger than 20 %. The incidence of missing values in the ammonia database is larger than in MHIDAS (Tamascelli et al., 2022), and the overall quality is thus lower.

4.3. Model training and Transfer Learning

The models described in Section 3.3.2 have been trained on the source database and evaluated on the target database, as described in Sections 3.3 and 3.4. The Deep Neural Network used in this study has three hidden layers, with 1024, 512, and 256 hidden units, respectively. The optimizers used in the Wide and Deep models are Ftrl and Adagrad (TensorFlow.org, 2020a, 2020b), respectively.

Two sets of binary classifications have been performed. The first set aims to identify the number of fatalities (i.e., NPK), the latter focuses on the number of people injured (i.e., NPI).

It is worth mentioning that each simulation has been performed using different iteration steps. Specifically, 200, 2000, 20000, and 200'000 steps have been used. Therefore, a set of 4 binary classifications have been performed for each combination of model (Wide or Deep), label category (NPI or NPK), and label ("NO", "1–10", "10–100"). Five steps have been followed to complete a simulation:

1. a model is selected;
2. a label category is selected;
3. a label is selected;
4. an iteration step is selected;
5. the model is trained on the source dataset;
6. the pre-trained model is evaluated on the target dataset.

The steps described above are reiterated in order to cover all possible combinations of model, label category, label, and iteration steps. Performance metrics and performance indicators are obtained for each simulation in order to evaluate the success of the Transfer Learning procedure. Finally, optimization strategies are assessed.

5. Results and discussion

The complete set of results of the transfer learning procedure is reported in the supplementary material. A selection of the most noteworthy simulations is displayed in Fig. 5 and Fig. 6, which focus on the category "NPI" and "NPK" respectively. In both figures, the performance indicators AUC PR, Recall, Accuracy, and Precision are displayed for each label and model.

The results displayed in Fig. 5 and Fig. 6 have been selected based on the AUC PR value. Specifically, the number of iteration steps that led to the largest AUC PR has been selected. The number of iteration steps used to obtain the metrics in Fig. 5 and Fig. 6 is shown in Table 4. The AUC PR has been chosen because it is independent of the decision threshold and representative of the potential model performance; in other words, it is

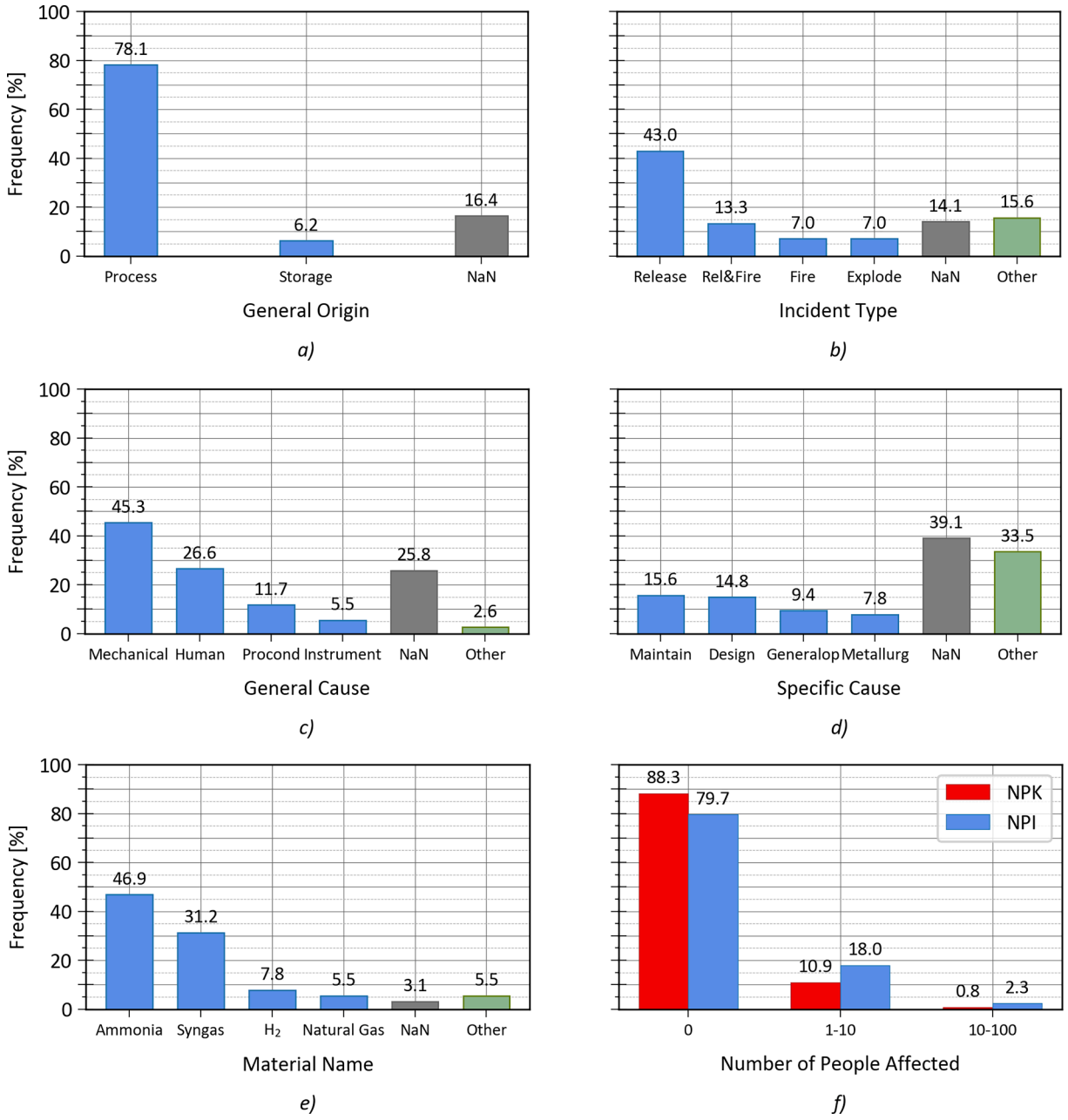


Fig. 4. Frequency distribution of the attributes GOG (a), IT (b), GC (c), SC (d), MN (e), NPK and NPI (f) (see Table 1 for the description of the attributes). Attribute codes are represented on the x-axis. "NaN" refers to missing values, "Other" refers to attribute codes that have not been represented in the figure for the sake of brevity.

one of the most comprehensive indicators of the success of the Transfer Learning procedure. Therefore, Fig. 5 and Fig. 6 provide a visual representation of the best performances achieved by the models in absolute terms, allowing a qualitative comparison between the algorithms.

From the data in Fig. 5 and Fig. 6, it is apparent that there is not a single model that outperforms the others in every simulation. For example, the Wide model shows an AUC PR higher than the Deep model in the category 1 – 10 NPI, while the opposite happens in category NO NPI (Fig. 5.a). Also, the Deep model outperforms the wide model in category 1 – 10 NPI (Fig. 5.a), while the opposite happens in category 1 – 10 NPK (Fig. 6.a). The same behavior is also evident in the complete set of results. Therefore, a scoring system has been used to rank the model performance and identify the best algorithm for this specific task. Briefly, the ranking system is designed to reward the model that produces the larger AUC PR in the most critical categories (i.e., those referring to events that caused a large number of fatalities or injuries). The procedure generates a score for each model and label category. By summing the scores of the two categories (i.e., the one calculated for "NPK" and the one for "NPI"), it is possible to obtain an overall measure of the model performance; larger scores indicate better performance. Table 5 reports the results of the scoring system. The Wide model offered the best performance in both categories and obtained the highest overall score. This finding may seem unexpected since DNNs are advanced models with inherent generalization and abstraction capabilities. In fact, the consequence of an accident results from the combination of many intermediate events. Thus, the Deep model was supposed to perform better on a Transfer Learning task due to the ability to capture inter-feature relationships and nonlinearities.

However, DNNs are prone to overfitting and overgeneralization, and they need high-quality input data to perform as intended. The quality of

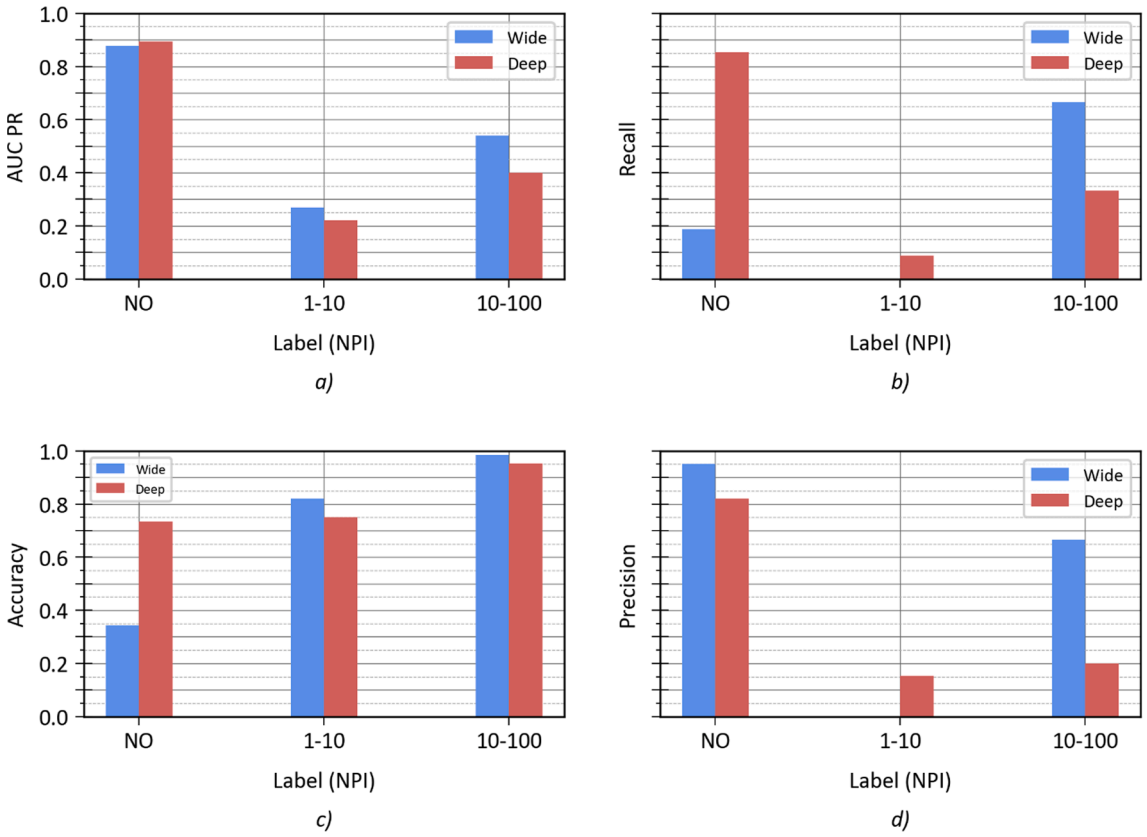


Fig. 5. Area Under the Curve Precision-Recall (a), Recall (b), Accuracy (c), and Precision (d) obtained from a small selection of simulations for the label category "Number of People Injured" (NPI). Labels are represented on the x-axis. Recall, Precision, and Accuracy are obtained at threshold = 0.5.

the source dataset is sufficient, but certainly not excellent considering the origin of the data (i.e., MHIDAS) and the limited details available. In addition, the target database has a high incidence of missing and uncertain values. The combination of relatively poor-quality input data and inherent model sensitiveness may have caused performance degradation. Differently, the more robust Wide model seems to have learned and assigned the right weights to the most significant and accessible features.

The results suggest that the approach benefits from a model capable of assessing the weights of each feature (or groups of features) independently, rather than generalizing over all the features. Linear models seem to be particularly suitable for addressing the problem considered in this study, especially when the dataset has uncertain and missing data. Future research should test whether higher-quality databases may improve the performance of the models. In addition, the Deep models may need more optimization and hyperparameters tuning to perform adequately. A different number of hidden units and layers, a different optimizer, learning decay, and optimization function may be tested to overcome the limitations of the Deep model and enhance its qualities.

In addition to these general considerations, the results in Fig. 5 and Fig. 6 offer interesting insights. A particular trend can be identified for the AUC PR curves: in most cases, the Area Under the Curve decreases as more critical events are considered. This fact is evident in Fig. 6.a, where AUCs decrease as a larger number of people involved is considered. Such behavior has also been observed and discussed by (Tamascelli et al., 2022). In fact, the knowledge gained by a classification algorithm largely depends on the quantity and quality of examples provided during the training phase. If the training database contains only a few examples of a particular label, the algorithms have little chance to learn. Since accidents with a high death toll are rare, the behavior of the AUC PR seems reasonable. Nevertheless, a few exceptions can be identified. For example, Fig. 5.a shows that the AUC PR produced by the models for the category 10–100 is larger than the AUCs for the category 1–10. In this case, the AUCs obtained for the most critical (and rare) label are unexpectedly large. This might be explained considering the extreme rarity of these events. In fact, only three events in the Ammonia database caused 10–100 injuries. Therefore, identifying two of these events would significantly improve the performance of the algorithm.

It is also worth noting that the accuracy follows a particular trend: the indicator tends to increase as more critical labels are considered. The reason for this is that when rare events are considered, high accuracy can be achieved by always performing a negative prediction. As an example, consider the label 10–100 in Fig. 6. Accuracy is almost 1 but Recall and Precision are 0 because the model never performed a positive prediction. In fact, the model made 127 correct predictions out of a total of 128 (only one event has 10–100 as a label). The accuracy is large, but the model failed to identify the critical event. This is an example of why accuracy alone is meaningless when considering unbalanced datasets.

As previously discussed, if the approach involves the identification of rare and critical events, a large Recall is desirable. Fig. 5 and Fig. 6 (and the rest of the results in the supplementary material) show Recall values obtained using a decision threshold equal to 0.5, which does not guarantee the best performance. A low Recall does not imply model inadequateness. Provided that the AUC PR is not zero, the decision threshold may be lowered to increase the Recall (as shown in Section 3.4.3).

As an example, consider the performance of the Deep model in Fig. 6. The Recall is zero for the label 1–10, and so is the Precision. This means that none of the fourteen events with label 1–10 were correctly identified. The model produced True Negatives and False Negatives only, as shown in Fig. 7 (i.e., the model never predicted the class "Y"). This happened because the raw probability values for the label "Y" were always smaller than 0.341, which is smaller than the standard decision threshold used to produce the metrics in Fig. 6.

However, the AUC PR is larger than 0 for the same model and label

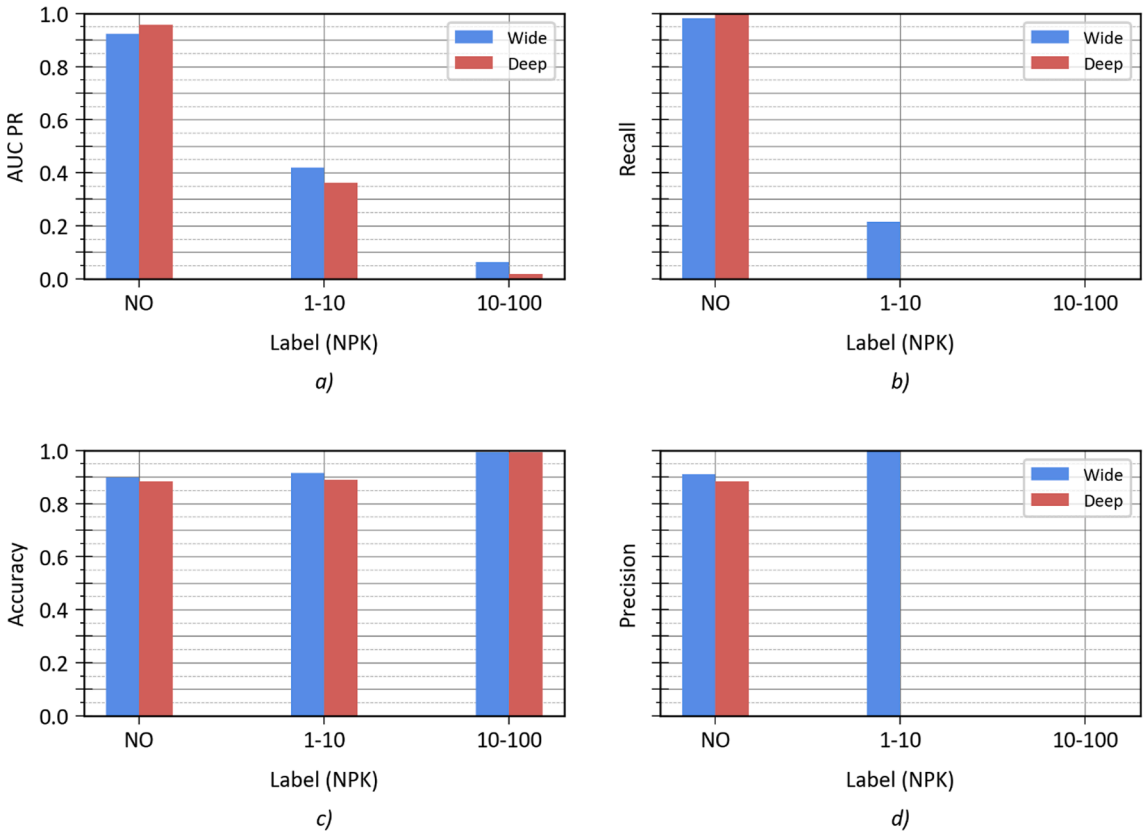


Fig. 6. Area Under the Curve Precision-Recall (a), Recall (b), Accuracy (c), and Precision (d) obtained from a small selection of simulations for the label category "Number of People Killed" (NPK). Labels are represented on the x-axis. Recall, Precision, and Accuracy are obtained at threshold = 0.5.

Table 4
Numbers of iteration steps used to obtain the results presented in Fig. 5 and Fig. 6. "NPI" and "NPK" respectively indicate the simulations for the Number of People Injured and Killed.

Models	Category	NO	1 – 10	10 – 100
Wide	NPI	200	200	2000
Deep	NPI	2000	20′000	20′000
Wide	NPK	200′000	20′000	2000
Deep	NPK	200	2000	200

Table 5
Scores assigned to the Wide and Deep model performances.

Model	Score NPI	Score NPK	Overall score
Wide	55	68	123
Deep	50	41	101

(Fig. 6.a). This suggests that Recall can be increased by lowering the decision threshold. The PR curve produced by this specific simulation has been shown in Fig. 3. The curve indicates that if the decision threshold is larger than 0.341, Precision and Recall will be zero because no positive prediction is generated (red mark in Fig. 3). If the decision threshold is decreased, more events are labeled as "Y", and more TP and/or FP are generated. In this example, the point at threshold = 0.317 (green in Fig. 3) appears to be a good balance between high Recall and acceptable Precision. The F-score analysis confirms this insight. Specifically, F_1, $F_{1.5}$, and F_2 curves are shown in Fig. 8.

A decision threshold equal to 0.317 maximizes the Recall-oriented $F_{1.5}$ and F_2 measures. Instead, F_1 shows a maximum for threshold = 0.3173, which has not been considered further. The number of TN, FP, TP, and FN obtained with threshold = 0.317 is displayed in Fig. 9.

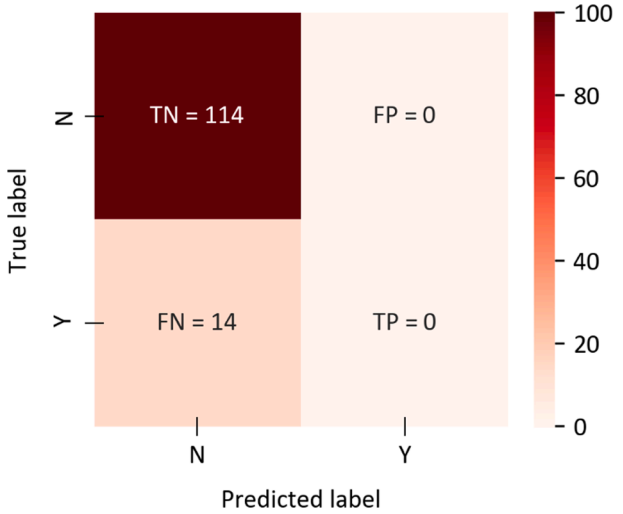


Fig. 7. Confusion Matrix produced by the Deep model for the label 1 – 10 (NPK) at 2000 integration steps. From top-left clockwise: True Negative (TN), False Positive (FP), True Positive (TP), and False Negative (FN) are obtained using a probability threshold equal to 0.5 and color-coded according to the color bar on the right.

The metrics in Fig. 9 indicate that 8 out of 14 events that caused 1–10 fatalities have been correctly identified (TP in Fig. 9). According to Eq. (5) and (4), the Recall is 0.57, and the Precision is 0.29, as shown in Fig. 3. As a drawback, reducing the threshold has generated 20 False Positives, whose nature has been studied:

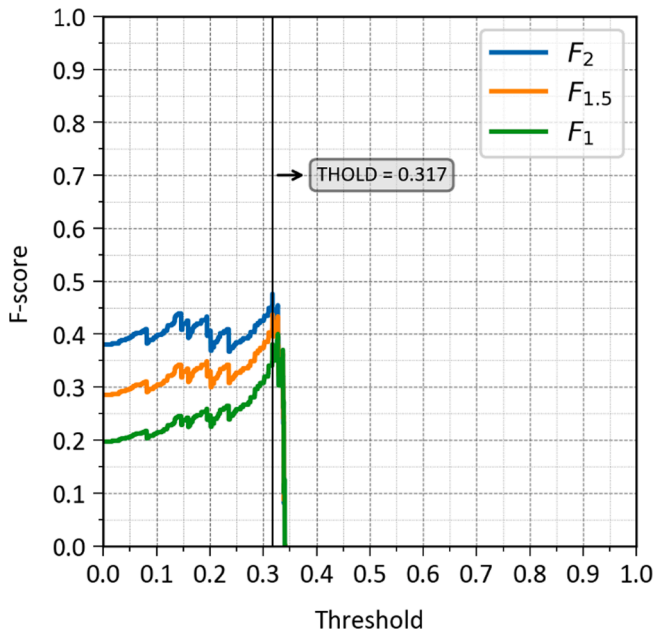


Fig. 8. F_1, $F_{1.5}$, and F_2 curves obtained by the Deep model for the label 1 – 10 (NPK) at 2000 integration steps. $F_{1.5}$, and F_2 show a global maximum for Threshold = 0.317. F_1 has a maximum at Threshold = 0.3173.

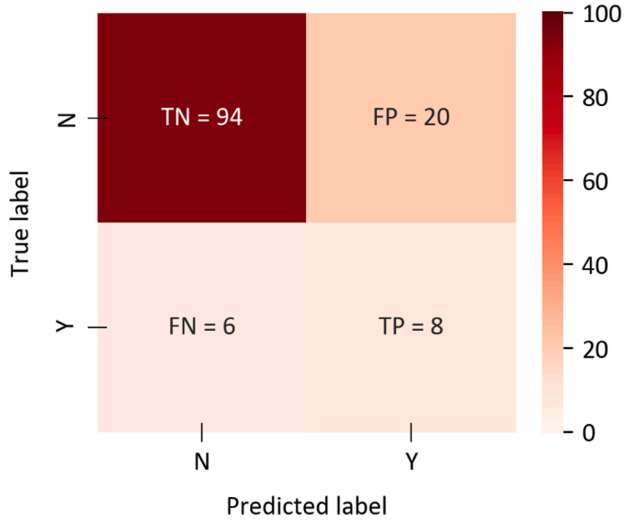


Fig. 9. Confusion Matrix produced by the Deep model for the label 1 – 10 (NPK) at 2000 integration steps. A decision threshold equal to 0.317 is used. From top-left clockwise: True Negative (TN), False Positive (FP), True Positive (TP), and False Negative (FN) are color-coded according to the color bar on the right.

- One of the False Positive involved the release of a relevant quantity of ammonia (10 to 100 tons) but did not cause any injuries or fatalities. In this case, the model might have mislabeled the event because large releases are more likely to cause at least one fatality. In fact, the model has labeled the event as potentially critical, which may be considered correct, even if the accident did not have tragic outcomes.
- Six events among the False Positives had caused 1–10 injured, which may indicate that those events had the potential to cause a low number of fatalities as well. Mislabeling these types of incidents is not deemed to be critical.
- Five False Positives have a large incidence of missing values (40 to 41 features out of 47 are not available). It seems reasonable and conservative to label uncertain events as potentially critical.

Hence, reducing the threshold to 0.317 has undeniably improved the performance considering that the same model produced null Precision and Recall (Fig. 7). It is worth mentioning that the model did not have any prior knowledge of the events included in the target task. Thus, the predictions rely entirely on the knowledge extracted from the source task. In this situation, a degree of uncertainty in the predictions appears to be reasonable. Therefore, it is not surprising that the metrics derived from the standard decision threshold are not satisfactory. However, the optimization process described in this example demonstrates that even if the pre-trained model may seem inadequate (i.e., a low Recall is produced), thresholding and F-score optimization may be used to improve the performance and fit the model to the target task. In general, the results suggest that the classification of rare and technology-specific accidents through Machine Learning may benefit from a meta-learning approach, which would enable knowledge transfer from generic and readily available accident databases. Furthermore, this approach may assist the industry in retaining the knowledge derived from past accidents more effectively.

In spite of the promising results, this study presents some limitations. Firstly, it must be recalled that this research has only considered accidents involving dangerous substances; therefore, the accident features described in Table 1 and the consequence categories proposed in Table 2 may not be suitable for different kinds of accidents. Nevertheless, the authors believe that the methodology is sufficiently generic to be extended to other industries and incidents. Another limitation is that the accident features presented in Table 1 may not be the most meaningful in this context. In fact, feature selection was manual and mainly guided by domain knowledge. Future research should investigate the effect of different sets of features and different feature representations. Secondly, this study has only examined two classification algorithms (i.e., linear and DNN); it may be worth testing different models such as Support Vector Machine (SVM), Decision Tree (DT), and Random Forest (RF). Furthermore, the DNN hyperparameters have not been optimized (e.g., number of layers and neurons, activation function, learning decay); more efforts should be directed toward hyperparameter tuning to identify the best network configuration. Thirdly, the Transfer Learning approach described in this study involves training on the source dataset and evaluation on the target database. This simple strategy was chosen to assess if the models could transfer knowledge to previously unseen events. It would be interesting to examine if an additional training phase on a small number of events drawn from the target dataset might significantly improve performance. Finally, further studies need to be carried out to investigate the potential of different learning strategies, such as regression or unsupervised learning.

Notwithstanding these limitations, the method described in this study might be used in combination with traditional techniques in different stages of the risk assessment and management framework. For example, the approach might be used to support the hazard identification phase, where information retrieval is critical, in order to avoid repeating mistakes in design or operations. Also, the ease of use and the intelligibility of results are interesting characteristics that may support the employees' training process and improve risk perception and awareness. Finally, the model might be a useful support for risk prioritization and residual risk management.

6. Conclusions

A data-driven method to extract, retain, and transfer knowledge from past industrial accidents involving dangerous substances is developed. Specifically, this study suggests that the knowledge extracted from generic accident databases might be used to predict the outcomes of technology-specific accidents in terms of injuries and fatalities. The

method has been tested on two datasets: MHIDAS and the Ammonia Plant Accident Database. Two different Machine Learning classification models (i.e., Wide and Deep) have been used. The Wide model offered the best performance in the Transfer Learning process. The challenges linked to the identification of rare and critical events have been discussed. An example of F-score optimization through thresholding has been described to stress the importance of threshold tuning in dealing with class-imbalanced datasets. Despite the limitations imposed by the quality and quantity of available data, the method leads to satisfactory performance. The results suggest that automated algorithms can learn from historical accident data sources and use the acquired knowledge to perform predictions on different types of accidents. The approach proposed in this study reduces the need for new data and improves the generalization capabilities of classification algorithms, and therefore makes an important contribution to the development of Machine Learning tools for improving process safety. More in general, the study indicates that improvements in IT and Industry 4.0 technologies offer interesting opportunities to integrate and support traditional risk assessment techniques with data-driven approaches, which are often faster to implement and cheaper in terms of working hours and required level of expertise. Furthermore, this study fits perfectly with the human-centric perspective of Industry 5.0 (Commission et al., 2021); ML techniques are not intended to substitute human judgment or threaten the role of safety practitioners. On the contrary, the methods proposed in this study have been designed to complement existing risk management techniques and provide practical support to workers.

CRediT authorship contribution statement

Nicola Tamascelli: Writing – review & editing, Supervision, Methodology, Conceptualization. **Nicola Paltrinieri:** Writing – review & editing, Supervision, Methodology, Conceptualization. **Valerio Cozzani:** Writing – review & editing, Supervision.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

References

AIChE, 2001. Ammonia Plant Safety (and Related Facilities), CEP technical manual. American Institute of Chemical Engineers.

Andrei, A.G., Balasa, R., Semenescu, A., 2022. Setting up new standards in aviation industry with the help of artificial intelligent-machine learning application. J. Phys. Conf. Ser. 2212 (1), 012014.

ARAMIS project team, 2004. Deliverable D.1.C.

Ashmore, R., Calinescu, R., Paterson, C., 2019. Assuring the machine learning lifecycle: Desiderata, methods, and challenges. arXiv.

Assi, K., Rahman, S.M., Mansoor, U., Ratrout, N., 2020. Predicting crash injury severity with machine learning algorithm synergized with clustering technique: A promising protocol. Int. J. Environ. Res. Public Health 17, 1–17. https://doi.org/10.3390/ijerph17155497.

BrinkS, H., Richards, J., Fetherolf, M., 2016. Real-World Machine Learning, first ed. Manning Publications, Shelter Island.

Bruha, I., 2017. Missing Attribute Values, in: Sammut, C., Webb, G.I. (Eds.), Encyclopedia of Machine Learning and Data Mining. Springer US, Boston, MA, pp. 834–841. https://doi.org/10.1007/978-1-4899-7687-1_954.

Bundesministerium für Umwelt Naturschutz Bau und Reaktorsicherheit, 2022. Central Reporting and Evaluation Office for Major Accidents and Incidents in Process Engineering Facilities - ZEMA [WWW Document]. URL https://www.infosis.uba.de/index.php/en/zema/index.html (accessed 8.28.20).

Bureau for Analysis of Industrial Risks and Pollutions, 2022. The ARIA Database - La référence du retour d'expérience sur accidents technologiques [WWW Document]. URL https://www.aria.developpement-durable.gouv.fr/the-barpi/the-aria-database/?lang=en (accessed 8.27.20).

Burnett, R.A., Si, D., 2017. Prediction of injuries and fatalities in aviation accidents through machine learning. ACM Int. Conf. Proceeding Ser. Part F1302, 60–68. https://doi.org/10.1145/3093241.3093288.

Cakir, E., Sevgili, C., Fiskin, R., 2021. An analysis of severity of oil spill caused by vessel accidents. Transp. Res. Part D Transp. Environ. 90, 102662 https://doi.org/10.1016/j.trd.2020.102662.

Carvalho, T.P., Soares, F.A.A.M.N., Vita, R., Francisco, R.d.P., Basto, J.P., Alcalá, S.G.S., 2019. A systematic literature review of machine learning methods applied to predictive maintenance. Comput. Ind. Eng. 137 https://doi.org/10.1016/j.cie.2019.106024.

Chebila, M., 2021. Predicting the consequences of accidents involving dangerous substances using machine learning. Ecotoxicol. Environ. Saf. 208, 111470 https://doi.org/10.1016/j.ecoenv.2020.111470.

Chen, T., Guestrin, C., 2016. XGBoost, in: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. ACM, New York, NY, USA, pp. 785–794. https://doi.org/10.1145/2939672.2939785.

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., 2016. Wide & deep learning for recommender systems. In: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, pp. 7–10.

Chinchor, N., 1992. MUC-4 Evaluation Metrics, in: Proceedings of the 4th Conference on Message Understanding, MUC4 '92. Association for Computational Linguistics, USA, pp. 22–29. https://doi.org/10.3115/1072064.1072067.

Chiong, M.-C., Kang, H.-S., Shaharuddin, N.M.R., Mat, S., Quen, L.K., Ten, K.-H., Ong, M. C., 2021. Challenges and opportunities of marine propulsion with alternative fuels. Renew. Sustain. Energy Rev. 149, 111397.

Choi, J., Gu, B., Chin, S., Lee, J.S., 2020. Machine learning predictive model based on national data for fatal accidents of construction workers. Autom. Constr. 110, 102974 https://doi.org/10.1016/j.autcon.2019.102974.

Chung, P.W.H., Jefferson, M., 1998. The integration of accident databases with computer tools in the chemical industry. Comput. Chem. Eng. 22 https://doi.org/10.1016/s0098-1354(98)00135-5.

Commission, E., Innovation, D.-G. for R. and, Breque, M., De Nul, L., Petridis, A., 2021. Industry 5.0 : towards a sustainable, human-centric and resilient European industry. Publications Office. https://doi.org/10.2777/308407.

European Commission, 2019. Ammonia release. URL https://emars.jrc.ec.europa.eu/en/emars/accident/view/891f340a-ac6d-11e9-bd0d-005056ad0167.

European Commission, 2022. eMARS Dashboard [WWW Document]. URL https://emars.jrc.ec.europa.eu/en/emars/content (accessed 8.27.20).

Donahue, J., Jia, Y., Vinyals, O., Hoffman, J., Zhang, N., Tzeng, E., Darrell, T., 2014. DeCAF: A deep convolutional activation feature for generic visual recognition. In: 31st Int. Conf. Mach. Learn. ICML 2014 2, pp. 988–996.

EU-OSHA, 1994. European Agency for Safety & Health at Work - Information, statistics, legislation and risk assessment tools. [WWW Document]. URL https://osha.europa.eu/en (accessed 8.28.20).

Gangadhari, R.K., Khanzode, V., Murthy, S., 2022. Application of rough set theory and machine learning algorithms in predicting accident outcomes in the Indian petroleum industry. Concurr. Comput. Pract. Exp. https://doi.org/10.1002/cpe.7277.

Ge, Z., Song, Z., Ding, S.X., Huang, B., 2017. Data Mining and Analytics in the Process Industry: The Role of Machine Learning. IEEE Access 5, 20590–20616. https://doi.org/10.1109/ACCESS.2017.2756872.

Gerassis, S., Saavedra, Á., Taboada, J., Alonso, E., Bastante, F.G., 2020. Differentiating between fatal and non-fatal mining accidents using artificial intelligence techniques. Int. J. Mining. Reclam. Environ. 34, 687–699. https://doi.org/10.1080/17480930.2019.1700008.

Goh, Y.M., Chua, D., 2013. Neural network analysis of construction safety management systems: a case study in Singapore. Constr. Manag. Econ. 31, 460–470. https://doi.org/10.1080/01446193.2013.797095.

Goldberg, D.M., 2022. Characterizing accident narratives with word embeddings: Improving accuracy, richness, and generalizability. J. Safety Res. 80, 441–455. https://doi.org/10.1016/j.jsr.2021.12.024.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning, Adaptive Computation and Machine Learning series. MIT Press.

Griffiths, T.L., Callaway, F., Chang, M.B., Grant, E., Krueger, P.M., Lieder, F., 2019. Doing more with less: meta-reasoning and meta-learning in humans and machines. Curr. Opin. Behav. Sci. 29, 24–30. https://doi.org/10.1016/j.cobeha.2019.01.005.

Han, J., Kamber, M., Pei, J., 2012. 8 - Classification: Basic Concepts. In: Han, J., Kamber, M., Pei, J.B.T.-D.M. (Eds.), The Morgan Kaufmann Series in Data Management Systems. Morgan Kaufmann, Boston, pp. 327–391. https://doi.org/10.1016/B978-0-12-381479-1.00008-3.

Harding, A.B., 1997. MHIDAS: The first ten years. Inst. Chem. Eng. Symp. Ser. 39–50.

Hastie, T., Friedman, R., Tibshirani, J., 2009. The Elements of Statistical Learning. Springer-Verlag New York. https://doi.org/10.1007/978-0-387-84858-7.

He, Q.P., Qin, S.J., Wang, J., 2005. A new fault diagnosis method using fault directions in Fisher discriminant analysis. AIChE J. 51, 555–571. https://doi.org/10.1002/aic.10325.

James, G., Hastie, T., Tibshirani, R., Witten, D., 2013. An Introduction to Statistical Learning: With Applications in R. Springer-Verlag, New York. https://doi.org/10.1007/978-1-4614-7138-7.

Japan Science and Technology Agency, 2005. Failure Knowledge Database [WWW Document]. URL http://www.shippai.org/fkd/en/index.html (accessed 8.27.20).

Jing, S., Liu, X., Gong, X., Tang, Y., Xiong, G., Liu, S., Xiang, S., Bi, R., 2022. Correlation analysis and text classification of chemical accident cases based on word embedding. Process Saf. Environ. Prot. 158, 698–710. https://doi.org/10.1016/j.psep.2021.12.038.

Kahraman, M.M., 2021. Analysis of Mining Lost Time Incident Duration Influencing Factors Through Machine Learning. Mining. Metall. Explor. 38, 1031–1039. https://doi.org/10.1007/s42461-021-00396-w.

Kalelkar, A.S., 1988. Investigation of large-magnitude incidents : Bhopal as a case study. IChemE. Prev. Major Chem. Relat. Process Accid. 553–575.

Khan, F.I., Abbasi, S.A., 1999. Major accidents in process industries and an analysis of causes and consequences. J. Loss Prev. Process Ind. 12, 361–378. https://doi.org/10.1016/S0950-4230(98)00062-X.
Khediri, I.B., Weihs, C., Limam, M., 2012. Kernel k-means clustering based local support vector domain description fault detection of multimodal processes. Expert Syst. Appl. 39, 2166–2171. https://doi.org/10.1016/j.eswa.2011.07.045.
Kurian, D., Sattari, F., Lefsrud, L., Ma, Y., 2020. Using machine learning and keyword analysis to analyze incidents and reduce risk in oil sands operations. Saf. Sci. 130, 104873 https://doi.org/10.1016/j.ssci.2020.104873.
Kushwaha, M., Abirami, M.S., 2022. Comparative Analysis on the Prediction of Road Accident Severity Using Machine Learning Algorithms. pp. 269–280. https://doi.org/10.1007/978-981-16-8721-1_26.
Landucci, G., Paltrinieri, N., 2016. A methodology for frequency tailorization dedicated to the Oil & Gas sector. Process Saf. Environ. Prot. 104, 123–141. https://doi.org/10.1016/j.psep.2016.08.012.
Langstrand, J.-P., Nguyen, H.T., McDonald, R., 2021. Applying Deep Learning to Solve Alarm Flooding in Digital Nuclear Power Plant Control Rooms, in: Ahram, T. (Ed.), Advances in Artificial Intelligence, Software and Systems Engineering. Springer International Publishing, Cham, pp. 521–527.
Le Coze, J.C., 2013. What have we learned about learning from accidents? Post-disasters reflections. Saf. Sci. 51, 441–453. https://doi.org/10.1016/j.ssci.2012.07.007.
Lees, F., 2004. Lees' Loss Prevention in the Process Industries, 3rd ed. Elsevier Butterworth_Heinemann, Burlington. https://doi.org/10.1016/C2009-0-24104-3.
Lemke, C., Budka, M., Gabrys, B., 2015. Metalearning: a survey of trends and technologies. Artif. Intell. Rev. 44, 117–130. https://doi.org/10.1007/s10462-013-9406-y.
Lonsdale, H., 1975. Ammonia Tank Failure - South Africa. Natal, South Africa.
Lu, J., Su, W., Jiang, M., Ji, Y., 2022. Severity prediction and risk assessment for non-traditional safety events in sea lanes based on a random forest approach. Ocean Coast. Manag. 225, 106202 https://doi.org/10.1016/j.ocecoaman.2022.106202.
Luo, X., Cruz, A.M., Tzioutzios, D., 2020. Extracting Natech Reports from Large Databases: Development of a Semi-Intelligent Natech Identification Framework. Int. J. Disaster Risk Sci. 11, 735–750. https://doi.org/10.1007/s13753-020-00314-6.
Makaba, T., Dogo, E., 2019. A Comparison of Strategies for Missing Values in Data on Machine Learning Classification Algorithms. In: Proc. - 2019 Int. Multidiscip. Inf. Technol. Eng. Conf. IMITEC 2019. https://doi.org/10.1109/IMITEC45504.2019.9015889.
Murphy, K.P., 2012. Machine Learning: A Probabilistic Perspective, Adaptive Computation and Machine Learning. MIT Press, Cambridge, Massachusetts, United States.
Nakhal, A., Patriarca, R., Di Gravio, G., Antonioni, G., Paltrinieri, N., 2021. Investigating occupational and operational industrial safety data through Business Intelligence and Machine Learning. J. Loss Prev. Process Ind. 73 https://doi.org/10.1016/j.jlp.2021.104608.
Palma, R., Martí, L., Sánchez-Pi, N., 2021. Predicting Mining Industry Accidents with a Multi-Task Learning Approach. In: 35th AAAI Conf. Artif. Intell. AAAI 2021 17B, pp. 15370–15376.
Paltrinieri, N., Comfort, L., Reniers, G., 2019. Learning about risk: Machine learning for risk assessment. Saf. Sci. 118, 475–486. https://doi.org/10.1016/j.ssci.2019.06.001.
Paltrinieri, N., Patriarca, R., Stefana, E., Brocal, F., Reniers, G., 2020. Meta-learning for safety management. Chem. Eng. Trans. 82 https://doi.org/10.3303/CET2082029.
Pan, S.J., Yang, Q., 2010. A survey on transfer learning. IEEE Trans. Knowl. Data Eng. 22, 1345–1359. https://doi.org/10.1109/TKDE.2009.191.
Pandya, J., 2020. Ammonia Gas Leaks At IFFCO Plant In Uttar Pradesh's Prayagraj; 2 Dead & 12 Admitted. Republicworld.com. URL https://www.republicworld.com/india-news/general-news/ammonia-gas-leaks-at-iffco-plant-in-uttar-pradeshs-prayagraj-2-dead-and-12-admitted.html.
Parmiggiani, E., Østerlie, T., Almklov, P.G., 2022. In the Backrooms of Data Science. J. Assoc. Inf. Syst. 23, 139–164. https://doi.org/10.17705/1jais.00718.
Pasman, H.J., 2009. Learning from the past and knowledge management: Are we making progress? J. Loss Prev. Process Ind. 22, 672–679. https://doi.org/10.1016/j.jlp.2008.07.010.
Pasman, H.J., Fouchier, C., Park, S., Quddus, N., Laboureur, D., 2020. Beirut ammonium nitrate explosion: Are not we really learning anything? Process Saf. Prog. 39 https://doi.org/10.1002/prs.12203.
Pattabathula, V., Richardson, J., 2016. Introduction to ammonia production. Chem. Eng. Prog. 112, 69–75.
Phark, C., Kim, W., Yoon, Y.S., Shin, G., Jung, S., 2018. Prediction of issuance of emergency evacuation orders for chemical accidents using machine learning algorithm. J. Loss Prev. Process Ind. 56, 162–169. https://doi.org/10.1016/j.jlp.2018.08.021.
Poh, C.Q.X., Ubeynarayana, C.U., Goh, Y.M., 2018. Safety leading indicators for construction sites: A machine learning approach. Autom. Constr. 93, 375–386. https://doi.org/10.1016/j.autcon.2018.03.022.
Rawson, A., Brito, M., 2022. A survey of the opportunities and challenges of supervised machine learning in maritime risk analysis. Transp. Rev. 1–23 https://doi.org/10.1080/01441647.2022.2036864.
Sarkar, S., Maiti, J., 2020. Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis. Saf. Sci. 131, 104900.
Sasaki, Y., 2007. The truth of the F-measure. Teach Tutor Mater 1–5.
Souza, P., Freitas, M.F., Machado, C., 1996. Major Chemical Accidents in Industrializing Countries: The Socio-Political Amplification of Risk. Risk Anal. 16, 19–29. https://doi.org/10.1111/j.1539-6924.1996.tb01433.x.
Stefana, E., Paltrinieri, N., 2021. ProMetaUS: A proactive meta-learning uncertainty-based framework to select models for Dynamic Risk Management. Saf. Sci. 138, 105238 https://doi.org/10.1016/j.ssci.2021.105238.
Tamascelli, N., Paltrinieri, N., Cozzani, V., 2020. Predicting Chattering Alarms: a Machine Learning Approach. Comput. Chem. Eng. 107122 https://doi.org/10.1016/j.compchemeng.2020.107122.
Tamascelli, N., Scarponi, G., Paltrinieri, N., Cozzani, V., 2021. A data-driven approach to improve control room operators' response. Chem. Eng. Trans. 86, 757–762. https://doi.org/10.3303/CET2186127.
Tamascelli, N., Solini, R., Paltrinieri, N., Cozzani, V., 2022. Learning from major accidents: A machine learning approach. Comput. Chem. Eng. 162, 107786 https://doi.org/10.1016/j.compchemeng.2022.107786.
Tanguy, L., Tulechki, N., Urieli, A., Hermann, E., Raynal, C., 2016. Natural language processing for aviation safety reports: From classification to interactive analysis. Comput. Ind. 78, 80–95. https://doi.org/10.1016/j.compind.2015.09.005.
Tauseef, S.M., Abbasi, T., Abbasi, S.A., 2011. Development of a new chemical process-industry accident database to assist in past accident analysis. J. Loss Prev. Process Ind. 24, 426–431. https://doi.org/10.1016/j.jlp.2011.03.005.
AEA Technology, 1999. MHIDAS (Major Hazard Incident Data Service.
TensorFlow.org, 2020a. tf.keras.optimizers.Ftrl | TensorFlow Core v2.1.0 [WWW Document]. URL https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Ftrl (accessed 4.25.20).
TensorFlow.org, 2020b. tf.keras.optimizers.Adagrad | TensorFlow Core v2.1.0 [WWW Document]. URL https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Adagrad (accessed 4.25.20).
TensorFlow.org, 2021. Overfit and underfit | TensorFlow Core [WWW Document]. URL https://www.tensorflow.org/tutorials/keras/overfit_and_underfit (accessed 6.28.21).
Tian, Y., Fu, M., Wu, F., 2015. Steel plates fault diagnosis on the basis of support vector machines. Neurocomputing 151, 296–303. https://doi.org/10.1016/j.neucom.2014.09.036.
Tixier, A.J.P., Hallowell, M.R., Rajagopalan, B., Bowman, D., 2016. Application of machine learning to construction injury prediction. Autom. Constr. 69, 102–114. https://doi.org/10.1016/j.autcon.2016.05.016.
Torrey, L., Shavlik, J., 2014. Transfer Learning, in: Handbook of Research on Machine Learning Applications and Trends. IGI Global, pp. 242–264. https://doi.org/10.4018/978-1-60566-766-9.ch011.
Union, E., 2012. L 197. Off. J. Eur. Union 55, 38–71. https://doi.org/10.3000/19770677.L_2012.197.eng.
United States Environmental Protection Agency, 2020. National Response System [WWW Document]. URL https://www.epa.gov/emergency-response/national-response-system (accessed 8.28.20).
Vanschoren, J., 2018. Meta-Learning: A Survey. arXiv.org 1–29.
Verma, R., Agnihotra, N., Dave, D., Naqvi, S., 2019. Ammonia, PEP Report 44C.
Wahab, L., Jiang, H., 2019. A comparative study on machine learning based algorithms for prediction of motorcycle crash severity. PLoS ONE 14, 1–17. https://doi.org/10.1371/journal.pone.0214966.
Wang, B., Zhao, J., 2022. Automatic frequency estimation of contributory factors for confined space accidents. Process Saf. Environ. Prot. 157, 193–207. https://doi.org/10.1016/j.psep.2021.11.004.
Weibull, B., Fredstrom, C., Wood, M.H., 2020. Learning lessons from accidents. Key points and conclusions for inspectors of major chemical hazard sites. Seveso Inspect. Ser. https://doi.org/10.2760/441934.
Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. IEEE Trans. Evol. Comput. 1, 67–82. https://doi.org/10.1109/4235.585893.
Xu, Z., Saleh, J.H., 2021. Machine learning for reliability engineering and safety applications: Review of current status and future opportunities. Reliab. Eng. Syst. Saf. 211, 107530 https://doi.org/10.1016/j.ress.2021.107530.
Xu, Z., Saleh, J.H., Subagia, R., 2020. Machine learning for helicopter accident analysis using supervised classification: Inference, prediction, and implications. Reliab. Eng. Syst. Saf. 204, 107210 https://doi.org/10.1016/j.ress.2020.107210.
Yang, X.-S., 2014. Introduction to Algorithms. Nature-Inspired Optim. Algorithms 1–21. https://doi.org/10.1016/b978-0-12-416743-8.00001-4.
Yedla, A., Kakhki, F.D., Jannesari, A., 2020. Predictive modeling for occupational safety outcomes and days away from work analysis in mining operations. Int. J. Environ. Res. Public Health 17, 1–17. https://doi.org/10.3390/ijerph17197054.
Zhang, X., Gweon, H., Provost, S., 2020. Threshold Moving Approaches for Addressing the Class Imbalance Problem and their Application to Multi-label Classification. ACM Int. Conf. Proceeding Ser. Part F16925, 72–77. https://doi.org/10.1145/3441250.3441274.
Zhang, J., Li, Z., Pu, Z., Xu, C., 2018. Comparing prediction performance for crash injury severity among various machine learning and statistical methods. IEEE Access 6, 60079–60087. https://doi.org/10.1109/ACCESS.2018.2874979.
Zhong, S., Wen, Q., Ge, Z., 2014. Semi-supervised Fisher discriminant analysis model for fault classification in industrial processes. Chemom. Intell. Lab. Syst. 138, 203–211. https://doi.org/10.1016/j.chemolab.2014.08.008.
Zhu, R., Hu, X., Hou, J., Li, X., 2021. Application of machine learning techniques for predicting the consequences of construction accidents in China. Process Saf. Environ. Prot. 145, 293–302. https://doi.org/10.1016/j.psep.2020.08.006.

Article IV.

Giannini, L., Tamascelli, N., Salzano, E., Paltrinieri, N. (2023). **Predicting the Consequences of Hydrogen Releases: how a Machine Learning Approach May Improve Risk-Based Inspection Planning**. Proceedings of the PSAM 2023 Topical Conference on AI & Risk Analysis for Probabilistic Safety/Security Assessment & Management. *Accepted for publication*.

This paper is submitted for publication and is therefore not included.

Article V.

Tamascelli, N., Scarponi, G.E., Amin, M. T., Sajid, Z., Paltrinieri, N., Khan, F., Cozzani, V., (2024). **A Neural Network Approach to Predict the Time-to-Failure of Atmospheric Tanks Exposed to External Fire**. Reliability Engineering & System Safety. https://doi.org/10.1016/j.ress.2024.109974.

A Neural Network Approach to Predict the Time-to-Failure of Atmospheric Tanks Exposed to External Fire

Nicola Tamascelli[1,2], Giordano Emrys Scarponi[1], Md Tanjin Amin[3], Zaman Sajid[3], Nicola Paltrinieri[2], Faisal Khan[3], Valerio Cozzani[1,*]

[1] *Department of Civil, Chemical, Environmental and Materials Engineering, Alma Mater Studiorum _ University of Bologna, via Terracini n.28, 40131 Bologna, Italy*

[2] *Department of Mechanical and Industrial Engineering, NTNU, Trondheim, Norway*

[3] *Mary Kay O'Connor Process Safety Center, Artie McFerrin Department of Chemical Engineering, Texas A&M University, College Station, TX 77843-3122, USA*

(*) corresponding author, e-mail: valerio.cozzani@unibo.it

Abstract

Domino scenarios triggered by fire pose severe risks to workers, assets, and the environment. Accurate quantitative models are needed to support mitigation actions addressing the prevention of fire escalation, especially considering sensitive targets such as atmospheric tanks containing large quantities of dangerous substances. A novel approach based on neural networks was developed, allowing the accurate quantification of the time-to-failure (TTF) of atmospheric tanks exposed to external fires accounting for mitigation actions. Data from a lumped parameter model were used to train and assess neural networks' performance. The toolbox of models obtained provides the TTF of atmospheric tanks both in the case of unmitigated fire scenarios and considering safety barriers and protection measures, such as water deluges and fire monitors. Model predictions are fast, accurate, and supplemented with confidence intervals. The comparative analysis demonstrated the better performance of the model developed compared to simplified correlations widely used in the literature to predict TTF. The approach developed, based on the integration of neural networks in consequence analysis tools, shows significant potential for the advancement of a quantitative assessment of domino scenarios, providing accurate and user-friendly tools for a quick evaluation of domino fire scenarios under both mitigated and unmitigated conditions.

Keywords: Domino effect; time-to-failure; Escalation; Cascading events; Fire; Artificial intelligence; Neural networks

1. Introduction

Fires are the most frequent type of accident in chemical facilities (Vipin et al., 2018). Fire scenarios have inherent characteristics that can jeopardize the safety of workers and nearby residents while also causing severe economic losses (Lees, 2012a). In addition, fire scenarios are a frequent cause of domino effects, triggering cascading events escalating into more severe accidents. Some of the most severe accidents in the last two decades indicate that fires play a leading role in domino scenarios (Abdolhamidzadeh et al., 2011; Huang et al., 2021). Remarkable examples are the accident that occurred in 2007 at the Valero refinery in Sunray, Texas, accounting for approximately $50 million in property damage and four injuries (Naderpour and Khakzad, 2018), the fires and explosions that occurred in 2009 in Jaipur, India, causing 13 deaths and more than 200 injuries (Fishwick, 2011), and the accident that took place in 2019 in Houston, USA, where 14 naphtha tanks burned for three days causing property damage exceeding $150 million (U.S. Chemical Safety and Hazard Investigation Board, 2023). Domino scenarios are described as low-probability high-impact (HILP) events and involve chains of events triggered by an initial incident, leading to a cascading effect with potentially severe consequences and elevating the potential for major accidents (Khan et al., 2021; Khan and Abbasi, 2001). The risk of domino effects escalates as chemical plants become more concentrated and densely packed, emphasizing the critical importance of proactive safety measures and risk assessments to prevent such cascading incidents (Cozzani and Reniers, 2021; Reniers and Cozzani, 2013).

Flammable substances are often stored in large atmospheric tanks (Lees, 2012a), which may contain up to 80000 m^3 of hazardous liquids. Domino scenarios involving storage tanks are considered particularly critical due to the quantity of the substance involved and the spacial vicinity between tanks in storage facilities. For this reason, it is vital to evaluate the response of atmospheric tanks when a fire occurs in their proximity.

When exposed to thermal loads, the tank shell is expected to undergo substantial deformation caused by thermal stresses and internal pressure build-up, which can lead to failure (Godoy, 2016; Godoy et al., 2023). For these reasons, passive (e.g., thermal insulation) and active devices (e.g., fixed water or foam sprays and water deluge systems) that mitigate the effect of the fire (Lees, 2012b) are frequently implemented. The time between the start of the fire and the possible failure of the tank, namely the time-to-failure (TTF), is thus influenced by the design features of the tank (e.g., diameter, height, and shell thickness), by operating parameters (e.g., filling degree, substance density, and vapor pressure), by the fire scenario (e.g., thermal radiation and view factor), and by the performance of passive and active safety systems installed.

The quantification of the TTF is critical for estimating the probability of failure and, eventually, for quantifying the risk associated with escalation resulting in domino scenarios. Hence, there is a compelling need to issue methods for the quantification of the TTF of atmospheric tanks considering (i) fire characteristics, (ii) tank characteristics, and (iii) safety barriers. Rigorous modeling through coupled computational fluid dynamics (CFD) and finite element modeling (FEM) can be used to evaluate the structural response of tanks exposed

to external fire (Yang et al., 2020). For example, Iannaccone et al. (2021) and Scarponi et al. (2019) utilized CFD modeling to simulate the pressure and temperature profiles of LNG and LPG pressurized tanks. Masum Jujuly et al. (2015) simulated the effect of a pool fire caused by LNG spill on multiple targets. A model for the dynamic evaluation of the fire response of steel storage tanks integrating CFD and FEM analysis was presented by (Li et al., 2022). Similar approaches were proposed by Wang et al. (2023) and Jianfeng Yang et al. (2023). However, the expertise in setting up the simulations and the substantial computational resources required to run them hinder their widespread application. In current practice, CFD and FEM simulations needs substantial data, and model validation is challenging and time-consuming within risk assessment studies. When considered, it is usually applied only for the deterministic analysis of single critical scenarios. To overcome this issue, Gubinelli (2005) proposed a lumped model called RADMOD to estimate the TTF of atmospheric tanks exposed to external fires.

The limited computational resources required to run RADMOD simulations result in a significant reduction of simulation time than in FEM simulations. However, the computational time required still makes this approach not suitable for advanced applications, such as dynamic risk analysis (Paltrinieri et al., 2015; Villa et al., 2016) and dynamic probabilistic risk assessment (Maidana et al., 2023), which often require the simulation of a large number of scenarios for the quantification of accident paths. To overcome this limitation, Landucci et al. (2009) used the results of RADMOD simulations to develop a simplified analytical correlation that explicitly correlates the dependency of the TTF on the heat load and tank volume. Recently, Yang et al. (2023) proposed an improved correlation where parameters were fitted on results from CFD and FEM simulations.

To the best of the authors' knowledge, the methodologies proposed by Landucci et al. (2009) and Yang et al. (2023) are the only currently available approaches capable of rapidly estimating the TTF for tanks exposed to external fire. This speed of estimation is crucial for integration into dynamic risk analysis frameworks of fire-induced scenarios – e.g., (Khakzad et al., 2014), (Ji et al., 2018), (Zeng et al., 2020), (Su et al., 2022), (Zhou and Reniers, 2022), (Ricci et al., 2024) – which all rely on the abovementioned correlations for the estimation of the TTF. However, notwithstanding the popularity of the approach proposed by Landucci et al. (2009), the following limitations must be acknowledged (Chen et al., 2018; Cui et al., 2022; Yang et al., 2018):

- The correlations above were derived from limited datasets based on a predefined set of tank geometries, which could restrict their generalization capabilities;
- The error in the TTF with respect to the original dataset values is relatively high, especially considering scenarios with large TTFs;
- They do not provide the confidence level of predictions;
- They tend to produce over-conservative forecasts;
- Only a single study attempted to incorporate the influence of safety barriers into the simplified correlations for the TTF (Landucci et al., 2015), still suffering from the limitations listed above.

In this context, Artificial Intelligence (AI) and Machine Learning (ML) present intriguing opportunities to address some of the limitations mentioned above. In fact, advanced ML algorithms can learn directly from failure data to develop predictive models that balance accuracy with the speed of predictions. In the field of safety and reliability, the application of ML is a relatively new yet promising area of research (Xu and Saleh, 2021). Recent studies have seen a surge in research on fault detection and diagnosis (Abid et al., 2021; H. Wang et al., 2023), anomaly detection (Nassif et al., 2021; Zhang et al., 2024), system prognosis (Huang et al., 2023; Xia et al., 2018), reliability analysis (Payette and Abdul-Nour, 2023; Roy and Chakraborty, 2023), and risk analysis (Hegde and Rokseth, 2020). Particularly in system prognosis, numerous studies have proposed using ML to estimate the Remaining Useful Life (RUL) of degrading equipment, including lithium batteries (Bai et al., 2023), bearings (J. Li et al., 2024), induction motors (Huang et al., 2023), turbofans (Arias Chao et al., 2022), wind turbines (Cao et al., 2023). However, there remains a significant gap in the literature regarding the application of ML for predicting the impact of fires on atmospheric tanks.

In addition, recent studies have explored the use of ML techniques to develop surrogate models based on data from Computational Fluid Dynamics (CFD) and Finite Element Method (FEM) simulations (Calzolari and Liu, 2021; Kudela and Matousek, 2022). For instance, Li et al. (2024) applied a graph neural network to predict the overpressure resulting from simulated Boiling Liquid Expanding Vapor Explosions (BLEVE). Similarly, Ye and Hsu (2022) used CFD and FEM data to model 1200 fire scenarios in a steel roof structure with fixed geometry and then employed this data to train a Long Short-Term Memory (LSTM) Neural Network to predict structural displacement. However, to the best of the authors' knowledge, there is a notable lack of research specifically targeting the estimation of the TTF of atmospheric tanks exposed to external fires using ML. Only one recent contribution, concurrent with our research, proposes a ML-based method to estimate the structural integrity of tanks under fire conditions (Amin et al., 2024a). Yet, this method primarily focuses on calculating failure probability rather than directly estimating TTF, and does not consider the effect of mitigation barriers.

The present study has developed a novel approach to bridge the aforementioned gaps in the literature. Neural networks (NN) were used to quantify the TTF of atmospheric tanks exposed to external fires. Simulated data from the RADMOD model were used to build an extensive dataset containing 4896 scenarios with different tank geometries and fire characteristics. The influence of safety barriers and systems is included in the model, considering a comprehensive set of protection measures characterized by various activation times and effectiveness. A model toolbox was obtained, allowing the estimation of the TTF of both unmitigated scenarios and considering the effect of different protection measures. Hyperparameters fine-tuning was performed to ensure optimal performance and good generalization capabilities.

Furthermore, to enhance the model output and account for prediction uncertainties, a model-agnostic approach was applied to estimate confidence intervals. This way, the model returns a single-point prediction for the TTF and a range of values with a specified confidence level. This comprehensive output captures the

inherent uncertainty in the data and the model, providing a more informative and robust assessment. The novelty contributions of this study can be summarized as follows:

- the proposed NN-based toolbox approach represents a novel, user-friendly, and computationally inexpensive for the estimation of the TTF of atmospheric tanks exposed to external fire;
- the approach allows for the explicit modeling of the impact of safety barriers, enabling a detailed and accurate analysis of their effectiveness in various scenarios;
- the use of a comprehensive dataset encompassing various tank geometries, fire characteristics, and barrier configurations improves model robustness and generalizability;
- confidence intervals enhance the interpretability, robustness, and credibility.

A comparative analysis between the model presented in this study and the simplified correlations proposed by Landucci et al. (2009) and Yang et al. (2023) shows that the new models developed have better performance.

The paper is organized into five sections, including this introductory paragraph. Section 2 outlines the methodology, addressing data generation and preprocessing, developing the NN models and their evaluation, fine-tuning hyperparameters, and generating confidence intervals. Section 3 presents the results, showcasing the performance obtained by the models and providing a comparative analysis with similar approaches found in the literature. Results are discussed in Section 4, and conclusions are drawn in Section 5. The models developed are included in the Supplementary materials and are freely available for use by researchers ("TTF_unmitigated.dill" and "TTF_mitigated.dill", "TTF_deluge.dill"). Also, a quick guide ("Model Configuration and Usage.pdf") is provided to prepare a Python environment for importing and using the models.

2. Methodology

The methodology used to build the NN models is schematized in Figure 1. First, a lumped parameter model known as RADMOD (Gubinelli, 2005) was used to create two datasets (step 1 in Figure 1): one containing failure data related to unmitigated fire scenarios ($\mathcal{D}$) and the other incorporating the effect of protective measures ($\overline{\mathcal{D}}$). These datasets were then preprocessed and split into two parts (step 2 in Figure 1). The first part (i.e., $\mathcal{D}_{train}$ and $\overline{\mathcal{D}}_{train}$ in Figure 1) was used to train the models, while the second part (i.e., $\mathcal{D}_{eval}$ and $\overline{\mathcal{D}}_{eval}$ in Figure 1) was used to evaluate their prediction performance. Subsequently, the models' hyperparameters were fine-tuned through a grid-search procedure to ensure optimal performance (step 5 in Figure 1). The resulting optimized models, namely $\mathcal{M}^*$ and $\overline{\mathcal{M}}^*$, can be utilized to predict the TTF of unmitigated and mitigated fire scenarios. In addition, a reference version of the model $\overline{\mathcal{M}}^*$ is offered to account for the influence of a standard generic water deluge system, useful in preliminary screening activities (step 6 in Figure 1). Finally, a procedure for estimating confidence intervals was implemented to provide

more informed judgments based on the confidence level associated with the predictions (step 7 in Figure 1). A detailed description of each step of the methodology is provided in the following.

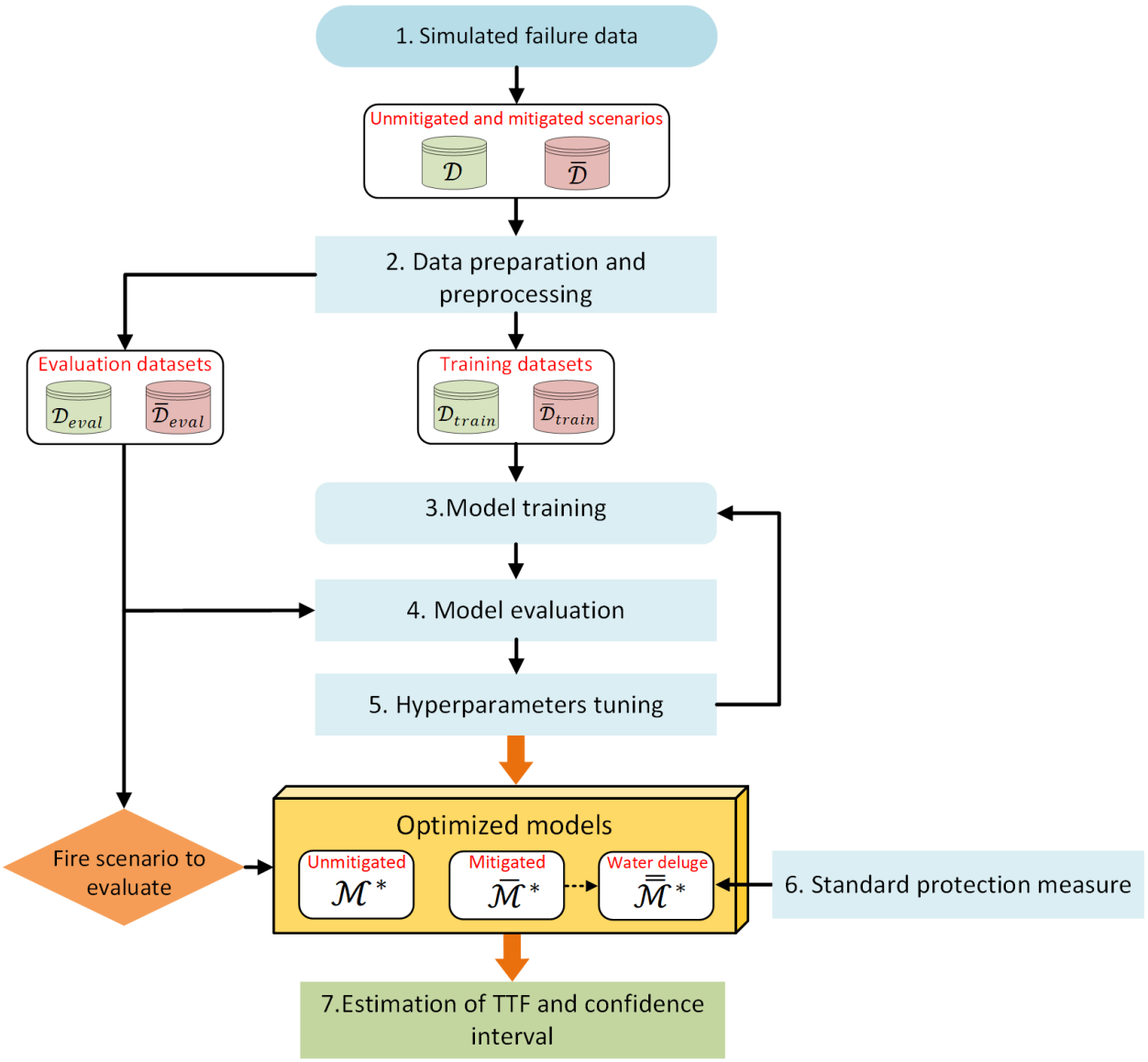


Figure 1. Flowchart of the proposed methodology. $\mathcal{D}_{train}$ and $\overline{\mathcal{D}}_{train}$ indicate the training datasets that contain respectively unmitigated and mitigated fire scenarios. Similarly, $\mathcal{D}_{eval}$ and $\overline{\mathcal{D}}_{eval}$ represent unmitigated and mitigated scenarios used to evaluate the models. $\mathcal{M}^*$ and $\overline{\mathcal{M}}^*$ indicate the models for the prediction of unmitigated and mitigated TTFs. $\overline{\overline{\mathcal{M}}}^*$ represents the simplified model that considers a single standard protection measure (water deluge), introduced as a case study.

2.1. Simulated failure data

Two series of simulations were carried out to generate failure data for different atmospheric tanks subject to various fire conditions. The first set of simulations focused on unmitigated scenarios, while the second set concentrated on mitigated scenarios. A widely used lumped parameter model, known as RADMOD (Gubinelli, 2005), was applied to simulate the behavior of atmospheric and pressurized tanks under fire exposure. Several studies are available in the literature in which this lumped model is applied in the assessment of the TTF for atmospheric tanks exposed to fire (e.g., (Cozzani et al., 2006), (Chen et al., 2018), (Su et al., 2022), (Jiahao Yang et al., 2023), (Amin et al., 2024a), (Amin et al., 2024b)). RADMOD partitions the computational

domain into two fluid nodes (the liquid and vapor) and two solid nodes (the liquid-wetted wall and the vapor-wetted wall). It requires design (shell diameter, height, and thickness), operational (filling level), and external (heat flux) parameters as inputs. The outputs of the simulations are node temperatures, tank pressure (accounting for the liquid head), axial stress, yield stress, and the equivalent von Mises stress in the steel structure. A detailed description of the RADMOD may be found in (Gubinelli, 2005; Landucci et al., 2009). The model was validated against TTF results obtained using finite element analysis (FEM). The results are shown in Figure 2, revealing an average relative error of 15% between the TTF calculated using RADMOD and the TTF obtained from FEM simulations.

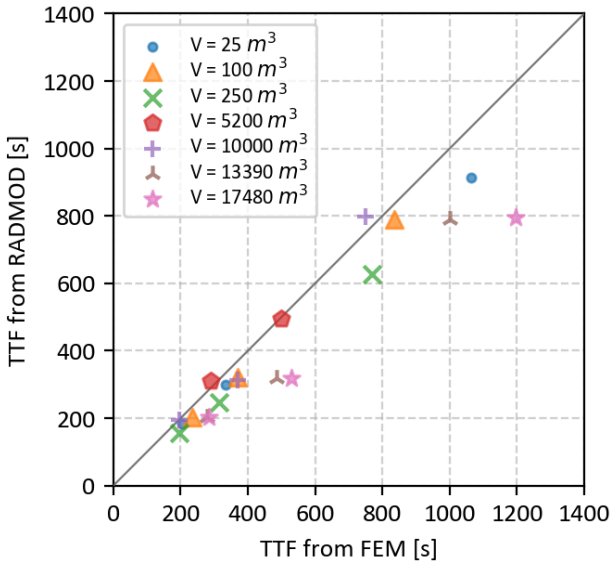


Figure 2. Comparison between the TTF obtained from RADMOD and the TTF obtained from FEM analysis for atmospheric tanks of various volumes (Landucci et al., 2009)

The simulation of fire scenarios requires the definition of a set of parameters to describe (i) the tank design and operating parameters, (ii) the characteristics of the external fire, and (iii) the mitigation measures. Table 1 presents a summary of these parameters, along with the corresponding values utilized in the simulations.

	Parameter	Values
Tank (geometric)	External diameter* [m]	20 values between 3 and 66
	Wall thickness* [m]	38 values between 0.005 and 0.012
	Height* [m]	10 values between 1.8 and 18
Tank (loading)	Filling level [%]	{20, 50, 80}
Fire	Total heat flux [kW/m^2]	{10, 20, 30, 40, 50, 60, 70, 80, 90, 100, 110, 125}
Mitigation measures	Activation time [s]	Randomly sampled between 30 and 90
	Damping factor	Randomly sampled between 0.2 and 0.8

*Table 1. The set of variables that defines a fire scenario and their values used in the simulations. * indicates that the values are taken from Table A-2a of API 650* (American Petroleum Institute, 2021)

The first three parameters in Table 1 define the geometry of the tank. Their values are taken from Table A-2a of API 650: "Welded Steel Tanks for Oil Storage" (American Petroleum Institute, 2021), which defines the shell-plate thickness of typical-sized tanks for oil storage. The table includes a comprehensive list of 136 tank sizes, each characterized by specific values of shell diameter, height, and thickness.

The fourth parameter, filling level, in Table 1 defines the tank filling level, for which three different values were considered: 20%, 50%, or 80%. Therefore, a total of 408 reference tanks were considered for the simulation, each defined by a set of geometric and filling parameters. The internal pressure and temperature were initialized to 1 bar and 20°C, respectively. The substance inside the tank was modeled as benzene, a liquid at 20°C with a density of 878 kg/m³ and a vapor pressure of 9913 Pa (Green and Perry, 2008).

The fifth parameter, total heat flux, in Table 1 describes the fire conditions defined by the heat flux to the tank wall. Twelve distinct heat flux values were considered, ranging from 10 kW/m² to 125 kW/m². The lower bound, 10 kW/m^2, corresponds to the radiation threshold at which the TTF of atmospheric equipment exceeds 30 minutes, providing adequate time for effective mitigation without incurring in critical damage (Cozzani et al., 2006). Conversely, the upper bound, 125 kW/m^2, represents a credible heat flux value to a target in severe tank fire scenarios, able to cause target damage in a limited time span, so that effective mitigation is not credible. Thus, the region between the lower and upper bound radiation values identified is that where the accuracy of TTF estimation is more critical in order to assess the time available for effective mitigation and the actual risk of escalation leading to domino effect.

The combination between reference tanks and the fire characteristics led to the definition of 4896 simulated fire scenarios using the RADMOD model. The results were collected in dataset $\mathcal{D}$, which comprises 4896 rows (i.e., number of fire scenarios) and 6 columns; the first 5 columns represent the features of the fire scenario (i.e., diameter, thickness, height, filling level, and total heat flux), while the last column indicates the TTF calculated by the RADMOD model.

An additional set of simulations was performed to incorporate the effects of mitigation measures on the same fire scenarios as defined above. Specifically, the last two parameters in Table 1 define the effect of the safety barriers adopted to mitigate the impact of the fire. The "activation time" represents the time required to activate the barrier, and the "damping factor" indicates the reduction in the total heat load caused by the activation of the safety barrier. Therefore, the net incident radiation at time t can be calculated as

$$\bar{q}(t) = \begin{cases} q & if\ t < t_a \\ q(1-\alpha) & if\ t \geq t_a \end{cases}, \quad (1)$$

Where $\bar{q}(t)$ indicates the net incident radiation on the tank wall at time t, q represents the total heat flux defined in Table 1, t_a indicates the barrier activation time, and $\alpha \in [0, 1]$ is the damping factor. Therefore, the effect of the mitigation measures is to reduce by a factor α the net incident radiation after t_a seconds. Activation time and dumping values were chosen based on reference values for automatic devices, such as fixed water sprays. Specifically, activation times are randomly sampled between 30s and 90s, while dumping

factors are sampled between 0.2 and 0.8. Each of the 4896 fire scenarios described earlier was associated with a specific barrier configuration (i.e., t_a and α), and RADMOD was used to simulate the mitigated scenarios. Results were collected in dataset $\overline{\mathcal{D}}$, which contains the TTF values of mitigated scenarios corresponding to the fire scenarios in dataset $\mathcal{D}$. Therefore, the $\overline{\mathcal{D}}$ dataset contains the same number of rows as $\mathcal{D}$, but it has two more columns, respectively reporting the activation time and the damping factor of the mitigation measures.

2.2. Data preparation and preprocessing

The two datasets described in the previous section served as the basis for the application of the methodology. However, before developing the machine learning (ML) models, the datasets needed to be preprocessed to arrange the data in a suitable format and facilitate the upcoming analysis. The datasets were split into two parts. The first part was used to train the ML models, while the second was used to evaluate their performance. The dataset $\mathcal{D}$ was split into $\mathcal{D}_{train}$ and $\mathcal{D}_{eval}$, where the former comprises 80% of the scenarios included in $\mathcal{D}$, and the latter contains the remaining observations. Similarly, $\overline{\mathcal{D}}$ was split into $\overline{\mathcal{D}}_{train}$ and $\overline{\mathcal{D}}_{eval}$, respectively comprising 80% and 20% of the events in $\overline{\mathcal{D}}$.

Finally, data were normalized to have zero mean and unit standard deviation. Specifically, a value x in the j-th column of the dataset is normalized as

$$z = \frac{x - \mu_j}{\sigma_j} \tag{2}$$

Where z indicates the normalized value, μ_j represents the mean of column j and σ_j is the standard deviation of column j. This procedure, called Z-score normalization, is a widely used data preprocessing technique that has been demonstrated to improve the generalization and convergence of neural networks (Goodfellow et al., 2016). It is worth mentioning that only the features of fire scenarios were normalized (i.e., the TTF was not normalized). In addition, the normalization of the evaluation datasets was performed using the mean and standard deviation of the training datasets to avoid any information leakage between training data and evaluation data.

2.3. Model training

The datasets $\mathcal{D}_{train}$ and $\overline{\mathcal{D}}_{train}$ were used to develop two distinct neural network models. The first model, trained on $\mathcal{D}_{train}$, aims at predicting the TTF of unmitigated fire scenarios. The second model, trained on $\overline{\mathcal{D}}_{train}$, aims at predicting the TTF of mitigated scenarios. The models were developed using Keras version 2.11.0 in Python version 3.10.11.

In this study, fully-connected feed-forward neural networks (FC-FFNNs) were used to model the relationship between scenario features and the TTF. FC-FFNNs were preferred over other traditional regression models (e.g., Linear regression, Support Vector Regression) due to their excellent abstraction and generalization

capabilities (Goodfellow et al., 2016; Hastie et al., 2009). In addition, FC-FFNNs offer a good balance between simplicity and effectiveness, making them an ideal choice over more advanced and complex algorithms. In fact, FC-FFNNs are less computationally demanding, making them suitable when resources are limited or for tasks that do not require complex models. FC-FFNNs also tend to generalize well and are less prone to overfitting, especially with limited data. Their training is more straightforward, involving fewer hyperparameters, which saves time and effort in model development.

A schematic representation of an FC-FFNN is presented in Figure 3.

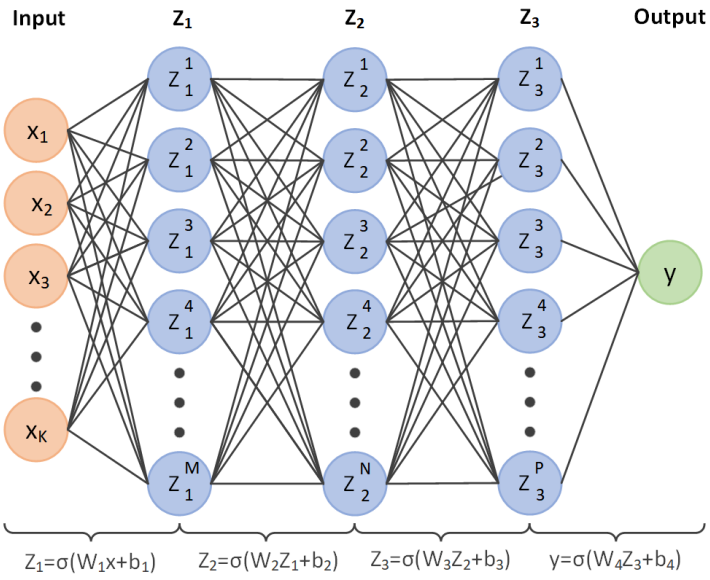


Figure 3. A schematization of an FC-FNN with three hidden layers. The input of the model (i.e., the features) is depicted in orange. Hidden layers (Z_i) and units (Z_i^j) are represented in blue. The model output (i.e., the response) is depicted in green.

FC-FFNNs are directed acyclic graphs that comprise an input layer (orange in Figure 3), one or more hidden layers (blue in Figure 3), and one output layer (green in Figure 3). Each layer comprises one or more units (circles in Figure 3), which are real-valued entities. The units of the input layer are the features of an observation (e.g., the parameters in Table 1). In contrast, the units of the hidden layers, also called hidden units or hidden neurons, are calculated through a nonlinear transformation of the linearly combined units in the previous layers (Hastie et al., 2009). Specifically, a generic layer $Z_i \in \mathbb{R}^{M\times 1}$ with M neurons is computed as follows:

$$Z_i = \sigma(W_i \cdot Z_{i-1} + b_i), \tag{3}$$

Where σ is the rectified linear unit (ReLu) function (Sharma et al., 2017), $W_i \in \mathbb{R}^{N\times M}$ is the matrix of the weights, $Z_{i-1} \in \mathbb{R}^{N\times 1}$ is the layer preceding Z_i, and $b_i \in \mathbb{R}^{M\times 1}$ is the vector of the biases. Weights and biases are learnable parameters that are tuned during training. The network's last layer is the output layer (green in Figure 3), representing the response variable (i.e., the TTF). Therefore, the NN aims at identifying the function f that approximates the relationship between features and response:

$$y \approx f(x) \tag{4}$$

The function f comprises a set θ of learnable parameters, namely the weights and the biases introduced in Eq. (3). Such parameters are randomly initialized and eventually tuned during the training procedure to minimize the error between the model's predictions and the actual value of the TTF. The training dataset (i.e., $\mathcal{D}_{train}$ or $\overline{\mathcal{D}}_{train}$) are used to train the models. Specifically, examples of both features and related TTFs are fed to the model. The model uses the features to calculate the predicted TTF, which is compared to the actual TTF, and the resulting error is back-propagated to update the learnable parameters θ. The procedure is iterative and aims to find the best set of hyperparameters, θ^*, that minimizes the error between predicted TTFs ($\tilde{y}$) and true TTFs (y),

$$\theta^* = \underset{\theta}{argmin}\left[\ell\left(y, \tilde{y}(\theta)\right)\right], \tag{5}$$

Where ℓ indicates the "loss", a function that measures the error between predictions and true responses. In this study, the mean squared error (MSE), defined as $MSE = {}^{1}/_{N}\sum_{i=1}^{N}(y_i - \tilde{y}_i)^2$, is used as a loss function, where $N \in \mathbb{N}$ is the number of scenarios included in the training datasets. The MSE is a differentiable loss function, which helps with network convergence. However, it is highly susceptible to the presence of outliers, as they have the potential to heavily distort the squared differences between predicted and actual values, resulting in inflated error measurements. In the datasets utilized for this study, outliers are not present, thereby making the MSE an appropriate choice as a loss metric.

As mentioned earlier, $\mathcal{D}_{train}$ or $\overline{\mathcal{D}}_{train}$ are utilized to construct two independent models, namely $\mathcal{M}$ and $\overline{\mathcal{M}}$, where their parameters are optimized to predict the TTF of unmitigated and mitigated fire scenarios, respectively. Assuming a successful training procedure, the models should be capable of predicting the TTF of fire scenarios contained in $\mathcal{D}_{train}$ or $\overline{\mathcal{D}}_{train}$ with minimal error. However, there is no guarantee that the models will maintain the same level of accuracy when considering new fire scenarios. Hence, evaluating the models using independent scenarios is crucial to ensure their ability to generalize the knowledge acquired during training to previously unseen events. The evaluation procedure will be described in the following section.

2.4. Model evaluation

The trained models are evaluated on their ability to predict the TTF of scenarios included in $\mathcal{D}_{eval}$ and $\overline{\mathcal{D}}_{eval}$. The features of the fire scenarios included in $\mathcal{D}_{eval}$ are fed to the model $\mathcal{M}$, which predicts the TTFs based on the knowledge extracted from $\mathcal{D}_{train}$. Similarly, the model $\overline{\mathcal{M}}$ is evaluated on $\overline{\mathcal{D}}_{eval}$. It is worth noting that only the features of fire scenarios are fed to the model during the evaluation phase, as opposed to the training phase, where both features and TTFs were used.

The evaluation of the models involves the calculation of performance metrics to quantify the quality of predictions. In this study, we utilize the following metrics:

- Coefficient of determination (R^2)

$$R^2 = 1 - \frac{\sum_{i=1}^{M}(y_i - \tilde{y}_i)^2}{\sum_{i=1}^{M}(y_i - \mu)^2} \quad (6)$$

Where, $M \in \mathbb{N}$ is the number of fire scenarios included in the evaluation dataset, y_i and $\tilde{y}_i$ are respectively the true and predicted TTF values, and μ is the mean of the true TTFs.

- Root mean squared error (RMSE)

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M}(y_i - \tilde{y}_i)^2} \quad (7)$$

Typically, the coefficient of determination assumes values between 0 and 1, although negative values are possible when the model performs worse than a constant function that simply predicts the mean of the data. Values of R^2 close to 1 are often interpreted as a good quality of the fit, which is generally true. However, it is important to remark that there may be specific circumstances in which large R^2 values do not necessarily guarantee the usefulness of the regression model (Hahn, 2007).

For this reason, the RMSE is also considered to overcome the limitations of R^2. The $RMSE$ may be interpreted as the square root of the variance of the residual. Therefore, it indicates how spread out these residuals are. This metric is bounded below by 0 but has no upper bound. Values of $RMSE$ close to 0 indicate good prediction performances. However, it must be noted that $RMSE$ shares the same units as the response variable (e.g., "seconds" for TTF). Therefore, when evaluating model performance, it is essential to consider the magnitude of the response variable in comparison to the $RMSE$. Often, to remove the effect of scale and enable a fair comparison between models trained on different tasks, the $RMSE$ is normalized using the mean of the response variable. This normalized version is usually defined as the normalized RMSE (NRMSE):

$$NRMSE = \frac{RMSE}{\mu} \quad (8)$$

If the objective is to compare the performance of two or more models on the same task, the RMSE can be used directly. In such cases, the model with the lowest RMSE is considered to outperform the other models.

2.5. Hyperparameters tuning

The procedures described in Sections 2.3 and 2.4 illustrate how to train and evaluate the models $\mathcal{M}$ and $\overline{\mathcal{M}}$ for predicting the TTF of unmitigated and mitigated scenarios. However, no details about the network hyperparameters have been provided. The hyperparameters of a NN may be defined as non-trainable parameters that define its structure and behavior during training (Yu and Zhu, 2020). For example, the number of hidden layers and neurons per layer are hyperparameters that can significantly influence the model's performance. Unfortunately, selecting the right set of hyperparameters is mainly guided by background knowledge and experimentation (Yang and Shami, 2020), and there is no "golden rule" to define an optimal set of hyperparameters in hindsight.

In this study, a "grid-search" procedure (Liashchynskyi and Liashchynskyi, 2019) was used to tune the network hyperparameters. This procedure involves (i) defining a search space for each parameter, (ii) training and evaluating one model for each unique combination of hyperparameters, and (iii) comparing the performance of the models and selecting the best set of hyperparameters.

The search space is defined by specifying the range or discrete values for each hyperparameter under consideration. In this study, we focus on three hyperparameters: the number of layers, the number of neurons per layer, and the learning rate (LR). The number of layers and neurons defines the network structures, as discussed in Section 2.3. The Learning Rate is a key parameter determining the step size at which the model updates its parameters during training. A large learning rate increases the convergence speed but raises the risk of the model diverging and failing to converge to an optimal solution. The search space used in this study is defined in Table 2.

Hyperparameter	Values
Number of layers	{1, 2, 3, 4}
Number of neurons per layer	{2, 52, 102, 152, 202, 252, 302}
Learning Rate	{0.01, 0.001}

Table 2. The search space used in the grid-search procedure.

The combination of the hyperparameters leads to the definition of 5600 model configurations, each characterized by several layers, one or more neurons per layer, and one learning rate. The grid search algorithm systematically evaluates the model performance for every possible combination within the defined search space. Specifically, each model configuration is used to train and assess the models $\mathcal{M}$ and $\overline{\mathcal{M}}$ as described in Sections 2.3 and 2.4. The training procedure was conducted with 300 epochs, which indicates the number of iterations made by the model over the training dataset. The selected number of epochs represents a good balance between computational complexity and model accuracy. A larger number of epochs would significantly increase the computation time, while fewer epochs may penalize the complex models with low learning rates because they typically require more training samples to perform adequately. As a result, every model configuration is associated with an $RMSE$ value that reflects its performance in predicting the TTF for unmitigated events, as well as another $RMSE$ value for predicting mitigated TTFs. The model that obtains the lowest $RMSE$ on $\mathcal{D}_{eval}$ identifies the best configuration for the prediction of unmitigated TTFs, while the model that obtains the lowest $RMSE$ on $\overline{\mathcal{D}}_{eval}$ identifies the best configuration for the prediction of mitigated TTFs. The two tasks (i.e., prediction of the TTF for unmitigated and mitigated scenarios) are treated independently because there is no guarantee that the best model for the prediction of unmitigated TTFs will also show superior performance on the prediction of mitigated TTFs (Wolpert and Macready, 1997).

The grid search procedure creates two models optimized for predicting the TTF of unmitigated and mitigated scenarios. The first model, namely $\mathcal{M}^*$, takes as an input the first five parameters in Table 1, while the second model, namely $\bar{\mathcal{M}}^*$, requires the complete set of parameters.

It is worth mentioning that the selection of the search space in Table 2 has been mainly guided by background knowledge, and it is not meant to identify the best model in absolute terms. In fact, there are additional hyperparameters that may require training, such as the choice of the activation function (σ in Eq. (3)), and other parameters not covered in this study, including batch size, regularization layers, and optimizers (Yu and Zhu, 2020). Nonetheless, the search space in Table 2 encompasses the parameters that the authors consider as the most crucial for addressing the specific task at hand, striking a balance between model performance and computational efficiency as the inclusion of additional hyperparameters in the search space can lead to a significant increase in computational requirements.

2.6. Definition of a standard protection measure

The model $\bar{\mathcal{M}}^*$ for evaluating mitigated scenarios, requires defining the activation time and damping factor of the protection measure, as discussed in Section 2.1. These characteristics are site-specific and depend on the particular fire detection and protection systems implemented. Therefore, there might be instances where they cannot be easily obtained. For these reasons, a third model was developed to predict the TTF of mitigated scenarios, namely $\bar{\bar{\mathcal{M}}}^*$, which considers a reference protection measure with standard activation time and damping factor. The model is intended to provide a rough estimate of the potential effect of mitigation systems based on the performance of a reference active barrier widely used in the current industrial practice. The model was obtained from $\bar{\mathcal{M}}^*$ selecting the activation time and the damping factor of a water deluge system. The activation time was set to 1 minute, which included 30 seconds to detect the fire and 30 seconds to activate water delivery to the nozzles (NORSOK, 2020). The damping factor is set to 0.2, which corresponds to a single row of nozzles, as reported by Lowesmith et al. (2007) and in accordance with the simulations carried out by Wu et al. (2020).

The third model offers a generic reference estimation of the TTF for mitigated tanks, based on the standard performance of water deluges. The model only intends to provide preliminary results when activation times and damping factors of the actual mitigation barriers are unavailable, e.g., when considering installing protection measures in a preliminary design phase. If the actual activation times and damping factors are known, model $\bar{\mathcal{M}}$ must be used since it provides far more accurate data.

2.7. Predictions and confidence intervals

The optimized models $\mathcal{M}^*$ and $\bar{\mathcal{M}}^*$, can predict the TTF of unmitigated and mitigated scenarios. Furthermore, the specialized model $\bar{\bar{\mathcal{M}}}^*$ can be used to simulate the effect of a standard water deluge system. However, in most practical applications, predicting a confidence interval rather than a single precise

value for the TTF is often preferable. This is because confidence intervals provide a range of potential values that captures the uncertainty and variability associated with the prediction. In fact, if a fire scenario differs significantly from those used during training, it is reasonable to expect larger uncertainties associated with the prediction of the TTF. In such cases, it becomes crucial to acknowledge the potential uncertainties and incorporate a larger confidence interval in the prediction.

This study used a model-agnostic method called "jacknife+" to estimate predictive confidence intervals (Barber et al., 2019). The confidence interval produced by jacknife+ can be summarized as follows. Let X_i and y_i respectively indicate the features and the response variable of the i-th observation in a training dataset that comprises n samples. The confidence interval of a new observation X_{n+1} is defined as:

$$\hat{C}_{n,\alpha} = \left[\hat{q}^{-}_{n,\alpha}\left\{\hat{\mu}_{-i}(X_{n+1}) - R_i^{LOO}\right\}, \hat{q}^{+}_{n,\alpha}\left\{\hat{\mu}_{-i}(X_{n+1}) - R_i^{LOO}\right\}\right], \quad (9)$$

where $\hat{C}_{n,\alpha}$ indicates the confidence interval in the form [TTF_{min}, TTF_{max}], $\alpha \in [0,1]$ indicates the uncertainty of the confidence interval, $\hat{q}^{-}_{n,\alpha}\{\blacksquare\}$ and $\hat{q}^{+}_{n,\alpha}\{\blacksquare\}$ respectively represents the α and 1- α quantiles, $\hat{\mu}_{-i}$ is the regression model fitted on all the examples in the training database except the i-th observation, and $R_i^{LOO} = [y_i - \hat{\mu}_{-i}(X_i)]$ is the residual of the i-th observation. A detailed description of the theoretical foundations and implementation of the method can be found in the original reference (Barber et al., 2019) and the Python library description (Taquet et al., 2022). In summary, the jackknife+ method extends the traditional jackknife resampling technique (Efron and Gong, 1983) by addressing algorithm instability, offering robust coverage guarantees without requiring any assumptions apart from having independent and identically distributed samples.

Therefore, incorporating the jacknife+ method, the output of the models $\mathcal{M}^*$, $\bar{\mathcal{M}}^*$, and $\bar{\bar{\mathcal{M}}}^*$ includes not only the predicted TTF but also the corresponding minimum and maximum TTF values, indicating that the true TTF is expected to lie between these two values with a confidence level of 95%.

3. Results

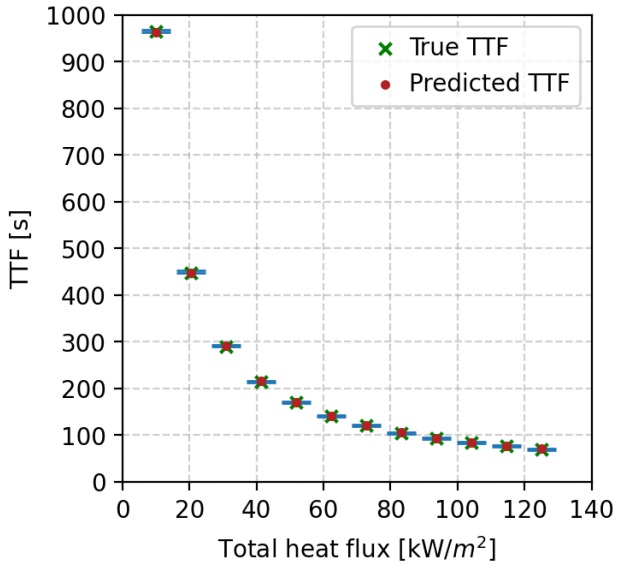
The best model configurations identified by the grid-search procedure (see Section 2.5) and their performance metrics are shown in Table 3. The results indicate that the model for predicting unmitigated scenarios ($\mathcal{M}^*$) achieves lower $RMSE$ and larger R^2 than the model that predicts mitigated scenarios ($\bar{\mathcal{M}}^*$ and $\bar{\bar{\mathcal{M}}}^*$). However, the TTFs of mitigated scenarios are expected to be higher than those of unmitigated scenarios. Therefore, the $RMSE$ was normalized to remove the effect of scale and to allow a fair comparison between the models.

Model	Layers	Neurons	R^2	$RMSE$ [s]	$NRMSE$
$\mathcal{M}^*$ (unmitigated scenarios)	2	{152, 52}	0.9999	1.66	0.006
$\bar{\mathcal{M}}^*$ and $\bar{\bar{\mathcal{M}}}^*$ (mitigated scenarios)	3	{102, 152, 2}	0.9925	71	0.124

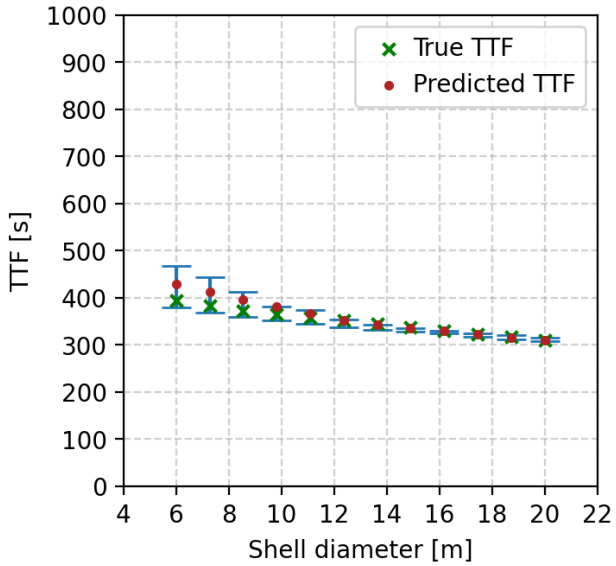
Table 3. Best model configurations for the prediction of mitigated and unmitigated TTFs. R^2 and RMSE are calculated according to Eq. (6) and (7), NRMSE is calculated according to Eq. (8).

Figure 4 and Figure 5 provide illustrative examples of the model output to exemplify the potential use of the unmitigated ($\mathcal{M}^*$) and mitigated ($\bar{\mathcal{M}}^*$, $\bar{\bar{\mathcal{M}}}^*$) models, and demonstrate their accuracy and robustness under different working conditions. A reference tank containing benzene at 20°C and 1 atm, with diameter = 6 m, height = 14.4 m, shell thickness = 0.005 m, filling level = 20 %, and a total heat flux = 50 kW/m^2 is used for the analysis. Each subfigure examines the effect of one or more scenario features (see Table 1) while holding the others constant. It is worth mentioning that the reference fire scenario was removed from the training dataset to ensure the fairness of the results.

In Figure 4, the TTF values obtained using model $\mathcal{M}^*$ for unprotected tanks are reported with respect to thermal radiation (a), shell diameter (b), shell thickness (c), tank height (d), and filling degree (e). The shell thickness of the reference tank was increased to 0.01 m in Figure 4.b in order to ensure physical integrity across the whole range of shell diameters. As shown in the figure, the model predicted values (predicted TTF) show a good agreement with the results of the RADMOD model (true TTF) across the entire feature space. The relatively small confidence intervals indicate low uncertainty in the predictions. The computation time required to perform one prediction is approximately 23.7 milliseconds for both models.



a)



b)

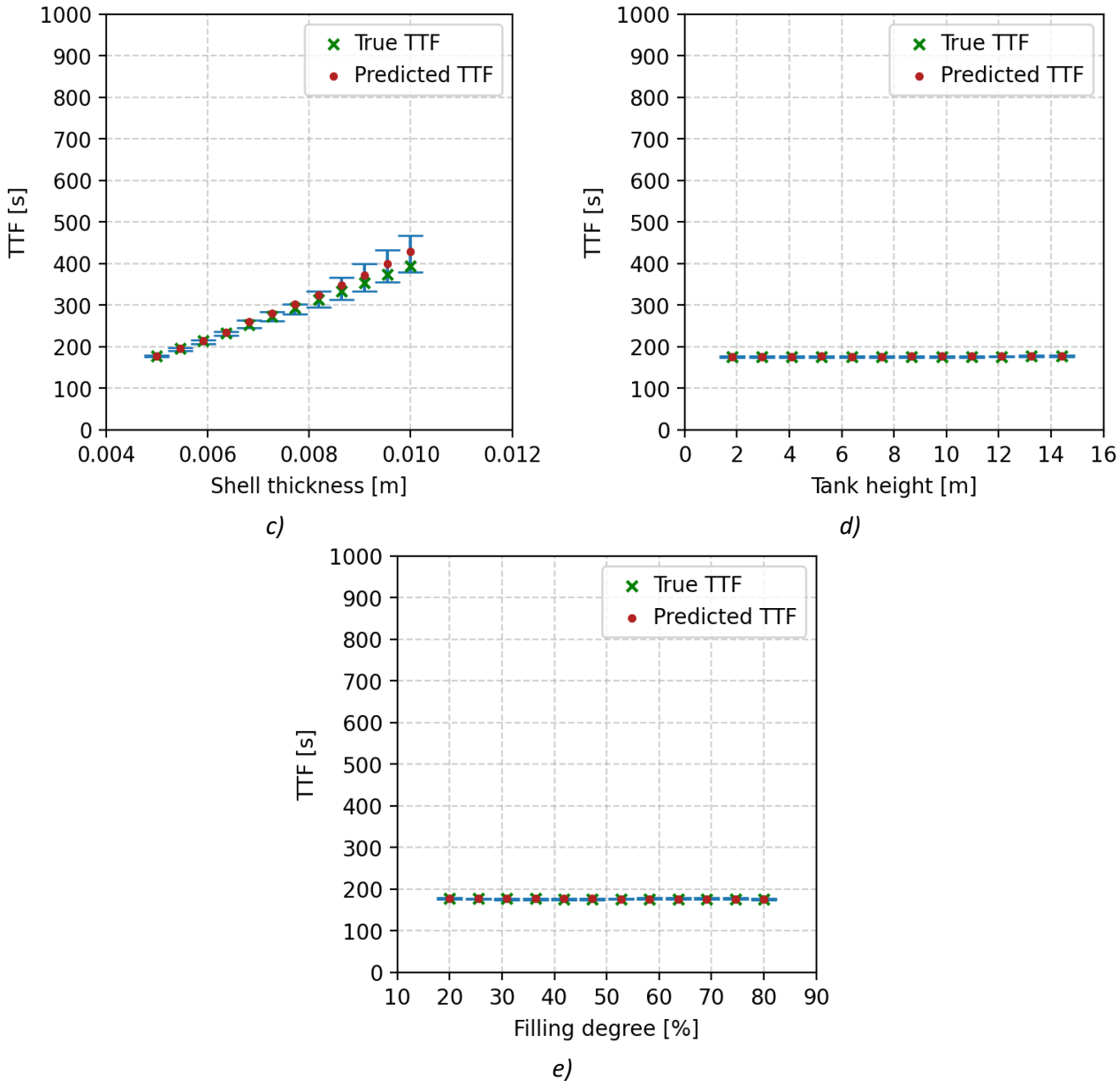


Figure 4. TTF values calculated by model $\mathcal{M}^*$ for an unmitigated tank with diameter = 6 m, height = 14.4 m, shell thickness = 0.005 m, filling level = 20 %, total heat flux = 50 kW/m^2, and containing benzene initially at 20°C and 1 bar. Each plot examines the effect of a specific tank feature while holding the others constant: (a) total heat flux, (b) diameter, (c) thickness, (d) height, and (e) filling degree.

The analysis of mitigated models is shown in Figure 5. Specifically, Figure 5.a illustrates the TTF values versus the damping factor calculated considering an activation time of 40 seconds. Figure 5.b shows the effect of activation time on a safety barrier with damping factor equal to 0.4. Also, in this case, a good agreement with simulated data is present, even if confidence intervals are larger, indicating that the model $\bar{\mathcal{M}}^*$ is associated with higher uncertainties. Figure 5.c compares the TTF values calculated by the model $\mathcal{M}^*$ for an unmitigated tank, and model $\bar{\bar{\mathcal{M}}}^*$ that considers the reference water deluge system introduced in section 2.6. As shown in the figure, the presence and activation of the water deluge results in a relevant increase of TTF values (up to 27 %) with respect to the values calculated for unmitigated tanks. The difference is larger for lower values of heat radiation and rapidly decreases as the heat load increases. The trend can be attributed to the intensified influence of higher heat loads before barrier activation (i.e., during the initial 60 seconds), rapidly

pushing the tank closer to its mechanical limits and thus diminishing the benefits of the water deluge system. The analysis has been further expanded to evaluate the joint effect of the heat radiation and the barrier's activation time. The findings, depicted in Figure 5.c, confirm a positive correlation between the barrier's effectiveness and the difference between its activation time and the unmitigated TTF. Specifically, the efficacy of a barrier increases when its activation time is greater than the unmitigated TTF. As the activation time approaches the TTF of the unmitigated scenario, the difference between mitigated and unmitigated TTFs decreases, eventually reaching zero when the unmitigated TTF equals the barrier's activation time. This observation also underscores a good agreement between the mitigated and unmitigated models, as they produce comparable TTF values when the barrier's activation time coincides with the unmitigated TTF.

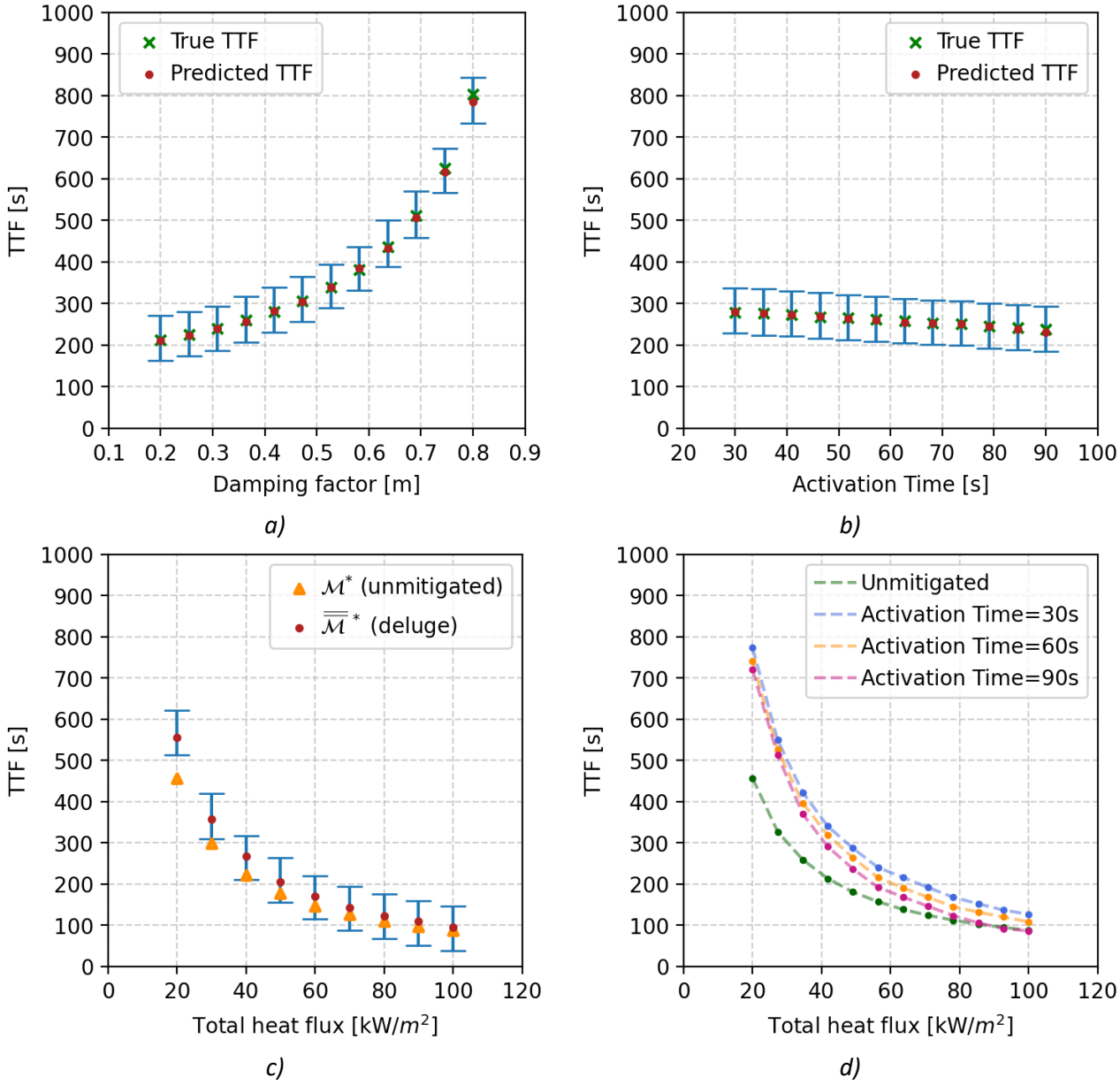
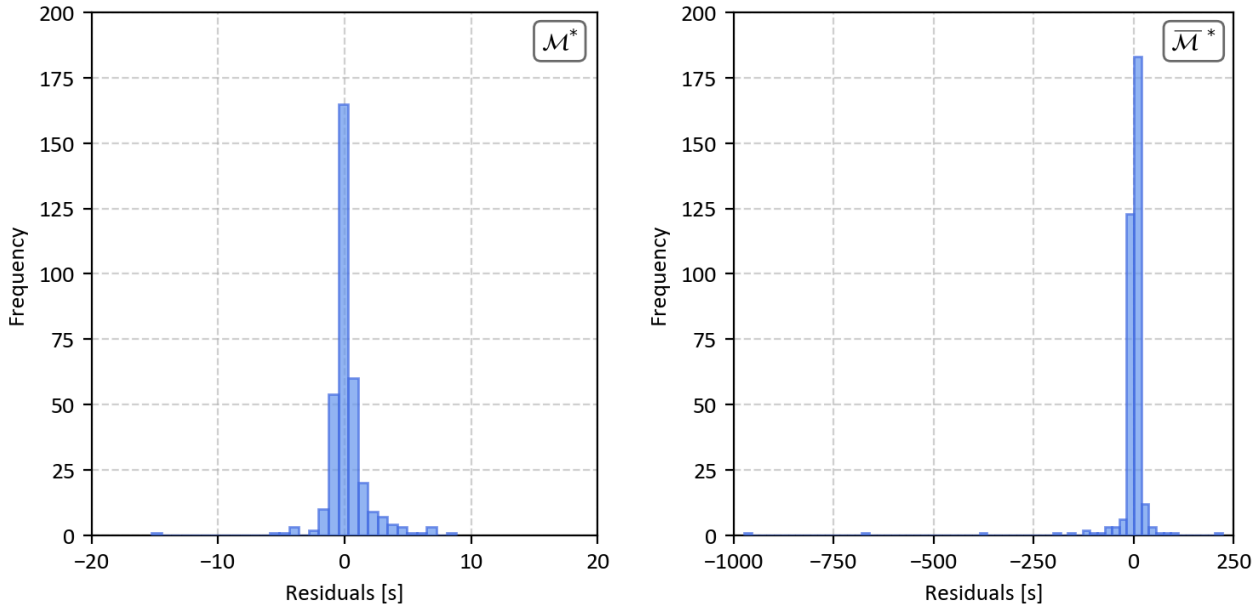


Figure 5. TTF values calculated by model $\overline{\mathcal{M}}^*$ (a, b ,c) and model $\overline{\overline{\mathcal{M}}}^*$ (d) for a mitigated tank with diameter = 6 m, height = 14.4 m, shell thickness = 0.005 m, filling level = 20 %, total heat flux = 50 kW/m2, activation time = 40 s, damping factor = 0.4, and containing benzene initially at 20°C and 1 bar. Figures a and b examine the effect of the damping factor (a) and activation time (b) while holding

the other tank characteristics constant. Figure c illustrates the combined effect of varying the activation time and the total heat flux. Figure d demonstrates the impact of the water deluge system in comparison to the unmitigated tank.

To gain further insights into the performance of the models, the distribution of residuals was calculated for both models $\mathcal{M}^*$ and $\overline{\mathcal{M}}^*$. The results are illustrated in Figure 6.a and Figure 6. b, respectively.



a) *b)*

Figure 6. Residuals of the model $\mathcal{M}^*$(a) and $\overline{\mathcal{M}}^*$(b). Residuals are calculated as the difference between the true TTF and the predicted TTF of fire scenarios included in the evaluation datasets.

Residuals represent the difference between actual and predicted TTFs of the fire scenarios in the evaluation dataset. Figure 6. a shows that most residuals of the model $\mathcal{M}^*$ are smaller than 10 seconds in absolute terms. Only one prediction returned a residual of -15 seconds. Also, the distribution appears centered around 0, with most residuals between -5 and +5 seconds. Similarly, Figure 6.b shows that most residuals of the model $\overline{\mathcal{M}}^*$ are close to 0, but the distribution appears more skewed toward negative values, with three outliers around -973, -673, and -378 seconds.

A comprehensive analysis was conducted to compare the model's performance $\mathcal{M}^*$, with the simplified correlations proposed by Landucci et al. (2009) and Yang et al. (2023). The models were tested based on the ability to predict the TTFs of the unmitigated fire scenarios included in $\mathcal{D}_{eval}$ dataset (see section 2.2). A comparison between the RMSE values obtained by the model $\mathcal{M}^*$ and the correlations by Landucci et al. (2009) and Yang et al. (2023) is shown in Table 4.

Table 4. Comparison between RMSE values obtained by the model presented in this study and state-of-the-art correlations. Only the events include $\mathcal{D}_{eval}$ were considered in the analysis.

Approach	RMSE [s]
The present study	1.66
Landucci et al. (2009)	162.9
Yang et al. (2023)	106.4

The results demonstrate that while Yang et al.'s (2023) method outperforms the approach introduced by Landucci et al. (2009), our model achieves even more remarkable results. Specifically, it reduces the Root Mean Square Error (RMSE) by an order of magnitude compared to these previous methods.

A visual comparison between the three methods is offered in Figure 7, showcasing 'predicted versus actual' plots of our approach compared to Landucci et al. (2009) (Figure 7.a), and Yang et al. (2023) (Figure 7.b)

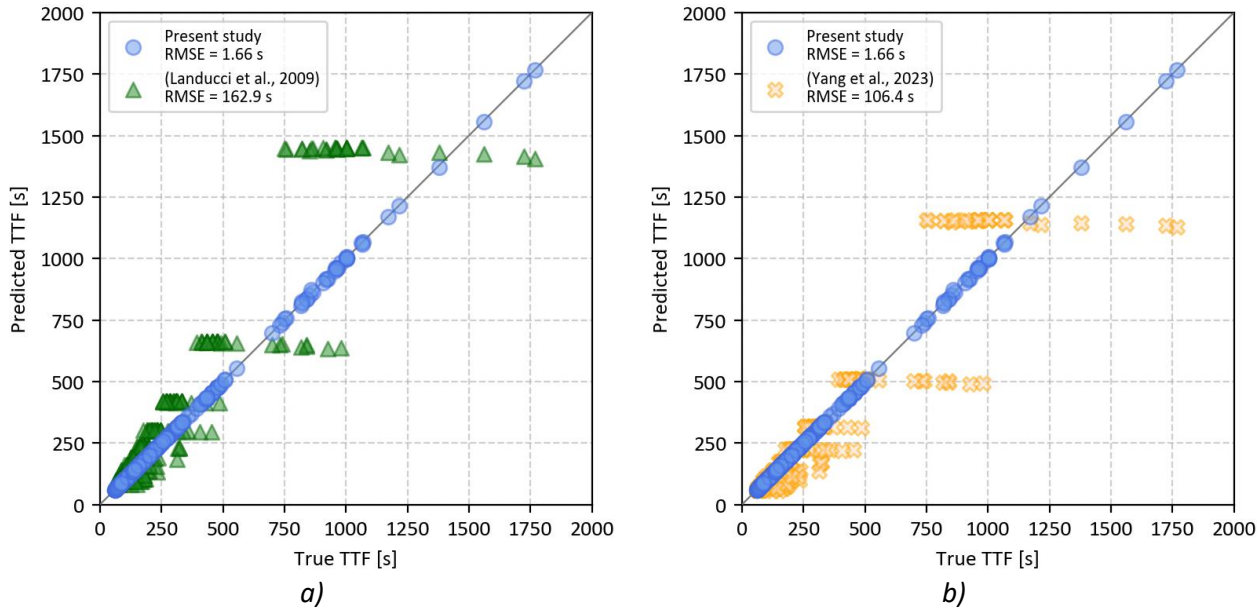


Figure 7. Comparison between the performance of the model $\mathcal{M}^*$ (blue), the simplified correlations proposed by Landucci et al. (2009) *(a) and the correlation proposed by Yang et al.* (2023) *(b). Displayed TTFs refer to unmitigated scenarios included in the $\mathcal{D}_{eval}$ dataset (see Section 2)*

The results confirm that the model proposed in this study aligns remarkably well with the RADMOD data, exhibiting high accuracy in reproducing RADMOD results. On the contrary, the correlations proposed by Landucci et al. (2009) and Yang et al. (2023) tend to produce larger errors, in particular when high values of TTFs are considered.

4. Discussion

The results presented above indicate that the models developed in the present study effectively predict the time-to-failure (TTF) of atmospheric tanks exposed to external fire, using a minimal set of input data and requiring a reduced computational effort. This approach provides safety practitioners and researchers with models that are not only user-friendly and accurate but also highly interpretable. Additionally, the seamless integration of safety barriers within this framework represents a notable advancement in researching risk analysis of fire-induced domino scenarios.

The results indicate that the models show a relatively good performance. However, model $\mathcal{M}^*$ (that considers unprotected tanks) exhibits a higher robustness and accuracy with respect to models $\bar{\mathcal{M}}^*$ and $\bar{\bar{\mathcal{M}}}^*$ (that considers the effect of safety barriers), as confirmed by the smaller confidence intervals in Figure 4 and

the distributions of the residuals shown in Figure 6. The better performance of the model for predicting the unmitigated TTF values can be attributed to the challenge of predicting the effects of safety barriers with varying activation times and damping factors. This complexity demands a more sophisticated approach to accurately capture the impact of safety measures, as evidenced by the need for additional layers and neurons per layer, as shown in Table 3.

The results presented in Figure 4 and Figure 5 reveal key insights into the factors influencing the TTF in the given scenarios. The analysis shows that the total heat flux, along with the shell diameter and thickness, have a great influence on the TTF. Conversely, the height of the tank and its filling degree have a less significant, sometimes negligible, impact. Considering the impact of safety barriers, Figure 5 suggests that the damping factor exerts a more substantial influence than the activation time. It is imperative to stress that a careful and accurate assessment of the activation time, particularly in comparison to the unmitigated TTF, is a key determinant of the barrier effectiveness.

The NN model addressing the simulation of unmitigated scenarios ($\mathcal{M}^*$) outperforms the simplified correlations reported in the literature (Landucci et al., 2009; Jiahao Yang et al., 2023), which are not conceived to capture the dynamics of the fire scenarios. Also, the simplified correlations do not consider the effect of the tank filling degree. In contrast, the proposed models may be easily tailored to the specific case of interest, is accurate, and provides results rapidly, requiring limited computational resources. Thus, the NN models are excellent candidates for dynamic frameworks to assess safety barriers and the risk generated by potential escalations resulting in domino scenarios.

Despite these promising results, it is important to acknowledge some limitations of the NN models developed and some potential areas for future improvements. Firstly, the models were trained using data from a simplified lumped model (i.e., RADMOD), which may, in turn, introduce errors in estimating the TTF values. Incorporating more rigorous TTF data, such as those obtained from large-scale experimental set-ups and/or validated CFD and FEM models, could enhance the performance of the NN models. In this context, the approach shown in the present study demonstrates the potential of coupling first principles modeling and ML for the construction of metamodels that can offer fast and reliable predictions, thereby decreasing the computational burden of techniques requiring many simulations.

Another limitation that needs to be pinpointed is the larger uncertainty associated with the TTF calculated from the model considering mitigated scenarios with respect to those obtained from the model considering unprotected tanks. Future work should focus on exploring a more extensive set of data and hyperparameters, including regularization layers, on enhancing the performance of $\bar{\mathcal{M}}^*$.

In addition, it is crucial to emphasize that the models developed in this study are intended to serve as surrogate tools for calculating the Time to Failure (TTF) and are not meant to replace comprehensive dynamic risk analysis procedures. Fire spread and evolution are intricate phenomena shaped by a multitude of factors, many of which cannot be explicitly integrated into our approach. These elements encompass both

environmental aspects, like atmospheric conditions and the tank's position relative to the fire, and human factors, such as individual behaviors and the timeliness and efficacy of emergency responses. Although our models can account for some of these elements (for instance, by simulating emergency response as a safety barrier with specific activation time and damping factor), they do not inherently address other factors that influence the fire intensity, such as the type of ignited substance, air transmissivity, and view factor, must be considered separately. To this end, existing literature, including van den Bosch and Weterings (2005), provides several correlations that can be used to link environmental aspects to total incident radiation. In other words, our models are not designed to encompass the entire spectrum of fire evolution scenarios; comprehensive frameworks have already been proposed to address the dynamic evolution of such events (Chen et al., 2022; Zeng et al., 2020). Instead, our models are tailored to enhance a specific aspect of these frameworks, focusing on the improved estimation of TTF. Environmental and human factors can be considered upstream, prior to the application of our models, allowing for a more targeted and effective integration into the overall risk assessment process.

Finally, it is worth mentioning that the models proposed in this study are not specifically designed to provide conservative predictions, as indicated by the residuals in Figure 6. The figure shows that the errors are likely to be both negative or positive, meaning that the predicted TTF is not always guaranteed to be smaller than the true TTF. This behavior is at least in part mitigated by the confidence intervals, which offer an estimation of the uncertainty affecting the TTF. The confidence interval allows a better understanding of the robustness of the predictions, allowing more informed judgments based on the level of confidence associated with the prediction.

Despite the abovementioned limitations, this study provides accurate and user-friendly tools, enabling a straightforward evaluation of fire scenarios under mitigated and unmitigated conditions. The toolbox of developed models may thus significantly benefit the assessment of domino scenarios triggered by the fire. The model's capabilities in predicting TTFs for atmospheric tanks and considering the effect of mitigation systems provide new opportunities to enhance the risk assessment and safety evaluation of cascading events leading to domino effects. From a practitioners' viewpoint, the availability of a toolbox of user-friendly models, able to incorporate the effect of mitigation measures without the need of a specific expertise, provides a crucial support to risk-informed decision-making in displaying safety barriers and safety systems aimed to prevent escalation and domino effect.

5. Conclusions

This study is pioneering in proposing using Neural Networks to estimate the TTF of atmospheric tanks exposed to external fires, also considering the effect of mitigative actions. A model toolbox was developed, including NN-based modes allowing the TTF calculation for unprotected tanks and for tanks protected by active and/or passive safety barriers. The models consider the effect of various types of safety measures and

safety systems in terms of activation time and effectiveness in reducing the heat load, thus allowing the simulation of a wide range of safety barriers. The models require only a minimal set of input parameters, thus resulting user-friendly and straightforward to configure. The predictions are fast and accurate, making the models suitable for the dynamic analysis of domino scenarios. The newly developed NN-based models outperform the simplified correlations used in the current practice to estimate TTF, allowing a more accurate calculation of TTF values. Overall, the method demonstrates the potential of coupling digital simulations and ML models to decrease the computational burden, enabling faster and more efficient predictions in complex accident scenarios.

6. References

Abdolhamidzadeh, B., Abbasi, T., Rashtchian, D., Abbasi, S.A., 2011. Domino effect in process-industry accidents – An inventory of past events and identification of some patterns. Journal of Loss Prevention in the Process Industries 24, 575–593. https://doi.org/10.1016/j.jlp.2010.06.013

Abid, A., Khan, M.T., Iqbal, J., 2021. A review on fault detection and diagnosis techniques: basics and beyond. Artificial Intelligence Review 54, 3639–3664. https://doi.org/10.1007/s10462-020-09934-2

American Petroleum Institute, 2021. API 650 - citeWelded Tanks for Oil Storage.

Amin, M.T., Scarponi, G.E., Cozzani, V., Khan, F., 2024a. Dynamic Domino Effect Assessment (D2EA) in tank farms using a machine learning-based approach. Computers & Chemical Engineering 181, 108556. https://doi.org/10.1016/j.compchemeng.2023.108556

Amin, M.T., Scarponi, G.E., Cozzani, V., Khan, F., 2024b. Improved pool fire-initiated domino effect assessment in atmospheric tank farms using structural response. Reliability Engineering & System Safety 242, 109751. https://doi.org/10.1016/j.ress.2023.109751

Arias Chao, M., Kulkarni, C., Goebel, K., Fink, O., 2022. Fusing physics-based and deep learning models for prognostics. Reliability Engineering & System Safety 217, 107961. https://doi.org/10.1016/j.ress.2021.107961

Bai, G., Su, Y., Rahman, M.M., Wang, Z., 2023. Prognostics of Lithium-Ion batteries using knowledge-constrained machine learning and Kalman filtering. Reliability Engineering & System Safety 231, 108944. https://doi.org/10.1016/j.ress.2022.108944

Barber, R.F., Candes, E.J., Ramdas, A., Tibshirani, R.J., 2019. Predictive inference with the jackknife+.

Calzolari, G., Liu, W., 2021. Deep learning to replace, improve, or aid CFD analysis in built environment applications: A review. Building and Environment 206, 108315. https://doi.org/10.1016/j.buildenv.2021.108315

Cao, L., Zhang, H., Meng, Z., Wang, X., 2023. A parallel GRU with dual-stage attention mechanism model integrating uncertainty quantification for probabilistic RUL prediction of wind turbine bearings. Reliability Engineering & System Safety 235, 109197. https://doi.org/10.1016/j.ress.2023.109197

Chen, C., Reniers, G., Yang, M., 2022. Dynamic Risk Assessment of Fire-Induced Domino Effects BT - Integrating Safety and Security Management to Protect Chemical Industrial Areas from Domino Effects, in: Chen, C., Reniers, G., Yang, M. (Eds.), . Springer International Publishing, Cham, pp. 49–68. https://doi.org/10.1007/978-3-030-88911-1_2

Chen, F., Zhang, M., Song, J., Zheng, F., 2018. Risk Analysis on Domino Effect Caused by Pool Fire in Petroliferous Tank Farm. Procedia Engineering 211, 46–54. https://doi.org/10.1016/j.proeng.2017.12.136

Cozzani, V., Gubinelli, G., Salzano, E., 2006. Escalation thresholds in the assessment of domino accidental events. Journal of Hazardous Materials 129, 1–21. https://doi.org/10.1016/j.jhazmat.2005.08.012

Cozzani, V., Reniers, G., 2021. Dynamic Risk Assessment and Management of Domino Effects and Cascading Events in the Process Industry. Elsevier, Amsterdam. ISBN:9780081028384

Cui, X., Zhang, M., Pan, W., 2022. Dynamic probability analysis on accident chain of atmospheric tank farm based on Bayesian network. Process Safety and Environmental Protection 158, 146–158. https://doi.org/10.1016/j.psep.2021.10.040

Efron, B., Gong, G., 1983. A Leisurely Look at the Bootstrap, the Jackknife, and Cross-Validation. The American Statistician 37, 36–48. https://doi.org/10.1080/00031305.1983.10483087

Fishwick, T., 2011. The fire and explosion at Indian Oil Corporation, Jaipur — a summary of events and outcomes. Loss Prevention Bulletin 9–13.

Godoy, L.A., 2016. Buckling of vertical oil storage steel tanks: Review of static buckling studies. Thin-Walled Structures 103, 1–21. https://doi.org/10.1016/j.tws.2016.01.026

Godoy, L.A., Jaca, R.C., Ameijeiras, M.P., 2023. On buckling of oil storage tanks under nearby explosions and fire, in: Above Ground Storage Tank Oil Spills. Elsevier, pp. 199–259. https://doi.org/10.1016/B978-0-323-85728-4.00004-8

Goodfellow, I.J., Bengio, Y., Courville, A., 2016. Deep Learning. MIT Press, Cambridge, MA, USA.

Green, D.W., Perry, R.H., 2008. Perry's Chemical Engineers' Handbook, Eighth Edition, 8th ed. /. ed. McGraw-Hill Education, New York. ISBN:9780071422949

Gubinelli, G., 2005. Models for the assessment of domino accidents in the process industry. University of Pisa.

Hahn, G.J., 2007. The coefficient of determination exposed !

Hastie, T., Tibshirani, R., Friedman, J., 2009. The Elements of Statistical Learning, Springer Series in Statistics. Springer New York, New York, NY. https://doi.org/10.1007/978-0-387-84858-7

Hegde, J., Rokseth, B., 2020. Applications of machine learning methods for engineering risk assessment – A review. Safety Science 122, 104492. https://doi.org/10.1016/j.ssci.2019.09.015

Huang, C., Bu, S., Lee, H.H., Chan, K.W., Yung, W.K.C., 2023. Prognostics and health management for induction machines: a comprehensive review. Journal of Intelligent Manufacturing.

https://doi.org/10.1007/s10845-023-02103-6

Huang, K., Chen, G., Khan, F., Yang, Y., 2021. Dynamic analysis for fire-induced domino effects in chemical process industries. Process Safety and Environmental Protection 148, 686–697. https://doi.org/10.1016/j.psep.2021.01.042

Iannaccone, T., Scarponi, G.E., Landucci, G., Cozzani, V., 2021. Numerical simulation of LNG tanks exposed to fire. Process Safety and Environmental Protection 149, 735–749. https://doi.org/10.1016/j.psep.2021.03.027

Ji, J., Tong, Q., Khan, F., Dadashzadeh, M., Abbassi, R., 2018. Risk-Based Domino Effect Analysis for Fire and Explosion Accidents Considering Uncertainty in Processing Facilities. Industrial & Engineering Chemistry Research 57, 3990–4006. https://doi.org/10.1021/acs.iecr.8b00103

Khakzad, N., Khan, F., Amyotte, P., Cozzani, V., 2014. Risk Management of Domino Effects Considering Dynamic Consequence Analysis. Risk Analysis 34, 1128–1138. https://doi.org/10.1111/risa.12158

Khan, F., Reniers, G., Cozzani, V., 2021. Domino Effect: Its Prediction and Prevention, Methods in chemical process safety. Elsevier, Amsterdam. ISBN:9780323915151

Khan, F.I., Abbasi, S.., 2001. An assessment of the likelihood of occurrence, and the damage potential of domino effect (chain of accidents) in a typical cluster of industries. Journal of Loss Prevention in the Process Industries 14, 283–306. https://doi.org/10.1016/S0950-4230(00)00048-6

Kudela, J., Matousek, R., 2022. Recent advances and applications of surrogate models for finite element method computations: a review. Soft Computing 26, 13709–13733. https://doi.org/10.1007/s00500-022-07362-8

Landucci, G., Argenti, F., Tugnoli, A., Cozzani, V., 2015. Quantitative assessment of safety barrier performance in the prevention of domino scenarios triggered by fire. Reliability Engineering & System Safety 143, 30–43. https://doi.org/10.1016/j.ress.2015.03.023

Landucci, G., Gubinelli, G., Antonioni, G., Cozzani, V., 2009. The assessment of the damage probability of storage tanks in domino events triggered by fire. Accident Analysis & Prevention 41, 1206–1215. https://doi.org/10.1016/j.aap.2008.05.006

Lees, F., 2012a. Fire, in: Lees' Loss Prevention in the Process Industries. Elsevier, pp. 1075–1366. https://doi.org/10.1016/B978-0-12-397189-0.00016-1

Lees, F., 2012b. Storage, in: Lees' Loss Prevention in the Process Industries. Elsevier, pp. 1889–1985. https://doi.org/10.1016/B978-0-12-397189-0.00022-7

Li, J., Mao, W., Yang, B., Meng, Z., Tong, K., Yu, S., 2024. RUL prediction of rolling bearings across working conditions based on multi-scale convolutional parallel memory domain adaptation network. Reliability Engineering & System Safety 243, 109854. https://doi.org/10.1016/j.ress.2023.109854

Li, Q., Wang, Y., Chen, W., Li, L., Hao, H., 2024. Machine learning prediction of BLEVE loading with graph neural networks. Reliability Engineering & System Safety 241, 109639.

https://doi.org/10.1016/j.ress.2023.109639

Li, X., Chen, G., Khan, F., Lai, E., Amyotte, P., 2022. Analysis of structural response of storage tanks subject to synergistic blast and fire loads. Journal of Loss Prevention in the Process Industries 80, 104891. https://doi.org/10.1016/j.jlp.2022.104891

Liashchynskyi, Petro, Liashchynskyi, Pavlo, 2019. Grid Search, Random Search, Genetic Algorithm: A Big Comparison for NAS.

Lowesmith, B.J., Hankinson, G., Acton, M.R., Chamberlain, G., 2007. An Overview of the Nature of Hydrocarbon Jet Fire Hazards in the Oil and Gas Industry and a Simplified Approach to Assessing the Hazards. Process Safety and Environmental Protection 85, 207–220. https://doi.org/10.1205/psep06038

Maidana, R.G., Parhizkar, T., Gomola, A., Utne, I.B., Mosleh, A., 2023. Supervised dynamic probabilistic risk assessment: Review and comparison of methods. Reliability Engineering & System Safety 230, 108889. https://doi.org/10.1016/j.ress.2022.108889

Masum Jujuly, M., Rahman, A., Ahmed, S., Khan, F., 2015. LNG pool fire simulation for domino effect analysis. Reliability Engineering & System Safety 143, 19–29. https://doi.org/10.1016/j.ress.2015.02.010

Naderpour, M., Khakzad, N., 2018. Texas LPG fire: Domino effects triggered by natural hazards. Process Safety and Environmental Protection 116, 354–364. https://doi.org/10.1016/j.psep.2018.03.008

Nassif, A.B., Talib, M.A., Nasir, Q., Dakalbab, F.M., 2021. Machine Learning for Anomaly Detection: A Systematic Review. IEEE Access 9, 78658–78700. https://doi.org/10.1109/ACCESS.2021.3083060

NORSOK, 2020. Technical safety - NORSOK S-001:2020+AC. Norway.

Paltrinieri, N., Khan, F., Cozzani, V., 2015. Coupling of advanced techniques for dynamic risk management. Journal of Risk Research 18, 910–930. https://doi.org/10.1080/13669877.2014.919515

Payette, M., Abdul-Nour, G., 2023. Machine Learning Applications for Reliability Engineering: A Review. Sustainability 15, 6270. https://doi.org/10.3390/su15076270

Reniers, G., Cozzani, V., 2013. Domino effects in the process industries: modelling, prevention and managing. Newnes.

Ricci, F., Misuri, A., Scarponi, G.E., Cozzani, V., Demichela, M., 2024. Vulnerability Assessment of Industrial Sites to Interface Fires and Wildfires. Reliability Engineering & System Safety 243, 109895. https://doi.org/10.1016/j.ress.2023.109895

Roy, A., Chakraborty, S., 2023. Support vector machine in structural reliability analysis: A review. Reliability Engineering & System Safety 233, 109126. https://doi.org/10.1016/j.ress.2023.109126

Scarponi, G., Landucci, G., Birk, A., Cozzani, V., 2019. CFD Study of the Fire Response of Vessels Containing Liquefied Gases. Chemical Engineering Transactions 77, 373-378 SE-Research Articles. https://doi.org/10.3303/CET1977063

Sharma, Sagar, Sharma, Simone, Athaiya, A., 2017. Activation functions in neural networks. Towards Data Sci

6, 310–316.

Su, M., Wei, L., Zhou, S., Yang, G., Wang, R., Duo, Y., Chen, S., Sun, M., Li, J., Kong, X., 2022. Study on Dynamic Probability and Quantitative Risk Calculation Method of Domino Accident in Pool Fire in Chemical Storage Tank Area. International Journal of Environmental Research and Public Health 19, 16483. https://doi.org/10.3390/ijerph192416483

Taquet, V., Blot, V., Morzadec, T., Lacombe, L., Brunel, N., 2022. MAPIE an open-source library for distribution-free uncertainty quantification.

U.S. Chemical Safety and Hazard Investigation Board, 2023. Storage Tank Fire at Intercontinental Terminals Company, LLC (ITC) Terminal -Investigation Report.

van den Bosch, C.J.H., Weterings, R.A.P.M., 2005. Heat flux from fires, in: Methods for the Calculation of Physical Effects Due to Releases of Hazardous Materials (Liquids and Gases).

Villa, V., Paltrinieri, N., Khan, F., Cozzani, V., 2016. Towards dynamic risk analysis: A review of the risk assessment approach and its limitations in the chemical process industry. Safety Science 89, 77–93. https://doi.org/10.1016/j.ssci.2016.06.002

Vipin, Pandey, S.K., Tauseef, S.M., Abbasi, T., Abbasi, S.A., 2018. Pool Fires in Chemical Process Industries: Occurrence, Mechanism, Management. Journal of Failure Analysis and Prevention 18, 1224–1261. https://doi.org/10.1007/s11668-018-0517-2

Wang, H., Zheng, J., Xiang, J., 2023. Online bearing fault diagnosis using numerical simulation models and machine learning classifications. Reliability Engineering & System Safety 234, 109142. https://doi.org/10.1016/j.ress.2023.109142

Wang, M., Wang, J., Yu, X., Zong, R., 2023. Experimental and numerical study of the thermal response of a diesel fuel tank exposed to fire impingement. Applied Thermal Engineering 227, 120334. https://doi.org/10.1016/j.applthermaleng.2023.120334

Wolpert, D.H., Macready, W.G., 1997. No free lunch theorems for optimization. IEEE Transactions on Evolutionary Computation 1, 67–82. https://doi.org/10.1109/4235.585893

Wu, Z., Hou, L., Wu, S., Wu, X., Liu, F., 2020. The time-to-failure assessment of large crude oil storage tank exposed to pool fire. Fire Safety Journal 117, 103192. https://doi.org/10.1016/j.firesaf.2020.103192

Xia, T., Dong, Y., Xiao, L., Du, S., Pan, E., Xi, L., 2018. Recent advances in prognostics and health management for advanced manufacturing paradigms. Reliability Engineering & System Safety 178, 255–268. https://doi.org/10.1016/j.ress.2018.06.021

Xu, Z., Saleh, J.H., 2021. Machine learning for reliability engineering and safety applications: Review of current status and future opportunities. Reliability Engineering & System Safety 211, 107530. https://doi.org/10.1016/j.ress.2021.107530

Yang, Jianfeng, Zhang, B., Chen, L., Diao, X., Hu, Y., Suo, G., Li, R., Wang, Q., Li, J., Zhang, J., Dou, Z., 2023. Improved solid radiation model for thermal response in large crude oil tanks. Energy 284, 128572.

https://doi.org/10.1016/j.energy.2023.128572

Yang, Jiahao, Zhang, M., Zuo, Y., Cui, X., Liang, C., 2023. Improved models of failure time for atmospheric tanks under the coupling effect of multiple pool fires. Journal of Loss Prevention in the Process Industries 81, 104957. https://doi.org/10.1016/j.jlp.2022.104957

Yang, L., Shami, A., 2020. On hyperparameter optimization of machine learning algorithms: Theory and practice. Neurocomputing 415, 295–316. https://doi.org/10.1016/j.neucom.2020.07.061

Yang, R., Khan, F., Neto, E.T., Rusli, R., Ji, J., 2020. Could pool fire alone cause a domino effect? Reliability Engineering & System Safety 202, 106976. https://doi.org/10.1016/j.ress.2020.106976

Yang, Y., Chen, G., Chen, P., 2018. The probability prediction method of domino effect triggered by lightning in chemical tank farm. Process Safety and Environmental Protection 116, 106–114. https://doi.org/10.1016/j.psep.2018.01.019

Ye, Z., Hsu, S.-C., 2022. Predicting real-time deformation of structure in fire using machine learning with CFD and FEM. Automation in Construction 143, 104574. https://doi.org/10.1016/j.autcon.2022.104574

Yu, T., Zhu, H., 2020. Hyper-Parameter Optimization: A Review of Algorithms and Applications.

Zeng, T., Chen, G., Yang, Y., Chen, P., Reniers, G., 2020. Developing an advanced dynamic risk analysis method for fire-related domino effects. Process Safety and Environmental Protection 134, 149–160. https://doi.org/10.1016/j.psep.2019.11.029

Zhang, C., Hu, D., Yang, T., 2024. Research of artificial intelligence operations for wind turbines considering anomaly detection, root cause analysis, and incremental training. Reliability Engineering & System Safety 241, 109634. https://doi.org/10.1016/j.ress.2023.109634

Zhou, J., Reniers, G., 2022. Dynamic analysis of fire induced domino effects to optimize emergency response policies in the chemical and process industry. Journal of Loss Prevention in the Process Industries 79, 104835. https://doi.org/10.1016/j.jlp.2022.104835

Article VI.

Tamascelli, N., Arslan, T., Shah, S.L., Paltrinieri, N., Cozzani, V. (2020). **A Machine Learning Approach to Predict Chattering Alarms**. Chem. Eng. Trans. 82. https://doi.org/10.3303/CET2082032.

CHEMICAL ENGINEERING TRANSACTIONS
VOL. 82, 2020
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Bruno Fabiano, Valerio Cozzani, Genserik Reniers

ISBN 978-88-95608-80-8; **ISSN** 2283-9216
DOI: 10.3303/CET2082032

A Machine Learning Approach to Predict Chattering Alarms

Nicola Tamascelli[a,b,*], Tufan Arslan[c], Sirish L. Shah[d], Nicola Paltrinieri[b], Valerio Cozzani[a]

[a] Department of Civil, Chemical, Environmental and Materials Engineering, University of Bologna, Bologna, Italy
[b] Department of Mechanical and Industrial Engineering, NTNU, Trondheim, Norway
[c] Scientific Computing group, IT Department, NTNU, Trondheim, Norway.
[d] Department of Chemical and Materials Engineering, University of Alberta, Alberta, Canada
nicola.tamascelli@gmail.com

The alarm system plays a vital role to ensure safety and reliability in the process industry. Ideally, an alarm should inform the operator about critical conditions only and provide guidance to a set of corrective actions associated with each alarm. During alarm floods, the operator may be overwhelmed by several alarms in a short time span, and crucial alarms are more likely to be missed during these situations. Most of the alarms triggered during a flood episode are nuisance alarms –i.e. alarms that do not convey any new information to the operator, or alarms that do not require operator actions. Chattering alarms that repeat three or more times in a minute and redundant or duplicated alarms are common forms of nuisance alarms. Identifying such nuisance alarms is a key step to improve the performance of the alarm system. Recently, advanced techniques for alarm management have been developed to quantify alarm chatter; although effective, these techniques produce relatively static results. Machine learning algorithms offer an interesting opportunity to analyse historical alarm data and retrieve knowledge, which can be used to produce more flexible and dynamic models, as well as to predict alarms behaviour. The present study aims to develop a machine learning-based algorithm for chattering prediction during alarm floods. A modified approach based on run lengths distribution has been developed to evaluate the likelihood of future alarm chatter. The method has allowed categorizing historical alarm events as alarms that will (or will not) show chattering in the future. Finally, categorized alarms have been used to train a Deep Neural Network, whose performance has been evaluated against the ability to predict alarm chatter. Overall, the Neural Network has shown good prediction capabilities and most of the chattering alarms were correctly identified.

1. Introduction

The advent of the Distributed Control System (DCS) has undeniably improved flexibility and safety of chemical plants, but some issues have arisen as well. In the analog days, installing new alarms used to cost around 1000 $/alarm (Katzel, 2007), including purchase and hard wiring of each alarm and the corresponding annunciator panels (Shaw, 1993). Nowadays, alarms are managed by the DCS. The cost for installing new alarms has dropped and physical panels are not required anymore (Katzel, 2007). The digitised installation has improved the flexibility of the alarm system but as a drawback, a large number of alarms are now present in most process system (Shaw, 1993). As a consequence, more than often the number of alarms displayed are unmanageable by the operator. Recently, standard manuals such as ANSI/ISA (2016) and EEMUA 191 (2013) have addressed the problem of poor alarm management in modern chemical plants, providing guidelines and suggestions. According to these standards, the average alarm annunciation rate should not exceed 6 alarms/hour per operator console to be considered manageable. Unfortunately, in most chemical plants, the alarms rate is much higher than the suggested value (Kondaveeti *et al.*, 2013).
Alarm floods are “conditions during which the alarm rate is greater than the operator can effectively manage (e.g. more than 10 alarms per 10 minutes)” (ANSI/ISA, 2016). During a flood episode, an operator may have to acknowledge and resolve hundreds of alarms in a short period. Clearly, an effective response is impossible in such a chaotic situation. Typically, a majority of the alarms in a flood episode are *nuisance* alarms (i.e. that

Paper Received: 30 January 2020; Revised: 3 May 2020; Accepted: 20 July 2020
Please cite this article as: Tamascelli N., Arslan T., Shah S., Paltrinieri N., Cozzani V., 2020, A Machine Learning Approach to Predict Chattering Alarms, Chemical Engineering Transactions, 82, 187-192 DOI:10.3303/CET2082032

do not communicate any new information) (ANSI/ISA, 2016). Several types of nuisance alarms exist (e.g. chattering, fleeting and stale alarms). Chattering alarms are alarms "that repeatedly transitions between active state and inactive state in a short period of time" (ANSI/ISA, 2016). Therefore, chattering alarms have the potential to produce a large count of alarms and reducing their number is a key step to improve the performance of the alarm system during alarm floods. Kondaveeti *et al.* (2013) proposed a method for quantifying alarm chatter based on run lengths distributions. Although effective, this technique produces static results (i.e. chattering is quantified based on historical alarm data, but no conclusion can be drawn about the alarm future behaviour). In a modern context, where computer technologies and Industry 4.0 solutions are rapidly expanding among different sectors, the need for more dynamic and flexible models is real. In the current scenario, chemical plants produce and store an immense amount of data (Balasko and Abonyi, 2007), modern computers have outstanding calculation capability, and data science techniques have come a long way. We now have the technical capability and the tools to process a vast amount of data. However, process data is mainly archived and not analysed or explored to mine for information and knowledge. The availability of multivariate statistical and Machine Learning techniques now offers the opportunity to "learn" and extract knowledge from past data (Liu *et al.*, 2018).
For the reasons mentioned above, the objective of this study is to overcome the limitations of the existing methods for chattering quantification and to propose a Machine Learning based method for chattering prediction. Specifically, the Chattering Index approach proposed by (Kondaveeti *et al.*, 2013) has been modified to obtain a Dynamic Chattering Index, whose results are then used to train a Deep Neural Network model. The efficacy of the proposed method is evaluated by application to an industrial case data set consisting of alarm data from an ammonia production plant.

2. Alarms from ammonia production plant

An industrial alarm database has been considered to support the analyses. Specifically, alarm data from a section of an ammonia production process (Topsoe.com, 2020) is analysed. Due to the large quantity of hazardous substances stored and handled during normal activity, the plant has been classified as an "upper tier" Seveso III establishment. Extensive use of methane, hydrogen, and ammonia (anhydrous and aqueous solution) occurs in the plant section. Furthermore, due to the intrinsic properties of the processes involved, severe operating conditions (i.e. high pressure and high temperature) are often associated with corrosive substances. Additional information about ammonia production and the considered site can be found at: (Aika *et al.*, 2012; Yara Italia S.p.A, 2016).

The alarm database consists of alarm data collected during an observation period of more than four months. Each row of the database represents an alarm event (26,473 observations in total), and each column (thirty-six in total) represents a piece of information about the alarm (i.e. an "attribute"). A list of the most meaningful attributes is presented in Table 1.

Table 1 - Alarm database attributes

Attribute	Meaning
Time Stamp	Date and time (GMT) of the alarm event.
Source	The source that triggered the alarm. It might be a measuring instrument or a PLC function.
Jxxx	The safety interlock logic associated with the alarm.
Message	The message that is shown to the operator contains the following five attributes: 1. the Source; 2. a concise description of the equipment involved; 3. the safety interlock logic (Jxxx); 4. the value and units of measures of the process variable; 5. the Alarm Identifier (e.g. HHH, HTRP, LLL, LTRP, ACK, etc.)
Active Time	Date and time (GMT) of the first alarm occurrence.
Data Value	The value of the process variable.
Eng. Unit	The units of measure of the process variable.

The *Alarm Identifier* (point 5. of the "Message" attribute) is a code that defines the alarm status. Examples of *Alarm Identifiers* are "HHH" (which means that the measured variable has exceeded the "high level" setpoint), "HTRP" (the measured variable has exceeded the "very high level" alarm setpoint and automatic block intervention procedures might be triggered), "IOP" (which indicates an instrumental failure or out-of-range measure), "LLL" and "LTRP" (same as "HHH" and "HTRP" but referring to a "low/very low level").

According to Kondaveeti *et al.* (2010), an alarm event is uniquely identified by three attributes only: *Time Stamp, Source, and Alarm Identifier*.
The combination of a "*source*" and an "*alarm identifier*" is called a "*unique alarm*". The time-distribution of the alarms has been assessed and represented in Figure 1.

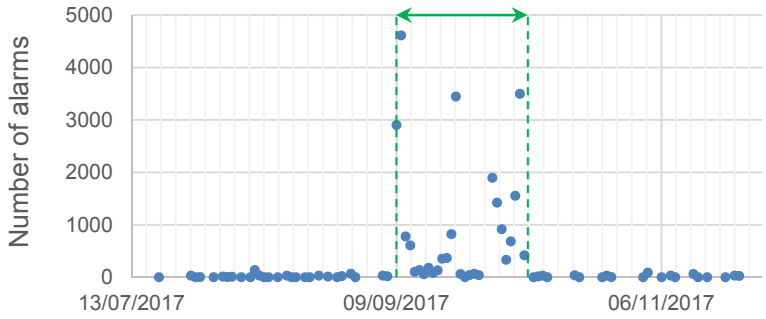


Figure 1 – Alarms time distribution

More than 96 % of the alarms registered in the database occurred within one month only (green line 'window' in Figure 1) when a considerable number of floods and chattering alarms must have occurred. In fact, only ten alarm sources (out of 194 in total) were responsible for more than 80 % of the alarms recorded.

3. Method

This section aims to describe the approach to define the Dynamic Chattering Index. Information about Deep Learning and the related simulations is provided in the sub-section that follows.

3.1 The Dynamic Chattering Index

Using the alarm database as a source of data, all the *Unique alarms* (e.g. FI209B IOP, LI318 LTRP, etc.) are identified, and alarm data are represented as binary sequences (Kondaveeti *et al.*, 2010). Given a generic *unique alarm* that raised *n* times during the observation period, each alarm event (i.e. 1 in the binary sequence) can be identified by an index *i* in such a way that the first occurrence has *i* = 1, the second has *i* = 2, …, the last one has *i* = *n.* The Dynamic Chattering Index related to a generic alarm event with index *i* can be obtained through the following steps:

1. All the alarm events occurred before the event *i* are removed from the binary sequence. The same is done to the events that occurred more than one hour after the event *i*. Data that have not been removed are stored in a new binary sequence, which contains the alarm event *i* and all the alarm events happened within one hour. For example, if the *unique alarm* event *i* occurred at 10:00:00, the reduced binary sequence will contain events that happened between 10:00:00 and 11:00:00.
2. Based on the reduced binary sequence identified during step 1, the *run-lengths* (i.e. the "time difference in seconds between two consecutive alarms on the same tag" (Kondaveeti *et al.*, 2013)) are calculated. Therefore, if the unique alarm occurs *n* times within one hour (i.e. the reduced binary sequence contains *n* 1's), and if the binary sequence does not contain the last alarm recorded during the observation period, *n run-lengths* are calculated. A run length is represented by the letter *r*.
3. The alarm count (i.e. the number of alarms with *run-length* equal to *r*) is obtained. The alarm count is represented by the symbol n_r.
4. The probability (P_r) of an alarm having a *run-length* equal to *r* is calculated:

$$P_r = \frac{n_r}{\sum_{r \epsilon N} n_r} \qquad \forall\, r \epsilon N \tag{1}$$

   One value of P_r is calculated for each unique *run-length* (e.g. P_2 for *r* = 2 s, P_3 for *r* = 3 s, etc.).
5. Finally, The Dynamic Chattering Index related to the alarm event *i* is calculated:

$$\psi_D = \sum_{r \epsilon N} P_r \frac{1}{r} \qquad \forall\, r \epsilon N \tag{2}$$

6. The steps above are repeated $\forall\, i \in [1, n - 1]$.

Through the steps above, each of the first *n - 1* occurrences of the *unique alarm* of concern is associated with a Dynamic Chattering Index (the last occurrence is excluded from the calculation). Then, the procedure is repeated for each unique alarm. The Dynamic Chattering Index assumes values between 0 and 1. The larger the index (i.e. the closer to 1), the higher the alarm chatter within one hour. According to Kondaveeti *et al.* (2013), an index value equal to 0.05 has been used as a threshold to categorise alarms into "Chattering" and "Not Chattering"; if an alarm event has $\psi_D \geq 0.05$, the alarm will show chattering in an hour.

3.2 Machine Learning simulations

A Deep Neural Network (DNN) has been trained and evaluated against the ability to predict alarm chatter. Specifically, the purpose of the algorithm is to classify alarms into two categories: "Chattering within one hour" or "Not Chattering within one hour". A database has been created containing both *features* (i.e. meaningful attributes of an alarm event) and *labels* (i.e. values or categories that the model must predict). Each row of the database represents an alarm event. The first thirteen columns represent an attribute of the alarm (i.e. a *feature*), the fourteenth column contains the *labels* associated with each alarm event. A *label* can be either "1" if the alarm will show chattering within one hour (i.e. $\psi_D \geq 0.05$) or 0 if the alarm will not show chattering within one hour (i.e. $\psi_D < 0.05$). The *features* are presented in Table 2.

Table 2 – Alarm's features

Attribute	Meaning
Y, M, d, H, m, S	Year, Month, Day, Hour, …, Second of the alarm event
SO	The alarm Source
ID	The alarm Identifier
CN	The alarm Condition Name (i.e. the alarm *identifier* of the original alarm from the same Source)
JX	The safety interlock logic associated with the alarm
ATD	Time between the alarm event and its recovery
VAL	The value of the process variable
UNI	The units of measure of the process variable

Next, the database has been shuffled (i.e. rows have been randomly rearranged to improve data distribution) and divided in two, to obtain two distinct databases: the first database (i.e. the *training database*) comprises ¾ of the original database, the remaining part constitutes the second database (i.e. the *evaluation database*). Finally, the *labels* have been removed from the evaluation database.
The databases have been used to *train* and *evaluate* the Deep Neural Network, whose generic architecture is shown in Figure 2.

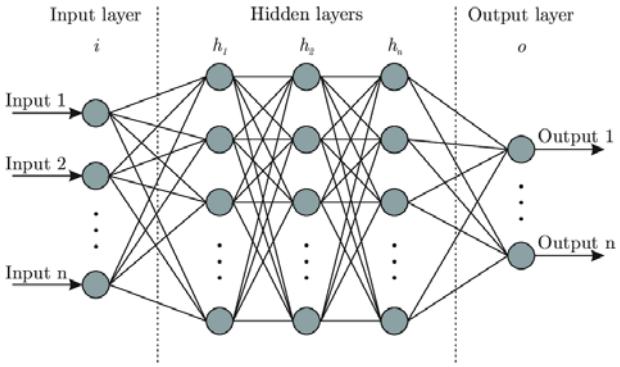


Figure 2 - Artificial neural network architecture (Bre *et al.*, 2018)

During the *training* phase, the algorithm receives as an input both the *features* (Input in Figure 2) and the associated *labels* (Output in Figure 2). During the process, the features are linearly combined and converted through non-linear functions (i.e. activation functions) into *derived features* (i.e. *hidden units*; h_1, h_2, h_n in Figure 2), which constitute the *hidden layer* of the Neural Network (Hastie *et al.*, 2009). ReLU rectifier has been used as an activation function in this work. The weights of the functions are optimised to best represent the relationship between *features* and *labels* (Hastie *et al.*, 2009). Adagrad optimiser has been used for this purpose.

The Deep Neural Network used in this work has three hidden layers with 1024, 512 and 256 hidden units, respectively. After the training, the algorithm is evaluated against the ability to predict the labels of the data included in the evaluation database (i.e. to predict the labels of alarm events that the algorithm has never "seen" before). The Machine Learning algorithm has been developed using TensorFlow r1.15.

4. Results

An example of the results obtained through the Dynamic Chattering Index approach is displayed in Table 3.

Table 3 – Dynamic Chattering Indices for FI227A LLL (Reduced version)

Time Stamps	FI227A LLL	ψ_D(FI227A LLL)
…	…	…
2017-09-09 16:18:09	1	0.072
2017-09-09 16:18:11	1	0.071
2017-09-09 16:24:01	1	0.051
2017-09-09 16:24:03	1	0.018
2017-09-09 16:24:47	1	0.012
…	…	…

Specifically, the table includes a small portion of the Dynamic Chattering Indices related to the unique alarm FI227A LLL. The alarm warns that the flow indicator FI227A has measured a value lower than the "low level" setpoint. The first two columns of the table are the binary representation of the unique alarm (zeroes have been removed from the binary sequence for visualisation purposes). The last column of the table contains the Dynamic Chattering Indices associated with each of the alarm events. The first three indices (marked in red) indicate that the alarm will show chattering behaviour within one hour after the alarm occurrence.

The results of the Machine Learning simulation are shown in the Confusion Matrix displayed in Figure 3.

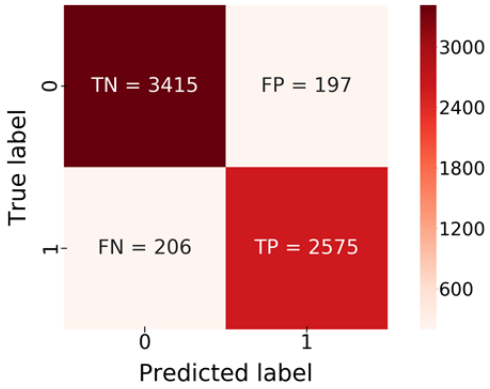


Figure 3 – DNN simulation Confusion Matrix

The metrics "TN" (i.e. True Negative) and "TP" (i.e. True Positive) together represent the number of correct predictions. "FP" (i.e. False Positive) and "FN" (i.e. False Negative) represent the number of wrong predictions. The total number of predictions can be obtained by summing all the metrics discussed above. Therefore, the algorithm produced 6393 predictions (i.e. number of alarm events in the evaluation database); 5990 of them were correct while 403 were incorrect. Besides, three additional metrics have been calculated:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \mathbf{0.937} \quad (3)$$

$$\text{Precision} = \frac{TP}{TP+FP} = \mathbf{0.929} \quad (4)$$

$$\text{Recall} = \frac{TP}{TP+FN} = \mathbf{0.926} \quad (5)$$

The *Accuracy* is the ratio between the correct predictions and the total number of predictions. The *Precision* is the fraction of correct positive predictions (i.e. predicted label = 1 and true label = 1). The *Recall* is the fraction of real positive correctly predicted. *Accuracy*, *Precision* and *Recall* are bounded between 0 and 1; the closer to 1, the better the algorithm performance.

5. Discussion

5.1 Dynamic Chattering Index

The Dynamic Chattering Index evaluates the likelihood of alarm chatter within a defined time interval (e.g. 1 hour). The method produces coherent results in most applications, but it may behave unexpectedly when few alarms occur within the time interval. Specifically, the index is sensitive to the combination of high probability and short run-lengths, a situation that may arise when few alarms occur in fast sequence within the time interval (Tamascelli, 2020). In these situations, just a couple of alarms with run-length less than 5 s could be enough to produce an index greater than 0.05 (i.e. chattering). Therefore, future research will be devoted to the development of a more reliable method for the dynamic quantification of alarm chatter.

5.2 Machine Learning simulations

The DNN model reveals excellent prediction capability. More than 93 % of the total predictions were correct, and more than 92 % of the chattering alarms were correctly identified. Despite the remarkable performance, the Deep Neural Network has not been optimised. For instance, future research will certainly investigate whether the use of a different set of *features*, as well as a different optimiser or a different set of hyperparameters (e.g. the number of hidden units), may lead to better results. As a long-term objective, future research will be devoted to the development of a method to integrate the Machine Learning model on a real industrial alarm system.

6. Conclusions

A method for Dynamic chattering assessment has been developed and the results have been used to train and evaluate a Deep Neural Network. The model has been tested against the ability to predict alarm chatter. Good results have been obtained using a “standard” model (i.e. not optimized). As previously argued, Poor alarm rationalization, chattering and alarm floods are common issues in chemical plants. In this context, Machine Learning models may meet the need for flexible, dynamic and Industry 4.0 oriented tools. Currently, chattering alarms are only addressed retrospectively; existing techniques can identify past alarm chatter but cannot predict future chattering based on actual plant conditions. Instead, the Machine Learning approach described in this work suggests that past alarm data can be used to extract knowledge and to predict alarms behaviour. These advanced models might be valuable tools in supporting the operator response during critical events.

References

Aika K., Christiansen L. J., Dybkjaer I., Hansen J. B., Nielsen P. E. H., Nielsen A., Stoltze P., Tamaru K. , 2012, Ammonia: catalysis and manufacture. Springer Science & Business Media.

ANSI/ISA , 2016, ‘ANSI/ISA–18.2–2016 Management of Alarm Systems for the Process Industries’, ANSI/ISA.

Balasko B., Abonyi J. , 2007, ‘What Happens to Process Data in Chemical Industry? From Source to Applications – An Overview’, Hungarian Journal of Industrial Chemistry, 35, pp. 75–84. doi: 10.1515/133.

Bre F., Gimenez J. M., Fachinotti V. D. , 2018, ‘Prediction of wind pressure coefficients on building surfaces using artificial neural networks’, Energy and Buildings, 158, pp. 1429–1441. doi: 10.1016/j.enbuild.2017.11.045.

EEMUA , 2013, ‘EEMUA Publication 191 Alarm systems - a guide to design, management and procurement’.

Hastie T., Friedman R., Tibshirani J. , 2009, The Elements of Statistical Learning. Springer-Verlag New York. doi: 10.1007/978-0-387-84858-7.

Katzel J. , 2007, Control Engineering | Managing Alarms. Available at: www.controleng.com/articles/managing-alarms (Accessed: 23 January 2020).

Kondaveeti S. R., Izadi I., Shah S. L., Black T. , 2010, ‘Graphical representation of industrial alarm data’, IFAC Proceedings Volumes. IFAC, 11(PART 1), pp. 181–186. doi: 10.3182/20100831-4-fr-2021.00033.

Kondaveeti S. R., Izadi I., Shah S. L., Shook D. S., Kadali R., Chen T. , 2013, ‘Quantification of alarm chatter based on run length distributions’, Chemical Engineering Research and Design. Institution of Chemical Engineers, 91(12), pp. 2550–2558. doi: 10.1016/j.cherd.2013.02.028.

Liu J., Kong X., Xia F., Bai X., Wang L., Qing Q., Lee I. , 2018, ‘Artificial intelligence in the 21st century’, IEEE Access, 6(April), pp. 34403–34421. doi: 10.1109/ACCESS.2018.2819688.

Shaw J. A. , 1993, ‘DCS-based alarms: Integrating traditional functions into modern technology’, ISA Transactions, 32(2), pp. 177–181. doi: 10.1016/0019-0578(93)90039-Y.

Tamascelli N. , 2020, A Machine Learning Approach to Predict Chattering Alarms. University of Bologna - NTNU.

Topsoe.com , 2020, Ammonia. Available at: www.topsoe.com/processes/ammonia (Accessed: 4 April 2020).

Yara Italia S.p.A , 2016, Relazione di riferimento della Yara Italia S.p.A. dello stabilimento di Ferrara. Available at: va.minambiente.it/it-IT/Oggetti/Documentazione/1905/10478.

Article VII.

Tamascelli, N., Paltrinieri, N., & Cozzani, V. (2020). **Predicting Chattering Alarms: A Machine Learning Approach**. Computers & Chemical Engineering, 143, 107122. https://doi.org/10.1016/j.compchemeng.2020.107122.

Computers and Chemical Engineering 143 (2020) 107122

Contents lists available at ScienceDirect

Computers and Chemical Engineering

journal homepage: www.elsevier.com/locate/compchemeng

Predicting chattering alarms: A machine Learning approach

Nicola Tamascelli [a,b], Nicola Paltrinieri [b,*], Valerio Cozzani [a]

[a] Department of Civil, Chemical, Environmental and Materials Engineering, University of Bologna, Bologna, Italy
[b] Department of Mechanical and Industrial Engineering, NTNU, Trondheim, Norway

ARTICLE INFO

Article history:
Received 29 July 2020
Revised 30 September 2020
Accepted 5 October 2020
Available online 6 October 2020

Keywords:
Machine Learning
Data Mining
Alarm management
Alarm floods
Chattering alarms
Chattering prediction

ABSTRACT

Alarm floods represent a widespread issue for modern chemical plants. During these conditions, the number of alarms may be unmanageable, and the operator may miss safety-critical alarms. Chattering alarms, which repeatedly change between the active and non-active states, are responsible for most of the alarm records within a flood episode. Typically, chattering alarms are only addressed and removed retrospectively (e.g. during periodic performance assessments). This study proposes a Machine-Learning based approach for alarm chattering prediction. Specifically, a method for dynamic chattering quantification has been developed, whose results have been used to train three different Machine Learning models – Linear, Deep, and Wide&Deep models. The algorithms have been employed to predict future chattering behavior based on actual plant conditions. Performance metrics have been calculated to assess the correctness of predictions and to compare the performance of the three models.

1. Introduction

The digital revolution and the advent of Distributed Control Systems have undeniably improved the flexibility of industrial alarm systems (Shaw, 1993). Installing new alarms has become relatively simple and economical (Katzel, 2007), but the misconception that more alarms would improve safety and reliability persists in some cases. On the contrary, too many alarms can negatively affect the performance of the alarm system and prevent an adequate operator's response (Kondaveeti et al., 2013; Laberge et al., 2014). Unsatisfactory alarm rationalization is expressed by episodes where an excessive number of alarms are triggered in a short period (ANSI/ISA, 2016; EEMUA, 2013; Laberge et al., 2014). A specific term is coined to define a period of intense alarm activity – an "alarm flood" (Beebe et al., 2012). Hundreds or even thousands of alarms may be triggered during a flood episode, causing a substantial distraction to the operators, and increasing the risk of missing critical alarms (Laberge et al., 2014).

Several studies, accident reports, and standard manuals have cited alarm floods as a contributing factor to financial loss, injuries, and deaths in the chemical industry (Beebe et al., 2012; EEMUA, 2013; Stanton and Barber, 1995), including the investigation report on the explosion in Pembroke Refinery on the 24 July 1994 (Health and Safety Executive, 1997). The accident was caused by a faulty control valve, which was stuck in a closed position. Unfortunately, the control system erroneously indicated that the valve was open, and operators had not been able to identify the problem. Due to the blocked valve, the liquid had accumulated inside a debutanizer column, causing the pressure to increase over the PSV setpoint. A liquid-vapor stream entered the flare pipe that eventually broke since it was not designed to handle liquids. Roughly 10 to 20 tons of partially vaporized flammable materials were released and mixed with air, forming a flammable cloud that ignited and exploded 4 hours after the valve failure, and 20 seconds after the pipe rupture. As a consequence, 26 workers were injured, and the refinery was severely damaged: £48 million were spent on rebuilding the damaged plant, to which the costs of prolonged business interruption should be added. During the accident, alarms were notified to the operators at the rate of one every two to three seconds. Approximately 275 alarms were triggered in the last eleven minutes before the accident, without a concrete effect on the possibility of preventing the accident.

Most of the alarm events within a flood episode are produced by alarms that oscillate between the active and not active state with high frequency –i.e., chattering alarms. Standard manuals have been published (ANSI/ISA, 2016; EEMUA, 2013), providing guidelines for proper alarm rationalization and management, suggesting strategies for chattering and floods reduction. Still, chattering alarms are only addressed and removed retrospectively. Rather than addressing the problem after chattering has happened, a method to predict future chattering based on actual process conditions would significantly improve the performance of the alarm

* Corresponding author.
E-mail address: nicola.paltrinieri@ntnu.no (N. Paltrinieri).

https://doi.org/10.1016/j.compchemeng.2020.107122

Definitions

Accuracy	the ratio of number of correct predictions to total number of predictions.
Alarm flood	a condition during which the alarm rate is greater than the operator can effectively manage (e.g., more than 10 alarms per 10 minutes).
Alarm identifier	a code defining the alarm status.
Alarm source	a field device, control system, or Human Machine Interface that can trigger a change in the alarm status.
Binary Database	a database where unique alarms data are represented as sequences of 0's and 1's at one-second sampling.
Chattering alarm	an alarm that repeatedly transitions between the active and the not active states in a short period (e.g., 3 or more alarm records in one minute).
Chattering Index	an index to quantify the amount of chattering that a unique alarm has shown over a certain time period.
Dynamic Chattering Index	an index to quantify the amount of chattering that a unique alarm has shown up to one hour after each alarm event.
Example	the description of an alarm event in terms of features and related label.
Feature	a meaningful attribute of an alarm event.
Label	the category of an alarm event – "Y" for "Chattering within one hour", "N" for "Not Chattering within one hour".
Nuisance alarm	an alarm that annunciates excessively, unnecessarily, or does not return to normal after the operator action is taken.
Precision	the fraction of positively predicted labels that are, in fact, positive.
Probability Threshold	an adjustable parameter used to convert raw predicted probabilities into predicted labels.
Recall	the fraction of real positive labels correctly predicted.
Run Length	the time difference in seconds between two consecutive alarm events from the same unique alarm.
Unique alarm	the unique combination of an alarm source and an identifier.
Unlabeled examples	the description of an alarm event in terms of a list of features.

system. Nevertheless, predictive methods based on first principles would be complicated to obtain because many variables influence the dynamics of the system (Ahmed et al., 2013). In this multivariate context, a statistical data-based approach appears to be more feasible. Chemical plants produce a large quantity of process and alarm data on a daily basis (Reis and Kenett, 2018). Thus, the use of Machine Learning techniques appears to be an interesting opportunity to extract knowledge from these data and to build predictive models. Various researches have focused on the development of Machine Learning algorithms for fault detection and diagnosis (Mahadevan and Shah, 2009; Miao et al., 2013; Zhong et al., 2014), risk assessment (Paltrinieri et al., 2019), process simulation (Aleixandre et al., 2015; Zhang et al., 2010), and dimensionality reduction (Ge et al., 2017). However, to the best of our knowledge, there is not a direct application of these algorithms for alarm chatter prediction.

The present study proposes a Machine Learning approach for chattering prediction. An industrial alarm database has been used to support the analysis. Initially, a modified version of the Chattering Index proposed by Kondaveeti et al. (2013) has been developed and used to classify historical alarm events as "Chattering within an hour" or "Not Chattering within an hour" (i.e., alarms that will/will not show chattering within one hour after an alarm event). The results of this method, named Dynamic Chattering Index, have been used to train and evaluate three different Machine Learning classification models –i.e., Linear, Deep, and Wide&Deep models. Each algorithm has been trained and assessed independently on the same dataset. Performance metrics have been calculated to assess the correctness of predictions and to compare the performance of the three models.

The paper is organized in 8 Sections. Section 2 provides an overview of industrial alarms and alarm databases, including definitions of nuisance alarms, chattering, and alarm floods. Section 3 focuses on the database used in this work; a brief description of the plant section that generated the alarms is also provided. Section 4 describes the methodology, which includes the preprocessing of alarm data, the development of the Dynamic Chattering Index, the Machine Learning models, and the performance metrics that have been used to evaluate the models. Section 5 provides a detailed description of the Machine Learning simulations. The results of the simulations are presented in Section 6 and discussed in Section 7. Finally, conclusions are summarized in Section 8.

2. Alarms in the chemical industry

Disturbances of various nature cause inherent process fluctuation during daily operations. Typically, minor deviations are managed by the Basic Process Control System (BPCS), and process oscillations are maintained to an acceptable level. However, situations may arise where automatic systems fail to restore normal operations, and human intervention is needed. In these circumstances, alarms inform the operator that process conditions are significantly deviating from their normal operating state (ANSI/ISA, 2016). Each alarm should support a timely and effective response by providing guidance to a set of corrective actions.

2.1. Nuisance, chattering and alarm floods

If an alarm does not convey any new information, or if no corrective action is possible, the alarm is ineffective. These types of alarms are called "nuisance" and are often caused by poorly managed alarm systems (ANSI/ISA, 2016; Kondaveeti et al., 2010). Different types of nuisance alarms can be identified (e.g., Chattering, Fleeting, Stale alarms)(ANSI/ISA, 2016), but in this study the attention has been directed to chattering alarms – i.e., alarms that "repeatedly transitions between the active state and the not active state in a short period of time" (ANSI/ISA, 2016). A rule of thumb

Table 1
Selection of the most common and significant alarm attributes presented to the operator.

Attribute	Description
Timestamp	Date and time (GMT) of the alarm event.
Source	The source that triggered the alarm. It might be a measuring instrument or a PLC function.
Jxxx	The safety interlock logic associated with the alarm, where "xxx" is a three digits code.
Alarm Identifier	A code that defines the alarm status (e.g. "HHH", "HTRP", "LLL", "IOP", "HHH Recover", "ACK").
Data Value	The value of the process variable.
Eng. Unit	The units of measure of the process variable (e.g. " % ", "°C ", " KPa ").

to determine chattering behavior is three or more alarm records (from the same alarm source) in one minute (Kondaveeti et al., 2010).

Besides, alarm floods –i.e., periods when the alarm rate exceeds 10 alarms/operator per ten minutes time interval– are another common issue in modern alarm systems (ANSI/ISA, 2016; Laberge et al., 2014). Due to the intense alarm activity, hundreds of alarm records may be produced in a short time. The workload caused by flood episodes is often overwhelming: the operator cannot provide an appropriate response, and crucial alarms are likely to be missed (Ahmed et al., 2013). Usually, most of the alarms inside a flood episode come from a limited number of alarm sources (ANSI/ISA, 2016). Furthermore, alarm floods are strongly related to chattering alarms due to their potential to cause a large number of alarm events in a short time span. For this reason, identifying and removing chattering alarms is a crucial step to improve the performance of the alarm system and to avoid flood episodes.

2.2. Alarm attributes

Alarm events are described through a list of attributes. Each attribute defines a characteristic of an event such as the time of the alarm occurrence (i.e., the Timestamp), the Source that triggered the alarm, the alarm status, and more. Table 1 describes a list of attributes that are most frequently presented to the operator. It is worth mentioning that different companies use different messages and different sets of alarm attributes. The table is thus a selection of the most common and significant alarm attributes.

When an alarm is triggered, a message appears on the operator console. An example is:

"LI01 LEVEL D01 J434 PV = 98, 0 % HHH"

The alarm message reports the source of the alarm (the level indicator 01), a brief explanation of the measured variable (the level in drum 01), the associated safety function (J434), the value of the process value (98 %) and finally, the alarm status (High Level –i.e., "HHH").

For a more comprehensive understanding of the following analyses, the alarm identifier must be described more in detail. The identifiers "HHH" and "HTRP" inform that the measured variable has exceeded the "high" and "very high" threshold respectively, "LLL" and "LTRP" refer to the "low" and "very low" threshold, "IOP" informs about an instrumental failure or out-of-range measure, "ACK" indicates that the operator has acknowledged the alarm. The alarm identifier may include the word "Recover" (e.g. "HHH Recover", "LTRP Recover"), that indicates that the original alarm has been recovered (i.e., the alarm is not active anymore). In addition, two more attributes must be described, the Active Time Delta and the Condition Name, which have been used in the analyses but are not listed in Table 1. The Active Time Delta (ATD) is the number of seconds between an alarm and its recovery. The Condition Name (CN) is the alarm identifier of the initial alarm event (e.g., if the alarm is an "LLL Recover", CN will be "LLL").

In spite of the variety of different attributes, an alarm event is uniquely identified by three attributes only (Kondaveeti et al., 2010):

1. Time Stamp;
2. Source;
3. Alarm Identifier.

Also, the combination of an alarm source and an identifier (e.g., "LI01 HHH", "PI103 LTRP") is called a unique alarm (Kondaveeti et al., 2010).

2.3. Alarm databases

Chemical plants produce a massive amount of data on a daily basis (Kordic et al., 2010). Alarm events are continuously recorded and stored in alarm databases, which are characterized by a large search-space and may contain years of alarm data (Kordic et al., 2010). Typically, alarm events are collected as chronologically ordered time sequences (Weiss, 2010). Each row of the database represents an event, and each column represents an attribute of the alarm event. Obviously, there is not a single database format: different companies use different Distributed Control Systems (DCS). The format, the codes and the set of displayed features may vary accordingly. Typically, an alarm database contains more features than those presented in Table 1, but most of these additional features are either redundant or not useful for the analyses.

The analysis of the alarm history is a crucial step in monitoring the alarm system performance (ANSI/ISA, 2016). Periodic study of the alarm database allows the production of performance metrics and the detection of nuisance alarms. An example of a performance metric suggested by ANSI/ISA (2016) is the "Percentage of time the alarm system is in a flood condition", which must be lower than 1 % to grant stable operations. Furthermore, the standard states that chattering alarms must not be tolerated, and actions must be taken to resolve any chattering that occurs. Nevertheless, due to the complexity and quantity of data in alarm databases, the extraction of relevant information is not trivial and usually requires time and resources (Kordic et al., 2010).

3. Case-study: ammonia production plant layout and alarms

The industrial alarm database used in this study is provided by an international chemical company and consists of alarm data that were collected in a plant section for ammonia synthesis. The process involves the manipulation of a significant amount of dangerous substances (e.g., methane, hydrogen, ammonia), and severe operating conditions are often required (e.g., high temperature, high pressure, corrosive fluids). According to the Directive 2012/18/EU of the European Parliament and of the Council (European Union, 2012), the plant has been classified as an upper-tier establishment due to its potential to cause major accidents.

The ammonia production plant comprises four sections:

1. Desulfurization and Reforming;
2. Water-Gas Shift, CO2 Removal, and Methanation;
3. Ammonia synthesis and Cooling circuit;
4. Anhydrous ammonia storage, Pipeline, and Loading/unloading tankers.

Fig. 1 shows a schematic representation of the plant layout for ammonia production, excluding storage, loading, and unloading

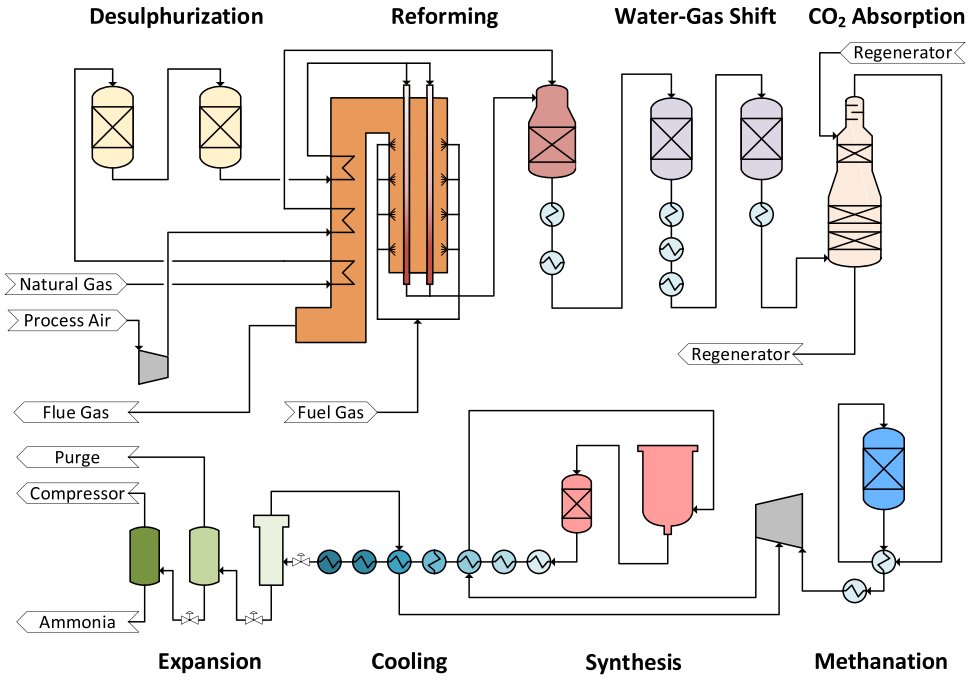


Fig. 1. Simplified process scheme of the ammonia production plant considered.

(Section 4). Natural Gas, Air, and Steam are used as raw materials for ammonia synthesis, according to the following exothermic reaction:

$$N_2 + 3\ H_2 \ \rightleftarrows\ 2\ NH_3 \qquad \Delta H_r^0[25°C] = -92.44\frac{KJ}{mol} \quad (1)$$

The reaction is carried out in two catalytic reactors arranged in series. The required nitrogen comes from process air, which enters the secondary reforming reactor. The hydrogen is produced through natural gas steam reforming in two distinct reforming stages. The first stage (primary reformer) is a vertical, side fired, proprietary reactor. The second stage (secondary reformer) is an autothermal adiabatic reactor. Typical temperature and pressure of the gas stream leaving the reforming section are 1000°C and 25 - 40 bar, respectively (Jennings, 1991). The catalysts used in the reforming reactors and in the downstream sections are sensitive to sulfur compounds (Aika et al., 1995). To avoid catalyst deactivation and poisoning, sulfur compounds are removed from natural gas in two reactors arranged in series. Similarly, carbon oxides must be removed because they are poisonous to the catalyst (Aika et al., 1995). For this reason, carbon monoxide is converted into carbon dioxide in two Water-Gas Shift reactors. Carbon dioxide is then removed in an absorption column where a Vetrocoke solution is used as a solvent (Giammarco and Giammarco, 1973). Finally, the residual amount of carbon oxides is removed in a Methanation reactor. The process stream leaving the methanator, which has the required purity for ammonia production, is compressed and sent to the ammonia synthesis loop, where ammonia is synthesized and liquefied through subsequent cooling and expansion units. Due to thermodynamic and kinetic constraints, the ammonia synthesis has low single-pass conversion (Jennings, 1991). Therefore, part of the gases released during the liquefaction process, which consists mainly of unreacted compounds, are recycled back to the reactors.

3.1. The alarm database

The alarm database of the ammonia plant contains alarm events collected between July 2017 and November 2017. In total, 26 473 alarm events (rows of the database) occurred during the observation period. Each event is described by a set of 39 attributes (columns of the database).

Alarms are not evenly spread over the observation period. The alarm daily annunciation rate is shown in Fig. 2. Over 96 % of the alarm events included in the database occurred between September and October 2017. The unusually high alarm rate was caused by a total power outage, which led to an unintended plant shut down. The plant instability and the abnormal alarm annunciation rate persisted over one month after the blackout, due to the emergency shutdown and the subsequent startup procedure. During the event, a significant number of alarm floods occurred. Therefore, the analyses described in this work have focused on that specific timelapse (September 9th to October 9th). Over the period of concern, 189 alarm sources produced a total of 25572 alarms. More than 72 % of them were triggered by ten alarm sources only, as shown in Fig. 3.

4. Methodology

The approach follows the steps depicted in Fig. 4.

4.1. Data preprocessing

Raw alarm data must be prepared for the analysis.

4.1.1. Attribute selection and data cleaning

The columns of the database that are either empty or not useful for the analysis have been removed (step 1.1 in Fig. 4). For example, the column “PlantHierarchy” contains standardized codes that refer to a specific plant inside the production site. The column has been removed because every alarm considered in this study comes from the ammonia production plant.

Machine Learning algorithms cannot process null (missing) values. For this reason, columns including null values were further analyzed, and, when relevant, null values were substituted by specific input values. Several techniques exist to impute missing values (Hastie et al., 2009). If the value is relevant for the analysis,

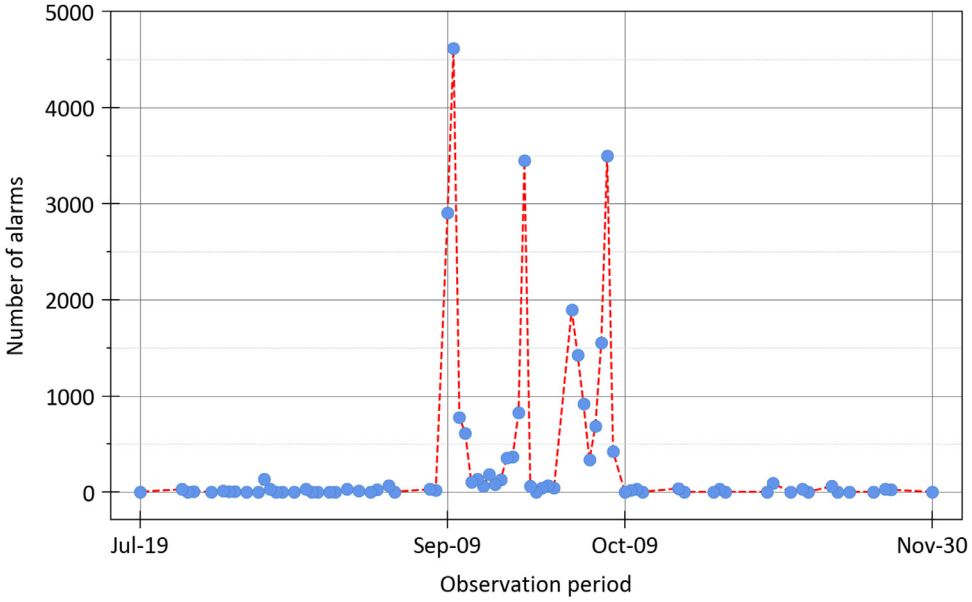


Fig. 2. Alarms' daily annunciation rate. Each circle represents the number of alarms that occurred during one day.

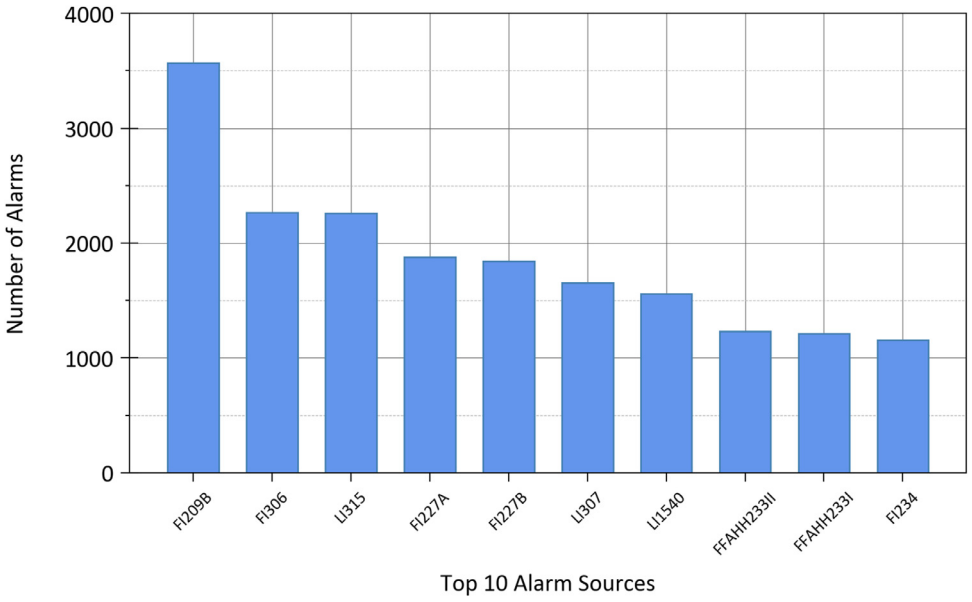


Fig. 3. The ten alarm sources with the larger alarm count over the observation period (September – October 2017).

one may decide to replace it with the mean or median of the non-missing values (Brink et al., 2016). If the missing value is not relevant, or if there is no way to guess the value through statistical calculations, it might be replaced with a user-defined global constant (Han et al., 2012). For example, the column "Eng. Unit" (see Table 1) contains a considerable amount of missing value due to alarms that are not associated with a measuring instrument (e.g. alarm generated by ad-hoc logics). Therefore, it has been decided to replace the missing values in the column with the symbol "– " (step 1.2 in Fig. 4).

As a result, a "clean" database is obtained (step 1.3 in Fig. 4), which contains only meaningful attributes and no missing values.

4.1.2. Binary Database creation

Unique alarms (i.e. the unique combination of an alarm source and an identifier) have been represented as binary sequences (step 1.4 in Fig. 4). According to Kondaveeti et al. (2010), the binary representation of a unique alarm is an array whose elements represent one-second-spaced time bins. For a one-month-long observation period, the array has 2592000 elements (i.e., seconds in one month). The value of an element of the array can be either "1" or "0". A "1" in the sequence indicates that the unique alarm occurred at that very moment. On the contrary, a "0" means that the unique alarm did not happen. In this way, alarm occurrences are represented as 1's in the array. Finally, binary sequences are grouped in a matrix (step 1.5 in Fig. 4). Rows containing zeroes only can be safely removed by the Binary Database (Kondaveeti et al., 2010).

Although it is not compulsory, representing alarm data as binary sequences will greatly simplify the calculation of the Dynamic Chattering Index.

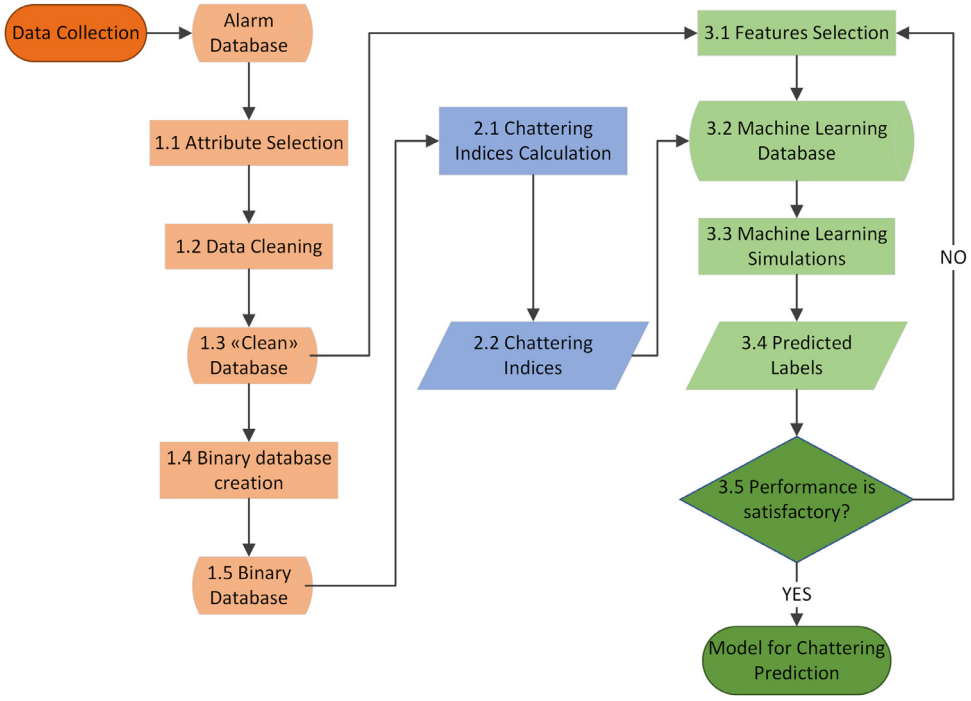


Fig. 4. Workflow of the analysis carried out. Colors represent three main stages. Stage 1 (orange): Data preprocess, Stage 2 (blue): Dynamic Chattering Indices calculation, Stage 3 (green): Machine Learning simulations. Each stage is divided into several steps, which are arranged chronologically and identified by two numbers (e.g., 1.1, 2.2, 3.5). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

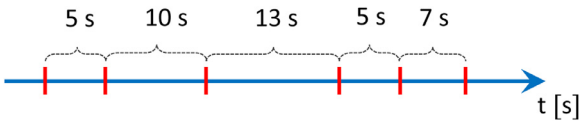


Fig. 5. Schematic representation of a unique alarm. Red sticks represent alarm events. Seconds between two subsequent sticks represent a run-length. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

4.2. The Dynamic Chattering Index

The mathematical formulation of the Dynamic Chattering Index is based on the method described by Kondaveeti et al. (2013) for the calculation of the Chattering Index. Both indices rely on the run-lengths distribution to quantify alarm chatter. A run-length is the "time difference in seconds between two consecutive alarms on the same tag" (Kondaveeti et al., 2013). As an example, the run-lengths related to a fictitious unique alarm are shown in Fig. 5.

In the figure, the red marks represent an alarm event (i.e., a "1" in the binary sequence), the number of seconds between two consecutive marks (i.e., the time between two subsequent "1's" in the binary sequence) represent a run-length. Intuitively, considering the whole observation period, if a unique alarm has a high number of short run-lengths, it is highly probable that the alarm has shown chattering. The alarm sequence described in Fig. 5 indicates Chattering because the alarm has occurred six times in less than 30 seconds. The Chattering Index in Kondaveeti et al. (2013) is

$$\psi = \sum_{r \in \mathbb{N}} P_r \frac{1}{r} \tag{2}$$

where

- r is a natural number that represents a run-length [s] (e.g. r = 5, 7, 10, 13 for alarm data in Fig. 5);
- P_r represents the probability that a run length is equal to r seconds, i.e.,

$$P_r = \frac{n_r}{\sum_{r \in N} n_r} \tag{3}$$

where

- n_r is the number of run lengths equal to r seconds (e.g. $n_5 = 2$ and $n_7 = n_{10} = n_{13} = 1$ for alarm data in Fig. 5);
- $\sum_{r \in \mathbb{N}} n_r$ represents the total number of run-lengths, which is one less than the unique alarm's occurrences over the observation period (e.g., the alarm in Fig. 5 occurred 6 times, the summation is equal to 5).

The Chattering Index indicates the mean frequency of annunciation of a unique alarm (units of ψ are $\frac{\text{alarms}}{\text{s}}$), and it assumes a value between 0 and 1. A threshold value is needed to assess whether a unique alarm has shown chattering during the observation period. Kondaveeti et al. (2013) propose a threshold value equal to 0.05 (i.e., $\psi \geq 0.05$ indicates alarm chatter). For instance, considering the example presented in Fig. 5, the probabilities (Eq. (3)) are $P_5 = {}^2/_5$ and $P_7 = P_{10} = P_{13} = {}^1/_5$. The Chattering Index (Eq. (2)) is:

$$\psi = \frac{2}{5} \cdot \frac{1}{2} + \frac{1}{5} \cdot \left(\frac{1}{7} + \frac{1}{10} + \frac{1}{13} \right) = 0.26 \geq 0.05 \tag{4}$$

which confirms that the alarm has shown Chattering behavior.

In fact, once the observation period is defined, a single ψ is obtained for each unique alarm. Although meaningful, the index is relatively static: observing the Chattering Index, one can determine whether the unique alarm has shown Chattering, but no further conclusion can be drawn (e.g., when exactly the alarm has shown Chattering). To overcome this limitation, the Chattering Index approach has been modified, and the Dynamic Chattering Index has been developed. The core idea is to calculate a regular Chattering Index every time a unique alarm occurs (i.e., every time a "1" is

found in the binary representation of the alarm). Another key feature is that the calculations of the Dynamic Chattering Index involve only alarm events that occurred up to one hour after the alarm event of concern. If a unique alarm has n 1's in the binary sequence, n-1 Dynamic Chattering Indices are calculated (the last event is excluded from the calculation). By this procedure, each "1" in the binary sequence is associated with a Dynamic Chattering Index, which quantifies the amount of chatter that the alarm has shown during the following hour.

Considering a generic alarm event with index i, the calculation of the Dynamic Chattering Index involves four steps (step 2.1 in Fig. 4).

1. The largest index k that meets the condition ($Timestamp_k - Timestamp_i) \leq 1\ h$ is selected, and the binary sequence is reduced in such a way that only alarm events having index $j \in [i, k]$ are taken.
2. The run-lengths (r) and the number of run-lengths (nr) of the reduced binary sequence are calculated. It is worth noting that as long as the reduced binary sequence does not include the last alarm event (i.e., if k < n), one run-length can be obtained for each of the alarms in the reduced sequence (i.e., the last element of the sequence is also included in the run-lengths calculations).
3. Probabilities are calculated according to Eq. (3).
4. The Dynamic Chattering Index of the alarm event is calculated according to Eq. (2).

The steps presented above are repeated $\forall\ i \in [0,\ n-1]$ to obtain the n - 1 Dynamic Chattering Indices of the unique alarm of concern. Finally, the procedure is repeated for each of the unique alarms in the Binary Database.

The same threshold value discussed above has been used for alarm classification. If an alarm event has $\psi_D \geq 0.05$, the unique alarm will show chattering within one hour.

Eventually, a Dynamic Chattering Index has been calculated for each alarm event (step 2.2 in Fig. 4). The use of a threshold allows classifying alarms into two categories, "Chattering within one hour" and "Not Chattering within one hour". This result has been used to train and evaluate the Machine Learning algorithms that will be described in the following section.

4.3. Machine Learning

Machine Learning (ML) can be defined as "computational methods using experience to improve performance or to make accurate predictions" (Mohri et al., 2012). Due to the ever-increasing computation capabilities of modern calculators and to the development of computer technologies, the number of ML algorithms and their applications have witnessed extraordinary growth during the last few years (Liu et al., 2018). Despite the immense number of different algorithms, there are only three categories of ML methods, which are: Supervised Learning, Unsupervised Learning, and Reinforcement Learning. Within the present work, Binary Classification algorithms have been used, which fall into the Supervised Learning category.

A Classification algorithm takes as an input a list of features (i.e., meaningful attributes) of the object that must be classified and returns a label (i.e., the class of the object). For instance, these algorithms are employed to classify emails into "Spam" and "Not Spam" while, in the present study, the algorithm aims to classify alarm events into "Chattering within one hour" or "No Chattering within one hour". If the objects are classified into two classes only, the problem is called Binary Classification.

The selection of the most relevant features is a crucial step, and it may significantly affect the performance of the algorithm. The selection of the set of features that best represent the problem under assessment is mostly guided by experience, and a trial and error approach is often required (Brink et al., 2016)(step 3.5 in Fig. 4).

The Machine Learning Classification workflow is presented in Fig. 6. Two distinct stages are necessary to build and test the algorithm: Training and Evaluation.

During the training stage (step 2 in Fig. 6), the algorithm receives a set of examples. An example is a list of features (e.g., the attributes of an alarm event) and the related label. From the examples, the algorithm "learns" the relation between features (Y) and labels (X) by optimizing the weights of an internal function (f).

$$Y = f\ (X) \tag{5}$$

The weights are adjusted by an optimization algorithm, which aims to minimize the "distance" between f(X) and Y. Different types of functions exist, as well as different optimization methods.

Later, during the evaluation phase (step 3 in Fig. 6), a new series of unlabeled examples (i.e., only features) are fed to the trained algorithm, which predicts the labels. Finally, the performance of the algorithm is quantified by comparing predicted labels with true labels (step 4 in Fig. 6).

It is worth mentioning that the raw output of a Classification algorithm is not a label, but the label's probability (Brink et al., 2016). For example, the algorithm used in this work returns the likelihood of a unique alarm being "Chattering within one hour" or "Not chattering within one hour". A threshold is needed to convert probabilities into the final label, which is 0.5 by default (i.e., if the "Chattering within one hour" label's probability is $\geq$ 0.5, the model will label the alarm as "Chattering within one hour"). The probability threshold is an adjustable parameter, and it can significantly affect the model's performance (Google, 2020a).

4.3.1. Models

A model can be defined as "a function with learnable parameters that maps an input to an output. The optimal parameters are obtained by training the model on data. A well-trained model will provide an accurate mapping from the input to the desired output" (TensorFlow.org, 2020a). Basically, the model defines the mathematical structure of the function *f* in Eq. (5). Three different models have been used in this study: a Linear model, a Deep Neural Network, and a Wide&Deep model.

4.3.1.1. Linear model. Linear models represent the labels as a linear combination of the features (Hastie et al., 2009).

$$Y = \beta_0 + \sum_{j=1}^{p} X_j \beta_j \tag{6}$$

where:

- Y = labels;
- X = [X_1, X_2, …, X_p] = the features vector;
- X_j = a feature;
- β_0 = intercept (or bias);
- β_j = coefficient (or weight).

In this representation, each feature has its own weight. Therefore, the model can assess how much a feature weights on the calculation of the label, but it cannot quantify the influence of combinations of features. This limitation is partially solved by crossing two or more features to create a new, more meaningful, synthetic feature (Google, 2020b). Despite that, the linear model lacks in generalization, and it cannot interpret the combination of features that never occurred during the training phase (Cheng et al., 2016).

Although simple, the model is widely used (James et al., 2013) because it is robust, fast, and performs well on large datasets. Furthermore, the weights values are easily accessible, allowing the

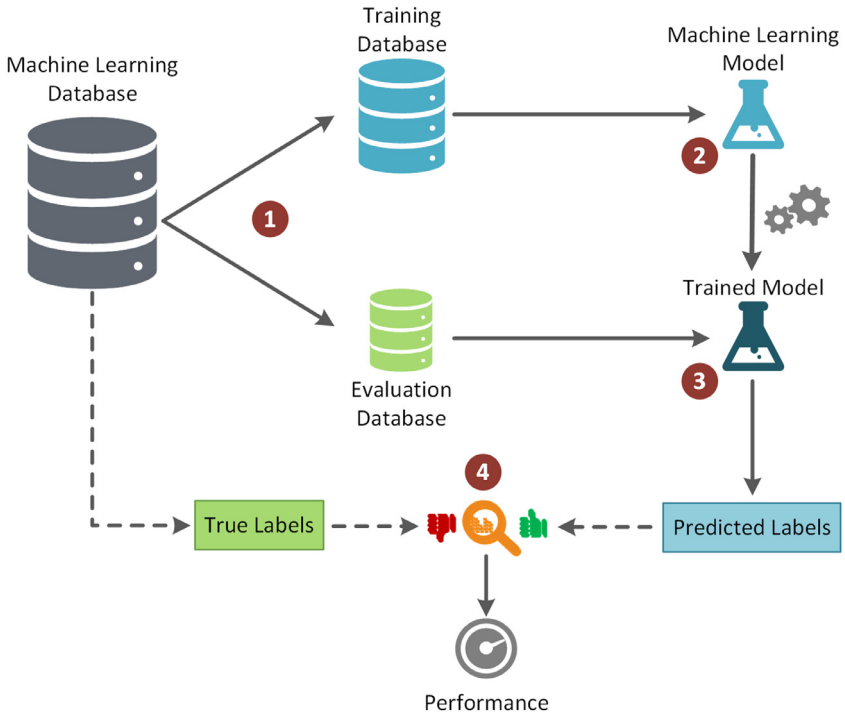


Fig. 6. Binary Classification Workflow. 1 - the database is divided into Training and Evaluation Databases. 2 - Training Database is fed to the ML model; a trained model is obtained. 3 - Evaluation Database is fed to the trained model, which predicts the labels. 4 - performance metrics are calculated.

user to evaluate which features are more meaningful for the problem under assessment (Brink et al., 2016).

The model employed in this work uses FTRL algorithm as an optimizer (TensorFlow.org, 2020b).

4.3.1.2. Deep Neural Network. Deep Neural Networks consist of interconnected layers. The first layer of the network is the vector of the features (X), and the last layer is the vector of the labels (Y). Between the first and the last layer there are the so-called hidden layers (H). Each hidden layer is made of a certain number of hidden units (Z). The number of hidden layers and hidden units is a design parameter that can greatly affect the performance of the algorithm. Generally speaking, it is better to use a large number of hidden layers and units. As a drawback, bigger networks require more computational effort than networks with few layers. The model used in this work has three hidden layers, with 1024, 512 and 256 hidden units, respectively. A schematic representation of a Neural Network is shown in Fig. 7.

The connections (i.e., solid lines) in Fig. 7 represent non-linear transformations. For example, the hidden units of the first and second hidden layers can be calculated as follows

$$Z_i^1 = \sigma\left(\alpha_{0i} + \alpha_i^T\ X\right) \qquad i\ =\ 1,\ \ldots,\ M \tag{7}$$

$$Z_i^2 = \sigma\ \left(\gamma_{0i} + \gamma_i^T\ Z^1\right) \qquad i\ =\ 1,\ \ldots,\ N \tag{8}$$

where:

- $\alpha_{0i},\ \gamma_{0j}\ =$ biases;
- $\alpha_i,\ \gamma_i\ =$ vectors of model coefficients;
- $Z_i^k\ =$ the i – th hidden unit of the k – th hidden layer;
- $Z^1 = [Z_1^1,\ Z_2^1,\ Z_3^1,\ \ldots,\ Z_M^1]$;
- $\sigma\ =$ activation function.

Biases and coefficients are optimized during the training of the algorithm. The model employed in this work uses Adagrad algorithm as an optimizer (TensorFlow.org, 2020c), and the activation

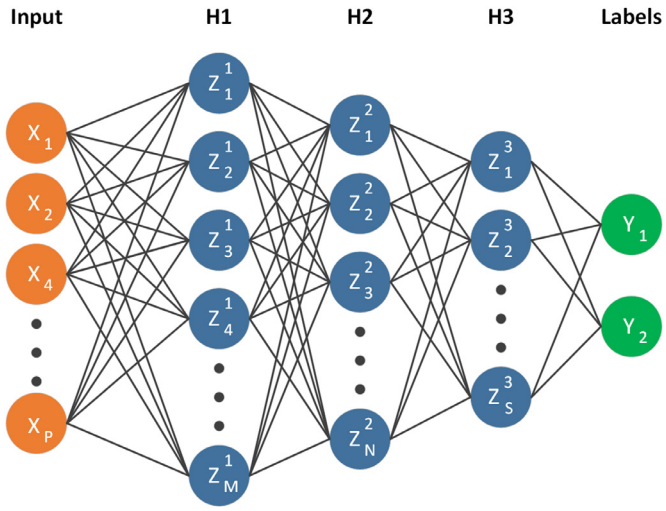


Fig. 7. Deep Neural Network with P features (orange circles), and three hidden layers (H1, H2, H3), which contain M, N, and S hidden units, respectively. The output layer (green circles) contains two labels (Y1 and Y2). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

function is the linear rectifier (TensorFlow.org, 2020d).

$$\sigma\,(x) = \max\,(0,\ x) \tag{9}$$

According to Eqs. (7) and (8), the activation function converts the linearly combined units of a layer into the hidden units of the following layer; this allows the model to capture non-linear inter-features relationships and strengthen its generalization capabilities.

Deep Neural Networks represent state-of-the-art algorithms for audio-video processing (i.e., speech and image recognition) (Brink et al., 2016; Hastie et al., 2009) and their applications are rapidly spreading among different sectors. Although flexible, these

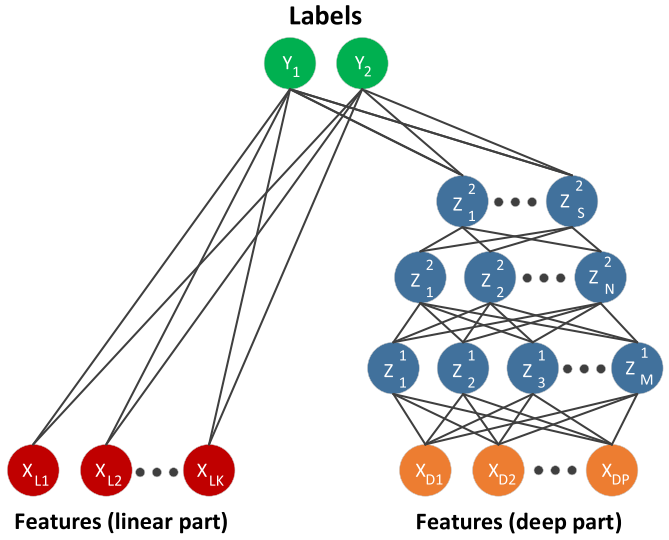


Fig. 8. Wide&Deep model, made of a linear (left) and a Deep (right) parts. The linear part takes K features (red circles). The Deep part is made of 3 hidden layers (blue circles) with M, N, and S hidden units, and takes P features (orange circles). The output layer (green circles) contains two labels (Y1 and Y2). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

models may over-generalize and detect non-existent relationships between features. Furthermore, they are harder to optimize compared with simpler models (e.g., the linear model).

4.3.1.3. Wide and Deep model. In an attempt to overcome the limitations of the models discussed above, Cheng et al. (2016) proposed a hybrid model, which is composed of a Wide part (i.e., linear) and a Deep part (i.e., Deep Neural Network), as shown in Fig. 8.

During the training phase, the parameters of the linear and deep parts are optimized simultaneously using FTRL and Adagrad algorithms. The linear part of the model could comprise both raw features and crossed-features; in this work, only crossed-features are used. The hybrid model has proven to combine the advantages of the linear model (e.g., robustness, memorization capability) and the Deep model (e.g., generalization, flexibility) minimizing their drawbacks (Cheng et al., 2016).

4.3.2. Performance indicators

The performance of a classification algorithm can be assessed by comparing predicted labels and true labels. For concision purposes, the label "No Chattering within one hour" will be referred to as the label "N", while "Chattering within one hour" will be referred to as the label "Y". Three metrics have been used to assess the performance

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{10}$$

$$Accuracy = \frac{TP}{TP + FP} \tag{11}$$

$$Accuracy = \frac{TP}{TP + FN} \tag{12}$$

where

- TP = True Positive –i.e. predicted label = Y, true label = Y;
- TN = True Negative –i.e. predicted label = N, true label = N;
- FP = False Positive –i.e. predicted label = Y, true label = N;
- FN = False Negative –i.e. predicted label = N, true label = Y.

Table 2
Alarm features used in the simulations. Features names have been coded for concision purposes. Codes are represented in the first column and described in the second column. The last two columns represent the features format used in the linear and Deep models.

Feature	Description	Format	
		Linear	Deep
Y	Year of the alarm event	Num.	Num.
M	Minute of the alarm event	Num.	Num.
D	Day of the alarm event	Num.	Num.
H	Hour of the alarm event	Num.	Num.
m	Minute of the alarm event	Num.	Num.
S	Seconds of the alarm event	Num.	Num.
SO	Source (see Table 1)	Categ	Dense
ID	Identifier (see Table 1)	Categ.	Dense
CN	Condition Name (see Section 2.2)	Categ	Dense
JX	Alarm Safety function (see Table 1)	Categ.	Dense
ATD	Active Time Delta (see Section 2.2)	Num.	Num.
VAL	Data Value (see Table 1)	Num.	Num.
UNI	Eng. Unit (see Table 1)	Categ.	Dense

The summation of TP and TN represents the number of correct predictions and the summation of FN and FP is the number of wrong predictions. The Accuracy is the number of correct predictions divided by the total number of predictions, the Precision is the fraction of correct positive predictions, and the Recall is the fraction of real positive correctly predicted; the metrics assume values between 0 and 1; the larger the value, the better the metric.

As it has already been discussed, Machine Learning algorithms use a probability threshold to determine the predicted label. Therefore, changing the probability threshold can greatly affect the algorithm's performance as it modifies the values of TP, TN, FP, and FN. Unfortunately, Precision and Recall are often in tension (Google, 2020a), changes in the threshold that aim to increase the Precision may cause the Recall to decrease, and vice versa.

It is worth noting that all the metrics discussed above must be considered to evaluate the performance of a Machine Learning algorithm (Google, 2020c). A high Accuracy alone is meaningless and does not necessarily indicate good performances. In this work, "legitimate" alarms (i.e., that are not going to show chattering) must not be labeled as chattering ones. Therefore, the Precision is the metric that must be optimized.

5. Simulations

Three simulations have been performed, one for each model described in Section 4.3.1. The Machine Learning algorithms have been built using TensorFlow r1.15 (TensorFlow.org, 2020e) running on Python 3.7.4 (Python.org, 2019). The first step to build the Machine Learning algorithms is the feature selection (step 3.1 in Fig. 4). A preliminary screening has already been performed during "Attribute selection and data cleaning" (section 11) when the not useful columns have been removed from the raw alarm database. Still, there is no guarantee that the algorithms will perform better if all columns of the "clean" database are used as features. As previously argued, feature selection often requires a trial and error approach. Different features have been tested. The best set (i.e., the one that has generated the best performance) is presented in Table 2, which contains the name and description of each feature.

After features selection, the alarm database has been reorganized and converted into a new database (step 3.2 in Fig. 4), which contains only the features listed in Table 2. Each row of the new database represents an alarm event, each of the first thirteen columns represents a feature, and the last column contains the labels. A label can be either "Y" (if $\psi_D \geq 0.05$ –i.e., the unique alarm will show chattering within one hour) or "N" (if $\psi_D < 0.05$ –i.e., the unique alarm will not show chattering within one hour).

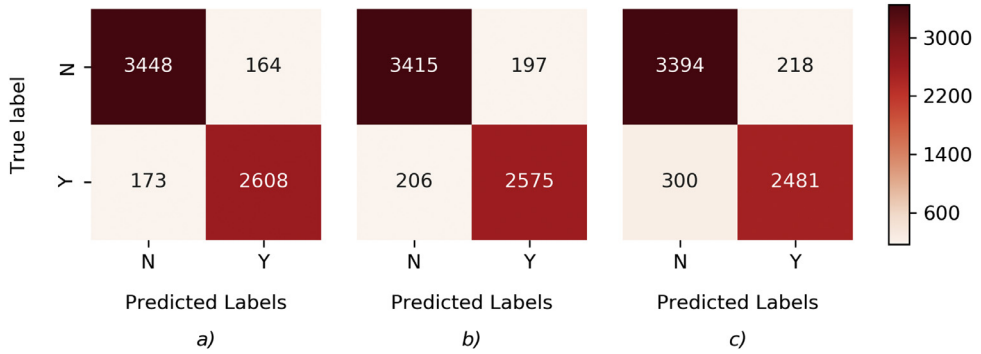


Fig. 9. Confusion matrices of the linear (a), Deep (b), and Wide&Deep (c) models. The label "N" means "No chattering within one hour", "Y" means "Chattering within one hour". TN, FP, TN, and FN are obtained using a probability threshold equal to 0.5 and color-coded according to the color bar on the right. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

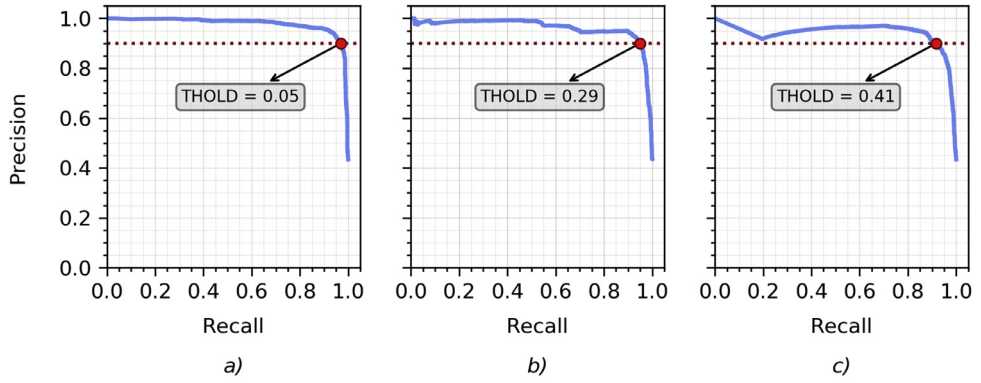


Fig. 10. Precision-Recall curves of the linear (a), Deep (b), and Wide&Deep (c) models. Probability thresholds between 0 and 1 have been used. Points of the curves represent the couple Precision – Recall at a specific threshold. Proceeding from Recall = 0 to Recall = 1, the threshold decreases from 1 to 0 in a non-linear fashion. Red markers indicate Precision = 0.9, which is obtained at Threshold = THOLD. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

Next, the database has been divided into two sections, to obtain the training database and the evaluation database (step 1 in Fig. 6). The first part, which contains ¾ of the alarm data, has been used to train the models, and the second part (¼ of the database) to evaluate them. The last column of the evaluation database has been removed and stored as a separate variable. This prevents the model from gaining access to the actual labels during the evaluation. Additionally, since chattering alarms are not spread evenly throughout the original database, the database has been shuffled before the division (i.e., rows have been randomly rearranged).

Prior to starting the simulations, one must ensure that the features are fed to the algorithm in a proper format. While Machine Learning models accept numerical features as an input, strings of characters (e.g., the features Identifier, Source, and Eng. Unit in Table 1) cannot be fed directly into the model and must be converted into a categorical or dense format. Non-numerical features that are fed to the linear model have been converted into categorical features (TensorFlow.org, 2020f). On the contrary, Deep Neural Networks do not accept categorical features as an input. Therefore, non-numerical features have been mapped into dense features (TensorFlow.org, 2020g). The last two columns of Table 2 summarize the features format used in the linear and deep models. Furthermore, three crossed features have been used in the linear model and in the linear part of the Wide&Deep model, which are [SO x CN x JX], [SO x CN x ID], and [VAL x UNI]. This enables the linear model to assess non-linear relationships between features. In summary, the Linear model uses three crossed features in addition to the features in Table 2. The Deep model and the deep part of the Wide&Deep model use the features in Table 2, but no crossed features (because Deep models have intrinsic generalization capabilities). The linear portion of the Wide&Deep model uses three crossed features only.

Finally, the models have been trained and evaluated, as shown in Fig. 6 (step 3.3 in Fig. 4). After the simulations, the algorithm provided raw label probabilities, which have been converted into predicted labels using 0.5 as a threshold (step 3.4 in Fig. 4). Next, predicted labels have been compared with true labels (i.e., the labels that the model should have predicted), the number of TP, TN, FP, and FN has been calculated. Accuracy, Precision, and Recall of the model have been obtained according to Eqs. (10)–(12). Finally, the performance metrics have been calculated again using different probability thresholds in order to study the effect of this parameter on the final results.

6. Results

The number of TP, TN, FP, and FN are presented in Fig. 9, which contains three confusion matrices, one for each model. The axes of a confusion matrix represent the true labels and the predicted labels. From top left clockwise, the elements of a confusion matrix are the number of true negatives, false positives, true positives, and false negatives. A probability threshold equal to 0.5 has been used.

Metrics in Fig. 9 indicate that the number of correct predictions is one order of magnitude higher than the number of wrong predictions. Moreover, the number of False Positives is always lower

Table 3
Accuracy, precision, and recall achieved by the Machine Learning models. Metrics are obtained using a probability threshold equal to 0.5.

Model	Accuracy	Precision	Recall
Linear	0.947	0.941	0.938
Deep	0.937	0.929	0.926
Wide&Deep	0.919	0.919	0.892

than the number of False Negatives. The Accuracy, Precision, and Recall achieved by each algorithm are shown in Table 3.

Values in Table 3 indicate that the linear model produces the largest metrics. Similarly, the Deep model achieves a better performance than the Wide&Deep model.

Fig. 10, shows the Precision-Recall (P-R) curves of the three models calculated using different probability thresholds.

As previously stated, classification algorithms provide the label probability of the events that are included in the evaluation database. A threshold value is needed to convert probabilities into labels. If the threshold is equal to 0, every alarm event in the evaluation database will be labeled as "Y". Oppositely, if the threshold is equal to 1, every alarm event in the evaluation database will be labeled as "N". Lowering the threshold causes the Recall to either decrease or to remain constant. Instead, Precision may increase or decrease when the threshold is reduced. Each point of the blue curves in Fig. 10 represents the Precision and Recall values obtained using a specific threshold. For a specific model (panels a, b, and c in Fig. 10), thresholds larger than THOLD ensures a Precision larger than 0.9.

7. Discussion

The Chattering Index proposed by Kondaveeti et al. (2013) is a valuable tool for addressing Chattering alarms retrospectively, but it does not fulfill the need for dynamicity required to achieve the objectives of the study. In fact, the Chattering Index quantifies the amount of chattering that an alarm has shown over the entire observation period: results are static, meaning that the index can be used to measure the chattering severity, but it does not provide any information about when, or why, the chattering has happened. For these reasons, the index has been modified, and a dynamic approach has been developed.

The Dynamic Chattering Index aims at quantifying the likelihood of alarm chatter after each alarm occurrence, linking past and actual process conditions to future alarm behavior. A threshold has been used to classify alarm events in two categories, "Chattering within one hour" and "Not chattering within one hour". The Dynamic version of the Chattering Index provides a more detailed picture of the alarm system performance if compared to the Chattering Index: the former classifies alarm events, the latter classifies unique alarms. In future works, the Chattering Index may be used to strengthen the Machine Learning simulations since it represents a meaningful piece of information about the past behavior of a unique alarm. For instance, one Chattering Index may be calculated for each alarm event in the database, taking into account only alarms that occurred before each event. This index may be used as a new feature in the Machine Learning simulations, allowing the model to learn the relation between past and future chattering. The approach has not been pursued in this study because the authors decided to exclude synthetic features (i.e., that requires calculation) and focus the attention on ready-to-use features (i.e., directly provided by the alarm system).

The Dynamic Chattering Index method requires to select a threshold (for alarm classification) and the length of the time interval (to obtain the reduced binary sequence, according to step 1 of the procedure described in Section 4.2). In this work, a time interval equal to 1 h has been used because it appears to be a good balance between dynamicity (that cannot be achieved using large time intervals) and statistical relevance (that cannot be achieved using short time intervals). However, the choice has been arbitrary and guided by general considerations. For example, longer time intervals (e.g., 2 hours or more) may cause the index to detect Chattering even if it only appears in the last minutes of the time sequence. As a result, the index would indicate chattering for two or more hours while the alarm would not exhibit chattering for most of the time. Oppositely, shorter time intervals (e.g., 30 minutes or less) may cause the index to overestimate short run lengths and to detect chattering were no – or low – chattering exists; this issue partially affects also the index used in this study. In fact, the Dynamic Chattering Index relies strongly on statistical methods, which perform better when a large amount of data is analyzed. Unlike the Chattering Index, which considers the entire observation period, the Dynamic Chattering Index calculations involve a relatively short time interval. It may happen that the unique alarm under assessment occurred a few times during the hour, and this could lead to unexpected results. For instance, few alarm events in the reduced binary sequence will produce relatively large probabilities, since the denominator in Eq. (3) will be small. Besides, if some of the few run-lengths involved in the calculation are short (e.g., 1 – 10 s), the combination of short run-lengths and large probabilities will cause Eq. (2) to produce a large Dynamic Chattering Index, most likely higher than 0.05 Tamascelli et al., 2020.

Table 4
Fictitious alarm sequence. Each row represents an alarm event. The last column contains the Dynamic Chattering Index of the event i (first row the table). The symbol " \ " indicates a value that is either not calculated or not relevant for the analysis.

Index	Timestamp	Run-length [s]	ψ_D
i	09/09/2017 16:07:24	3	0.069
i+1	09/09/2017 16:07:27	234	\
i+2	09/09/2017 16:11:21	133	\
i+3	09/09/2017 16:13:34	1559	\
i+4	09/09/2017 16:39:33	2160	\
i+5	09/09/2017 17:15:33	\	\

As an example, consider alarm data represented in Table 4. The calculation of the Dynamic Chattering Index of the event i includes all the alarms in Table 4 except the last one (because it happened later than one hour after the event i). Observing the run-lengths, one may conclude that the alarm did not show chattering since they appear to be long enough, and the 3 seconds long run-length alone does not seem sufficient to suggest chattering. Despite that, the calculation of the Dynamic Chattering Index leads to an unexpected result. In particular, the run-length count and the probabilities are $n_r = 1$ and $P_r = 1/5 \ \forall$ r. Therefore, the Dynamic Chattering Index is

$$\psi_D = \frac{1}{5}\cdot\left(\frac{1}{3}+\frac{1}{234}+\cdots+\frac{1}{2160}\right) = 0.067 + 8.5\cdot 10^{-4} + \cdots + 9.2\cdot 10^{-5} = 0.069 \quad (13)$$

Which is greater than 0.05, and suggests chattering within one hour. Focusing on how each run-length impacts the index calculation, one can observe that a run-length equal to 3 s alone produces a contribution of 0.067, which is greater than 0.05. This behavior is due both to an extremely short run length and to large probabilities (caused by few alarms being triggered during the observation period). Usually, if many alarms occurred within the observation period, the effect of a few short run-lengths is mitigated by a small probability value. Instead, if few alarms were triggered, the probability increases, the mitigation effect stops, and an unreliably

high ψ_D is produced. The issue might be avoided by excluding extremely short run-lengths or extending the time interval. The first solution has the disadvantage of ignoring extreme chattering behavior, and the second may cause loss of dynamicity. For the reasons mentioned above, future research should be devoted to the improvement of the Dynamic Chattering Index calculation in order to achieve higher reliability. For example, alarm sequences affected by the issues described above might be isolated, and different indices with different time intervals (e.g., 30 minutes, 1 hour, and 2 hours) might be tested on these sequences to assess which one performs better (i.e., which one does not overestimate the effect of few, extremely short, run lengths). Also, the Dynamic Chattering Index described in this study might be modified to take into account the number of alarm events considered in the calculation –e.g., a weighting function might be created to dampen the effects of the combination of few alarms and short run lengths. In addition, indices calculated with different time intervals may be aggregated and used together to obtain a single, more comprehensive, and informative index.

The results of the Dynamic Chattering Index approach have been used to train three Machine Learning models for Chattering prediction. The models have shown good performances in predicting alarm chatter: results in Table 3 indicate that a high Accuracy can be achieved while maintaining high Precision. These flexible and dynamic tools may significantly improve the operators' response in different situations. During alarm floods, early warning of chattering may be delivered to the operator, who may decide to silence the alarm before it becomes a nuisance. In addition, the models could warn that the chattering is going to end (i.e., the model predicts an "N" after a sequence of "Y"), and the operator may decide to restore the alarm without the burden of checking it periodically. During normal operations, early warnings of chattering may allow the operator to investigate the issue in advance, and the ability to detect the end of a chattering sequence would prevent the alarm from being forgotten in a silenced status. In general, the models could help to increase risk awareness by providing quick and ready-to-use information and by reducing the need for manual intervention.

When the standard probability threshold is used (i.e., 0.5), the linear model qualifies as the best model since it produces the largest metrics (Table 3). Deep and Wide&Deep models show slightly smaller metrics and may need more optimization to improve their performance. On the contrary, the simpler but more robust linear model has performed better without the need of a specific optimization. The reasons why this has happened are diverse. For instance, DNN and Wide&Deep models are prone to overgeneralization and may detect inter-feature relationships where no relationship exists. The problem described in this study may need a model that is better at memorizing (e.g., Linear) rather than generalizing (e.g., DNN, Wide&Deep). Future research should investigate whether different optimization strategies (e.g., different hyperparameters, learning decay, activation functions) could improve the performance of advanced but sensitive models such as the Deep and Wide&Deep.

P-R curves in Fig. 10 suggest that precisions larger than 0.9 can always be achieved while maintaining the Recall close to 0.9 by varying the probability threshold. If the threshold is further reduced (i.e., below 0.05 for the linear model, 0.29 for the Deep, and 0.41 for the Wide&Deep), the Precision drops significantly. The selection of the best threshold (i.e., threshold tuning) strongly depends on the specific problem under assessment (e.g., unbalanced/balanced dataset, cost-sensitive/insensitive classification)(Brink et al., 2016; Google, 2020d; Ling and Sheng, 2008). Misclassifying legitimate alarms (FP) is more critical than misclassifying chattering alarms (FN) as a False Positive may cause the operator to silence a legitimate alarm. Therefore, False Positives must be avoided, and Precision must be increased. Unfortunately, increasing the Precision often causes the Recall to decrease (Brink et al., 2016). The best threshold must ensure a high Precision while maintaining the Recall to an adequate level. Acceptable thresholds may be identified by selecting minimum values of Precision and Recalls, but selecting the best threshold requires more considerations. Often when classification errors have different criticality, a process similar to cost-benefit analysis is needed to identify the best threshold value (Ling and Sheng, 2008; Sheng and Ling, 2006). Other approaches involve the optimization of the weighted harmonic mean between Precision and Recall (F_β - measure)(Chai, 2005; Paltrinieri et al., 2020).

As a final note on thresholds and P-R curves, it is worth noting that the linear model provides Precisions greater than 0.90 when thresholds between 1 and 0.05 are used. This means that whenever the model predicts the label "1" (i.e., "Chattering within one hour"), it produces a large probability value, which is often larger than 0.95 (i.e., 1 - 0.05). In other words, the linear model is extremely "confident" when predicting chattering alarms.

Focusing on the Linear model, the nature of wrong predictions (i.e., FN and FP) has been studied more in detail. Three leading causes of error have been identified:

1. The model could not identify the beginning of a chattering series.
2. The model could not identify the end of a chattering series.
3. The model labels all the events of the unique alarm of concern as "Y" or "N".

Cause 1 occurs when the model fails to identify the first element of a Chattering sequence or, in other words, it fails to detect the first unique alarm event labeled as "Y" after one or more events labeled as "N". Fig. 11 clarifies this insight. As one might notice, the first event of the chattering sequence (the red dot in Fig. 11) has been incorrectly labeled (true label is Y, predicted label is N), and a False Negative has been produced as a consequence. Later in time, the model has correctly identified chattering (green dots). Also, the model has correctly predicted the end of the chattering sequence, which occurred at 13:56:00 (not displayed in Fig. 11).

Cause 2 occurs when the model fails to identify the last element of a Chattering sequence. Fig. 12 provides an example of this. The last two unique alarm events of the series (red dots) have been incorrectly labeled (the true label is N, while the predicted label is Y), and two False Positive have been produced as a consequence.

Regarding cause 3, it may happen that if the true labels related to a unique alarm are strongly unbalanced (i.e., mostly "Y" or "N"), the model will deduce that all the events produced by that particular unique alarm must be labeled as "Y" or "N", depending on which is the most frequent. For instance, this behavior has been observed for both the unique alarms "LI315 IOP" and "TI542 IOP": the first produced a total of 18 alarm events and only 5 of them were "Not Chattering within one hour", the latter produced only 4 "Chattering within one hour" events out of 38 in total. As a consequence, the algorithm has predicted that all the events produced by "LI315 IOP" must be labeled as "Y", and events produced by "TI542 IOP" must be labeled as "N".

Poor data distribution, as well as the use of too small datasets, may play a crucial role in causing the issues described above. For this reason, it might be worthwhile to consider a more extensive database for further analyses. Besides, different sets of features should be tested to resolve the misidentification of chattering sequences boundaries.

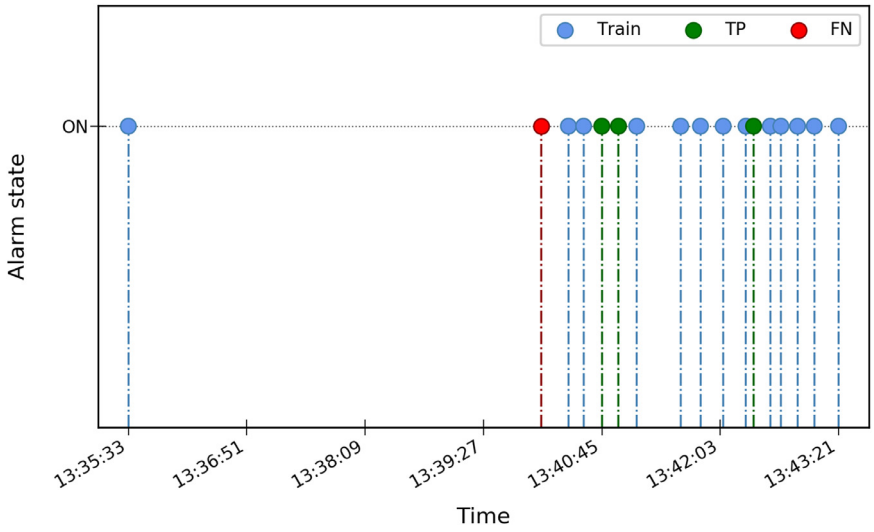


Fig. 11. Detail of a Chattering sequence produced by the unique alarm FI234 LTRP. Colored dots represent alarm events (alarm state = "ON"). True label is "Y" for all the events in the figure. Blue dots refer to alarm events included in the Training database, other colors refer to events included in the Evaluation database. Red dots indicate a wrong prediction (a False Negative), green dots indicate a correct prediction (a True Positive). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

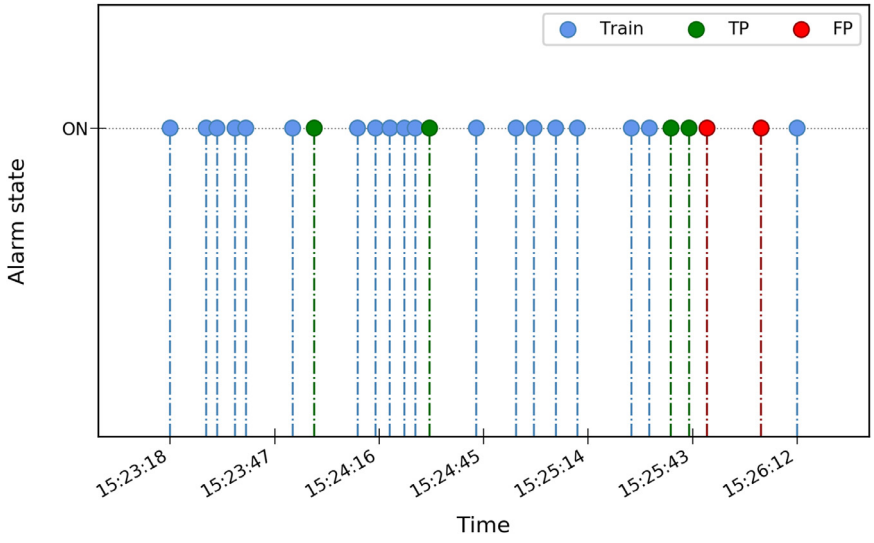


Fig. 12. Detail of a Chattering sequence produced by the unique alarm FI234 LTRP. Colored dots represent alarm events (alarm state = "ON"). True labels of the last three elements are "N", other events have "Y" as True labels. Blue dots refer to alarm events included in the Training database, other colors refer to events included in the Evaluation database. Red dots indicate a wrong prediction (a False Positive), green dots indicate a correct prediction (a True Positive). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.).

8. Conclusions

A Machine Learning method for chattering prediction was developed. Analyses have involved the formulation of the Dynamic Chattering Index to perform a preliminary classification of historical alarm data. This new index overcomes the limitations of the Chattering Index, providing more flexible and dynamic results, which can be used to link actual process conditions to future alarm behavior.

Three different Machine Learning models –Linear, Deep, and Wide&Deep– have been trained and evaluated. The models have been tested on the ability to predict future chattering behavior based on actual process conditions. The performance metrics and the P-R curves indicate robustness and good prediction capability of the models. The method may be used to build an online tool for chattering prediction and decision-making support. For instance, the algorithm could provide early warnings of possible chattering, and actions might be taken by the operator to avoid this event. Consequently, the workload would be reduced, and the risk of alarm floods would be minimized. In general, the approach demonstrates that advanced analysis techniques can be used to extract knowledge from historical data and perform accurate predictions. A data-driven approach for process monitoring and control appears to be a valuable and interesting opportunity to exploit process data and increase process safety and stability.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

CRediT authorship contribution statement

Nicola Tamascelli: Conceptualization, Methodology, Software, Validation, Formal analysis, Writing - original draft, Writing - review & editing, Visualization. **Nicola Paltrinieri:** Conceptualization, Methodology, Software, Data curation, Supervision, Project administration, Funding acquisition. **Valerio Cozzani:** Supervision, Project administration, Funding acquisition.

Acknowledgments

The authors would like to gratefully acknowledge Yara International for the support and the data provided.

References

Ahmed, K., Izadi, I., Chen, T., Joe, D., Burton, T., 2013. Similarity analysis of industrial alarm flood data. IEEE Trans. Autom. Sci. Eng. 10, 452–457. https://doi.org/10.1109/TASE.2012.2230627.

Aika, K., Christiansen, L.J., Dybkjaer, I., Hansen, J.B., Nielsen, P.E.H., Nielsen, A., Stoltze, P., Tamaru, K., 1995. Ammonia. Springer Berlin Heidelberg, Berlin, Heidelberg https://doi.org/10.1007/978-3-642-79197-0.

Aleixandre, J., Alvarez, I., García, M., Lizama, V., 2015. Application of multivariate regression methods to predict sensory quality of red wines. Czech J. Food Sci. 33, 217–227. https://doi.org/10.17221/370/2014-CJFS.

ANSI/ISA, 2016. ANSI/ISA–18.2–2016 management of alarm systems for the process industries. ANSI/ISA.

Beebe, D., Ferrer, S., Logerot, D., 2012. Alarm floods and plant incidents 17.

Brink, H., Richards, J., Fetherolf, M., 2016. Real-World Machine Learning, first ed. Manning Publications, Shelter Island.

Chai, K.M.A., 2005. Expectation of f-measures: tractable exact computation and some empirical observations of its properties. SIGIR 2005 Proc. 28th Annu. Int. ACM SIGIR Conf. Res. Dev. Inf. Retr., pp. 593–594.

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., 2016. Wide & deep learning for recommender systems,. In: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, pp. 7–10.

Directive 2012/18/EU of the European Parliament and of the Council of 2012 on the control of major-accident hazards involving dangerous substances, OJ L 197, 2012. https://doi.org/10.3000/19770677.L_2012.197.eng.

EEMUA, 2013. EEMUA Publication 191 alarm systems - a guide to design, management and procurement.

Ge, Z., Song, Z., Ding, S.X., Huang, B., 2017. Data mining and analytics in the process industry: the role of machine learning. IEEE Access 5, 20590–20616. https://doi.org/10.1109/ACCESS.2017.2756872.

Giammarco, G., Giammarco, P., 1973. Process for eliminating CO2 and/or H2S from gaseous mixtures. 3725592.

Google, 2020a. Classification: precision and recall [WWW Document]. URL https://developers.google.com/machine-learning/crash-course/classification/precision-and-recall (accessed 1.24.20).

Google, 2020b. Feature crosses [WWW Document]. URL https://developers.google.com/machine-learning/crash-course/feature-crosses/video-lecture (accessed 1.24.20).

Google, 2020c. Classification: accuracy [WWW Document]. URL https://developers.google.com/machine-learning/crash-course/classification/accuracy (accessed 1.24.20).

Google, 2020d. Classification: thresholding [WWW Document]. URL https://developers.google.com/machine-learning/crash-course/classification/thresholding (accessed 6.15.20).

Han, J., Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques, Data Mining: Concepts and Techniques. Elsevier Inc https://doi.org/10.1016/C2009-0-61819-5.

Hastie, T., Friedman, R., Tibshirani, J., 2009. The Elements of Statistical Learning. Springer-Verlag, New York https://doi.org/10.1007/978-0-387-84858-7.

Health and Safety Executive, 1997. The Explosion and Fires at the Texaco Refinery, Milford Haven, 24 July 1994. HSE Books, Incident Report Series.

James, G., Hastie, T., Tibshirani, R., Witten, D., 2013. An Introduction to Statistical Learning: with Applications in R. Springer-Verlag, New York https://doi.org/10.1007/978-1-4614-7138-7.

Jennings, J.R., 1991. Catalytic Ammonia Synthesis. Springer Science & Business Media https://doi.org/10.1007/978-1-4757-9592-9.

Katzel, J., 2007. Control engineering | Managing alarms [WWW Document]. URL www.controleng.com/articles/managing-alarms (accessed 1.23.20).

Kondaveeti, S.R., Izadi, I., Shah, S.L., Black, T., 2010. Graphical representation of industrial alarm data. IFAC Proc. Vol. 11, 181–186. https://doi.org/10.3182/20100831-4-fr-2021.00033.

Kondaveeti, S.R., Izadi, I., Shah, S.L., Shook, D.S., Kadali, R., Chen, T., 2013. Quantification of alarm chatter based on run length distributions. Chem. Eng. Res. Des. 91, 2550–2558. https://doi.org/10.1016/j.cherd.2013.02.028.

Kordic, S., Lam, C.P., Xiao, J., Li, H., 2010. Patterns Relevant to the Temporal Data–Context of an Alarm of Interest, in. In: Dynamic and Advanced Data Mining for Progressing Technological Development. IGI Global, pp. 18–39.

Laberge, J.C., Bullemer, P., Tolsma, M., Reising, D.V.C., 2014. Addressing alarm flood situations in the process industries through alarm summary display design and alarm response strategy. Int. J. Ind. Ergon. 44, 395–406. https://doi.org/10.1016/j.ergon.2013.11.008.

Ling, C.X., Sheng, V.S., 2008. Cost-sensitive learning and the class imbalance problem. Encycl. Mach. Learn. 231–235 https://doi.org/10.1.1.15.7095.

Liu, J., Kong, X., Xia, F., Bai, X., Wang, L., Qing, Q., Lee, I., 2018. Artificial intelligence in the 21st century. IEEE Access 6, 34403–34421. https://doi.org/10.1109/ACCESS.2018.2819688.

Mahadevan, S., Shah, S.L., 2009. Fault detection and diagnosis in process data using one-class support vector machines. J. Process Control 19, 1627–1639. https://doi.org/10.1016/j.jprocont.2009.07.011.

Miao, A., Ge, Z., Song, Z., Zhou, L., 2013. Time neighborhood preserving embedding model and its application for fault detection. Ind. Eng. Chem. Res. 52, 13717–13729. https://doi.org/10.1021/ie400854f.

Mohri, M., Rostamizadeh, A., Talwalkar, A., 2012. Foundations of Machine Learning. Adaptive Computation and Machine Learning series, first ed MIT Press, Cambridge.

Paltrinieri, N., Comfort, L., Reniers, G., 2019. Learning about risk: Machine Learning for risk assessment. Saf. Sci. 118, 475–486. https://doi.org/10.1016/j.ssci.2019.06.001.

Paltrinieri, N., Patriarca, R., Stefana, E., Brocal, F., Reniers, G., 2020. Meta-learning for safety management. Chem. Eng. Trans. 82. DOI https://doi.org/10.3303/CET2082029.

Python.org, 2019. Python Release Python 3.7.4 | Python.org [WWW Document]. URL https://www.python.org/downloads/release/python-374/ (accessed 4.23.20).

Reis, M.S., Kenett, R., 2018. Assessing the value of information of data-centric activities in the chemical processing industry 4.0. AIChE J. 64, 3868–3881. https://doi.org/10.1002/aic.16203.

Shaw, J.A., 1993. DCS-based alarms: integrating traditional functions into modern technology. ISA Trans. 32, 177–181. https://doi.org/10.1016/0019-0578(93)90039-Y.

Sheng, V.S., Ling, C.X., 2006. Thresholding for making classifiers cost-sensitive. Proc. Natl. Conf. Artif. Intell. 1, 476–481.

Stanton, N.A., Barber, C., 1995. Alarm-initiated activities: an analysis of alarm handlingby operators using text-based alarm systems in supervisory control systems. Ergonomics 38, 2414–2431. https://doi.org/10.1080/00140139508925276.

Tamascelli, N., Arslan, T., Shah, S.L., Paltrinieri, N., Cozzani, V., 2020. A Machine Learning Approach to Predict Chattering Alarms. Chem. Eng. Trans. 82. DOI https://doi.org/10.3303/CET2082032 (accessed 1.24.20).

TensorFlow.org, 2020a. Models and layers | TensorFlow.js [WWW Document]. URL https://www.tensorflow.org/js/guide/models_and_layers (accessed 1.24.20).

TensorFlow.org, 2020b. tf.keras.optimizers.Ftrl | TensorFlow Core v2.1.0 [WWW Document]. URL https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Ftrl (accessed 4.25.20).

TensorFlow.org, 2020c. tf.keras.optimizers.Adagrad | TensorFlow Core v2.1.0 [WWW Document]. URL https://www.tensorflow.org/api_docs/python/tf/keras/optimizers/Adagrad (accessed 4.25.20).

TensorFlow.org, 2020d. tf.nn.relu | TensorFlow Core v2.1.0 [WWW Document]. URL https://www.tensorflow.org/api_docs/python/tf/nn/relu (accessed 4.23.20).

TensorFlow.org, 2020e. Tensorflow [www document]. Url https://www.tensorflow.org/ (accessed 4.23.20).

TensorFlow.org, 2020f. tf.feature_column.categorical_column_with_vocabulary_list [WWW Document]. URL https://www.tensorflow.org/api_docs/python/tf/feature_column/categorical_column_with_vocabulary_list (accessed 4.23.20).

TensorFlow.org, 2020g. tf.feature_column.indicator_column | TensorFlow Core v2.1.0 [WWW Document]. URL https://www.tensorflow.org/api_docs/python/tf/feature_column/indicator_column (accessed 4.23.20).

Weiss, J., 2010. Protecting Industrial Control Systems from Electronic Threats, first ed. Momentum Press, New York.

Zhang, Y., Teng, Y., Zhang, Y., 2010. Complex process quality prediction using modified kernel partial least squares. Chem. Eng. Sci. 65, 2153–2158. https://doi.org/10.1016/j.ces.2009.12.010.

Zhong, S., Wen, Q., Ge, Z., 2014. Semi-supervised Fisher discriminant analysis model for fault classification in industrial processes. Chemom. Intell. Lab. Syst. 138, 203–211. https://doi.org/10.1016/j.chemolab.2014.08.008.

Article VIII.

Tamascelli, N., Scarponi, G.E., Paltrinieri, N., Cozzani, V., (2021). **A data-driven approach to improve control room operators' response**. Chem. Eng. Trans. 86, 757–762. https://doi.org/10.3303/CET2186127.

CHEMICAL ENGINEERING TRANSACTIONS
VOL. 86, 2021
A publication of AIDIC
The Italian Association of Chemical Engineering
Online at www.cetjournal.it
Guest Editors: Sauro Pierucci, Jiří Jaromír Klemeš

ISBN 978-88-95608-84-6; **ISSN** 2283-9216
DOI: 10.3303/CET2186127

A Data-driven Approach to Improve Control Room Operators' Response

Nicola Tamascelli[a,b,*], Giordano Scarponi[a], Nicola Paltrinieri[b], Valerio Cozzani[a]

[a]Department of Civil, Chemical, Environmental and Materials Engineering, University of Bologna, Bologna, Italy
[b]Department of Mechanical and Industrial Engineering, NTNU, Trondheim, Norway
nicola.tamascelli2@unibo.it

Digitalization has significantly improved productivity and efficiency within the chemical industry. Distributed Control Systems and extensive use of sensor networks enable advanced control strategies and increase optimization opportunities. On the other hand, chemical plants are increasingly complex, equipment is highly interlinked, and it is more difficult to describe the system dynamics through first principles. Finding the root causes of process upsets and predicting dangerous deviations in process conditions is often challenging. Advanced and dynamic tools are needed to grant safe and stable operations in such a complex and multivariate environment. In this context, Machine Learning techniques may be used to exploit and retrieve knowledge from the large amount of data that chemical plants produce and store on a daily basis. Data-driven methods may be adopted to develop predictive models and support a proactive approach to process safety. The study aims to develop Machine Learning techniques to improve the response of control room operators during critical events. Specifically, alarm data originated in an upper-tier Seveso site have been collected, cleaned, and analyzed to identify periods of intense alarm activity. Alarm behavior following operator responses has been evaluated to assess whether the actions were adequate to prevent future alarm occurrences. In doing so, alarm events that reoccur within 30 minutes after an operator acknowledgment have been identified and labeled. Subsequently, a hybrid classification algorithm was trained to predict the probability that a critical alarm reoccurs after being acknowledged by the operator. This predictive tool might be used to support the operator's decision-making process and focus his/her attention on critical alarms that are more likely to occur again in the near future.

1. Introduction

The alarm system is one of the first layers of protection to prevent process deviations from escalating into accidents (Stauffer and Clarke, 2016). Still, there are inherent difficulties in designing, operating, and maintaining an efficient alarm system (Goel et al., 2017), which includes both technical (e.g., sensors, DCS, actuators) and human functions (e.g., operators). Alarms inform control room operators about dangerous deviation from normal operating conditions so that appropriate corrective actions could be taken. On the other side, operators should be provided with enough time to detect the issue, diagnose the situation, and determine/implement corrective actions (ANSI/ISA, 2016). Still, manual intervention by operators is subject to human error; improper procedures, worker fatigue, and lack of operator training may prevent an adequate response (Exida, 2009). In fact, several accident reports have highlighted that improper alarm management and inaccurate operator actions play a significant role in the development of process accidents. For example, poor alarm prioritization and an excessive alarm annunciation rate contributed to the Texaco Milford Haven explosion, where 26 workers were injured (Health and Safety Executive, 1997). Also, the non-detection of a loss of coolant led to the Three Mile Island Accident (United States Nuclear Regulatory Commission, 2018), where alarms rang, warning light flashed, but no operator could diagnose the situation. In an attempt to rationalize and provide a methodology for a more effective design and management of alarm systems, standard manuals have been published, such as ISA 18.2 (ANSI/ISA, 2016) and EEMUA 191 (EEMUA, 2013). Still, much remains to be done (Goel et al., 2017). The advent of the Third Industrial Revolution has already

Paper Received: 28 August 2020; Revised: 7 February 2021; Accepted: 26 April 2021
Please cite this article as: Tamascelli N., Scarponi G.E., Paltrinieri N., Cozzani V., 2021, A Data-driven Approach to Improve Control Room Operators' Response, Chemical Engineering Transactions, 86, 757-762 DOI:10.3303/CET2186127

brought changes and improvements to the industry. Chemical plants are more productive, more automated, more flexible, and safety systems are more advanced. Nevertheless, some issues have arisen as well. Modern DCS allows the configuration of new alarms with few clicks of mouse (Katzel, 2007). As a result, a larger number of alarms are installed, and the workload for operators has increased to the point where they are often overwhelmed by nuisance alarms (Kondaveeti et al., 2013). In addition, chemical plant complexity has increased, and control/safety functions are more intricate. As complexity increases, failures are more likely to occur, and root causes of process upsets are more difficult to detect (Wall, 2009). Therefore, it may be challenging for control room operators to find the appropriate set of corrective actions in a reasonable time. An intelligent tool to assess and predict the effectiveness of operators' actions would be of great support in dealing with critical situations. In this context, advancements in IT, IoT, and computer science have led to the development of intelligent computer-based algorithms to extract knowledge from data and support knowledge-based decision-making. In fact, a massive amount of process and alarm data are produced and stored on a daily basis (Reis and Kenett, 2018). Machine Learning algorithms may be used to "mine" these data and create predictive models for, e.g., fault detection e fault diagnosis (Tian et al., 2015), predictive maintenance (Carvalho et al., 2019), Dynamic Risk Assessment (Paltrinieri et al., 2019), modeling and simulation (Aleixandre et al., 2015).
This work focuses on the application of Machine Learning techniques for predicting the probability that a critical alarm reoccurs after being acknowledged by an operator. In this way, the operator's attention would be driven to alarms that are more likely to occur again in a short time. Alarm data from an ammonia production site have been used to evaluate the proposed methodology.

2. Alarm database: structure and features

Alarm records originated in an ammonia production plant have been used to test the proposed methodology. Alarms that occurred between July and November 2017 have been extracted and stored in a CSV file (i.e., the alarm database), which contains 26,473 alarm records described by means of 39 different attributes. That is, the database may be considered a 26,473x39 matrix, where each row represents an alarm event, and each column represents an attribute (i.e., a feature) of an alarm event. A reduced version of the database was described and used by Tamascelli et al. (2020a, 2020b). In the present study, the whole dataset has been used.Most of the alarms in the database occurred between 09/09/2017 and 09/10/2017, when a total power outage forced an emergency plant shutdown. During this critical event, the alarm annunciation rate often exceeded 1000 alarms/day, and the workload for control room operators increased drastically.
Each alarm event is described by a list of attributes. A comprehensive description of the attributes may be found in Tamascelli et al. (2020b). However, only three attributes are needed for uniquely identifying an alarm event:

1. Timestamp;
2. Source;
3. Identifier.

The Timestamp is the date and time of the alarm occurrence. The Source represents the instrument or logic function that generated the alarm. An example of a Source is LI315 (i.e., the level indicator in the control loop 315). The Identifier indicates the alarm status. Nine different identifiers are found in the database, as shown in Table 1.

Table 1: Alarm Identifiers

Identifier	Meaning
HHH	The measured variable has exceeded the high alarm setpoint
HTRP	The measured variable has exceeded the very-high alarm setpoint
LLL	The measured variable has exceeded the low alarm setpoint
LTRP	The measured variable has exceeded the very-low alarm setpoint
ALM	Generic alarm
IOP	Instrument failure or out-of-range measure
ACK	The operator has acknowledged the alarm
NR	A generic alarm is terminated (it refers to an earlier ALM alarm)
Recover	Alarm terminated (it refers to an earlier HHH, HTRP, LLL, LTRP, or IOP alarm)

In addition to alarms per se, the database keeps track of two different events: the acknowledgment of an alarm by an operator and the recovery of an alarm. The former is described by the Indicator “ACK”, the latter by “Recover” or “NR” depending on the Identifier of the original alarm.

3. Methodology

The method follows three main steps: Data pre-processing, Target identification, and Machine Learning simulations.

3.1 Data pre-processing

Alarm attributes (i.e., columns of the database) that have not been deemed relevant for the analyses have been discarded. For example, empty columns have been removed, as well as columns that show the same value for each event in the database. Also, redundant attributes have been removed.
Next, missing values have been substituted by the value “0”. This has been done because most Machine Learning algorithms do not tolerate missing values (Brink et al., 2016). The choice of the value “0” is arbitrary; a different numerical or categorical value (i.e., a text string) would be equally effective (Han et al., 2012). In doing so, one must ensure that the chosen value is outside of the domain of the attribute affected by missing values (i.e., the attribute should never take values equal to the one selected for the substitution).
Finally, it may happen in industrial databases that different measurements have different units. For example, one may find that some of the pressure measurements are expressed in “bar”, while others in “atm”. Whenever this happens, it is critical to ensure that attribute values referring to homogeneous physical quantities are expressed into common measurement units. Also, numerical values should be normalized in order to suppress scale effects (e.g., using min-max scaling) (Brink et al., 2016).

3.2 Target identification

The database must be analyzed to find and highlight events where an operator has acknowledged an alarm, but still, another alarm from the same Source occurs within 30 minutes. Events that meet this criterion are called Target events. The time window has been selected in accordance with the approach mentioned in the PETRO-HRA Guideline, which evaluates the 30 minutes criterion as the time required for action from the operator (Stauffer and Clarke, 2016). A binary categorical variable is assigned to each event in the database to highlight Targets. The binary variable is called the Label of an alarm event, and it assumes the values “YES” or “NO” depending on whether the event is a Target or not. Therefore, if the database contains n events, n Labels are generated, stored in a vector, and appended to the alarm database. Table 2 clarifies the role of Labels.

Table 2: Fictitious alarm sequence from LI315.

Timestamp	Source	Identifier	Attribute 4	. . .	Attribute n	Label
01/01/2021 00:00:00	LI315	LLL	---	. . .	---	NO
01/01/2021 00:03:00	LI315	LLL ACK	---	. . .	---	NO
01/01/2021 00:19:00	LI315	LLL Recover	---	. . .	---	NO
01/01/2021 01:30:00	LI315	LLL	---	. . .	---	NO
01/01/2021 01:15:00	LI315	LLL ACK	---	. . .	---	YES
01/01/2021 01:40:00	LI315	LTRP	---	. . .	---	NO

The table shows a fictitious alarm sequence from LI315. Data are organized as described in section 2, except for the last column, which contains the Labels. The second-last event of the series has “YES” as a label since another alarm from the same Source (LTRP) has occurred less than 30 minutes after acknowledging the previous low-level alarm (LLL). On the contrary, the first ACK event in Table 2 has “NO” as a label because the alarm has been recovered after 16 minutes (LLL Recover).

3.3 Machine Learning simulations

A Wide&Deep classification model has been trained and evaluated on the alarm database. The purpose of the algorithm is to classify alarm events into two categories: Target (i.e., Label = “YES”) and Not Target (i.e., Label = “NO”). That is, the model would predict whether an acknowledgment will be followed by another alarm from the same Source (Label = “YES”) or not (Label = “NO”). The workflow to set up and perform the Machine Learning simulations is illustrated in Figure 1. Two steps must be followed to complete a classification task: training and evaluation. During training, ⅔ of the alarm database is fed to the Wide&Deep model (arrow T1 in Figure 1), which “learns” the relationship between the features of an event (i.e., the attributes) and its Label.

The process involves the joint optimization of two distinct models (Cheng et al., 2016): a Linear model and a Deep Neural Network. The structure of a Wide&Deep model is illustrated in Figure 1. Mathematics and technical details behind the model are out of the scope of this work and may be found in Cheng et al. (2016).

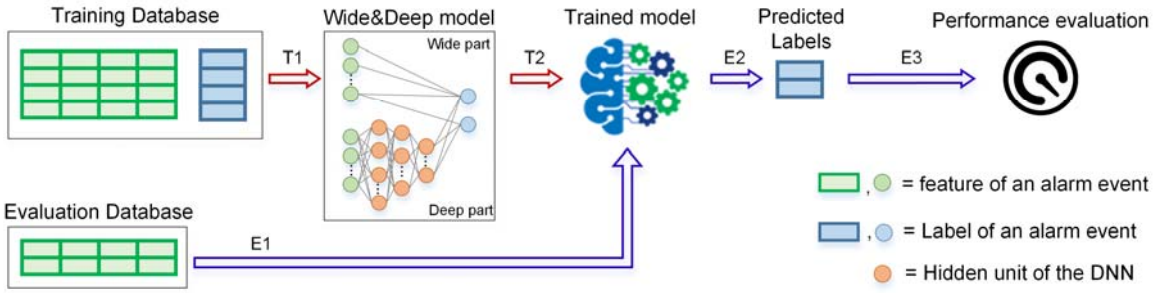


Figure 1: Training and evaluation of the Wide&Deep model.

In general, the algorithm aims at optimizing the internal parameters of a function f to best represent the relationship between features of an event (X), and its Label (Y):

$$f(X) \approx Y \tag{1}$$

The function f in Eq(1) comprises a linear part, where features are linearly combined and mapped into Labels, and a Deep part, where features are linearly combined and transformed into derived features (i.e., hidden units or "neurons" of the DNN) through nonlinear transformations. The parameters used to set up the model are the same as those used by Tamascelli et al. (2020b). After training, a trained model is obtained (T2 in Figure 1). Next, the model is evaluated. Labels are removed from the rest of the database, which is fed to the trained model (E1 in Figure 1). The algorithm takes as an input the features of each event and returns the probability of the Label being "YES" or "NO" (E2 in Figure 1). By default, a probability decision threshold equal to 0.5 is used to convert Label probabilities into Labels (i.e., if the probability of Label "YES" is greater than 0.5, the event will be labeled as "YES"). Finally, predicted Labels are compared with true Labels to assess the model performance.

4. Results

The target identification procedure (step 3.2) highlighted that a total of 119 events meet the requirements to be classified as Target. The training database comprises 17,649 alarm events, of which 78 belong to the Target category. The evaluation database contains 8824 events, of which 41 belong to the Target category.
After the evaluation phase, three metrics have been calculated in order to assess the performance of the model:

$$\text{Accuracy} = \frac{TP+TN}{TP+TN+FP+FN} = \mathbf{0.995} \tag{2}$$

$$\text{Precision} = \frac{TP}{TP+FP} = \mathbf{0.5} \tag{3}$$

$$\text{Recall} = \frac{TP}{TP+FN} = \mathbf{0.049} \tag{4}$$

Where TP identifies a True Positive (i.e., the model predicted the label "YES", and the true Label of the event is "YES"), a TN identifies a True Negative (i.e., predicted Label = "NO", real Label = "NO"), FP identifies a False Positive (predicted Label = "YES", real Label = "NO"), and FN identifies a False Negative (i.e., predicted Label = "NO", real Label = "YES"). Therefore, the sum of TP and TN indicates the number of correct predictions, while FN and FP show the number of wrong predictions.

5. Discussion

Metrics presented in Eq(2), Eq(3), and Eq(4) indicates that 99.5 % of the prediction were correct. Nevertheless, this result does not imply satisfactory performance. In fact, only 41 out of 8824 events in the evaluation database have "YES" as a true Label. Therefore, the model would have achieved an Accuracy greater than 99 % by always predicting the Label "NO". This happens because de dataset used for the simulation is imbalanced, meaning that there are only a few examples of Target events within the database. In

this situation, Precision and Recall offer more information than Accuracy. Furthermore, it should be considered that mislabelling a Target event as a non-Target one is more critical than labeling a “NO” event as “YES”. Thus, Recall is the most meaningful metric to consider in this specific task. A large precision is desirable but not as important as a high Recall. Precision in Eq(2) indicates that 50 % of the “YES” predictions were correct, and the Recall in Eq(4) shows that only 4.9 % of the Target event were correctly identified (i.e., 2 out of 41). Evidently, the rarity of Target events must have affected the learning process and contributed to the uncertainty of predictions. However, these metrics have been calculated using a standard probability decision threshold equal to 0.5. There is no guarantee that this value will lead to the best performance. Thus, the decision threshold has been varied from 0 to 1; every time the threshold has been changed, predicted labels are calculated again, and so are Precision and Recall. Figure 2 illustrates how Precision and Recall change with the decision threshold. Lowering the threshold to 0.012 would increase the Recall from 0.049 to 0.9 and decrease the Precision from 0.5 to 0.34. As previously mentioned, Recall is the most important metric to consider in this problem. Thus, the performance would significantly improve since the increase in Recall is five times larger than the decrease in Precision.

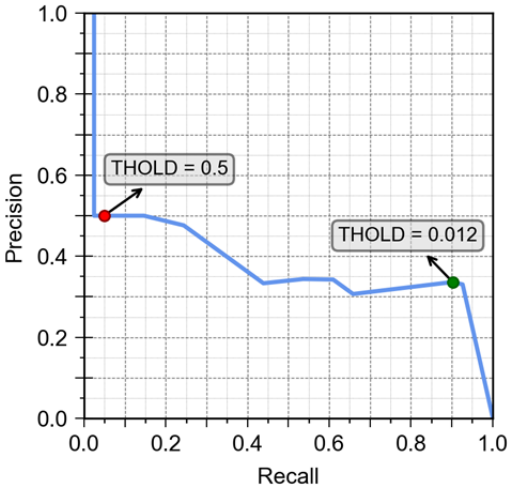


Figure 2: Precision–Recall curve produced by the Machine Learning simulation. Coordinates of points on the curve represent Precision and Recall obtained using different decision thresholds (THOLDs). The Red mark represents the point at threshold = 0.5. The green mark represents the point at threshold = 0.012.

It is worth noting that 0.012 is a relatively small threshold. However, most of the predicted probabilities are lower than 0.012. More than 80 % of the prediction probabilities are smaller than 0.1, and only 110 events over 8824 have a probability larger than 0.012. This suggests that the model is relatively unconfident, which may be due to the rarity of the event considered. Still, 90 % of the Target event lies within those 110 events, which is an encouraging result considering the size of the dataset. In this situation, lowering the threshold to such a low value seems an acceptable compromise considering how probabilities are distributed. Future works should investigate whether training the algorithm with more alarm data would partially overcome the issues related to the rarity of Target events and eventually improve the model confidence. Additional tests should also be performed to assess whether different sets of features or different Machine Learning models would be better suited for the problem under assessment. Moreover, it is worth stressing that the analyses rely entirely on historical alarm data. Further tests are needed to assess the algorithm performance in a real environment. For example, the model may be integrated into the plant DCS in order to analyze live streams of alarm data; this would allow evaluating the model effectiveness and highlighting its possible limitations.

6. Conclusions

This work proposes a data-driven method to extract knowledge from historical alarm data and perform predictions on the effectiveness of control room operators’ actions. A real industrial database has been used to support the analyses. A Wide&Deep classification model has been trained and evaluated on the database. The model aims at predicting whether or not the operator’s acknowledgment of an alarm will be followed by another alarm from the same Source within 30 minutes. In this way, the model would indirectly predict the effectiveness of the operator’s action and eventually drive his/her attention to alarms that are more likely to occur again in a short time. The issues related to the identification of rare unwanted events (such as those

considered in this work) have been discussed. Results show that even if performance may seem inadequate, a high Recall value may be obtained by lowering the decision threshold. After this simple adjustment, the model performance has improved considerably, and more than 90 % of the Target events have been correctly identified. Further investigations should be performed to evaluate the viability of the study in real-time applications. However, the approach suggests that Machine Learning may be used to extract relevant information from historical alarm data and use the acquired knowledge to support the control room operators proactively.

Acknowledgments

The authors would like to acknowledge Yara International for supplying the data and for the valuable support. Also, we wish to extend our thanks to Davide Santini, whose work has contributed significantly to the analyses described in this paper.

References

Aleixandre, J., Alvarez, I., García, M., Lizama, V., 2015. Application of Multivariate Regression Methods to Predict Sensory Quality of Red Wines. Czech Journal of Food Sciences 33, 217–227. https://doi.org/10.17221/370/2014-CJFS

ANSI/ISA, 2016. ANSI/ISA–18.2–2016 Management of Alarm Systems for the Process Industries. ANSI/ISA.

Brink, H., Richards, J., Fetherolf, M., 2016. Real-World Machine Learning, first. ed. Manning Publications, Shelter Island.

Carvalho, T.P., Soares, F.A.A.M.N., Vita, R., Francisco, R. da P., Basto, J.P., Alcalá, S.G.S., 2019. A systematic literature review of machine learning methods applied to predictive maintenance. Computers and Industrial Engineering 137, 106024. https://doi.org/10.1016/j.cie.2019.106024

Cheng, H.-T., Koc, L., Harmsen, J., Shaked, T., Chandra, T., Aradhye, H., Anderson, G., Corrado, G., Chai, W., Ispir, M., 2016. Wide & deep learning for recommender systems, in: Proceedings of the 1st Workshop on Deep Learning for Recommender Systems. pp. 7–10.

EEMUA, 2013. EEMUA Publication 191 Alarm systems - a guide to design, management and procurement.

Exida, 2009. Saved by the Bell : Using Alarm Management to make Your Plant Safer. Sellersville, PA.

Goel, P., Datta, A., Mannan, M.S., 2017. Industrial alarm systems: Challenges and opportunities. Journal of Loss Prevention in the Process Industries 50, 23–36. https://doi.org/10.1016/j.jlp.2017.09.001

Han, J., Kamber, M., Pei, J., 2012. Data Mining: Concepts and Techniques, Data Mining: Concepts and Techniques. Elsevier Inc. https://doi.org/10.1016/C2009-0-61819-5

Health and Safety Executive, 1997. The Explosion and Fires at the Texaco Refinery, Milford Haven, 24 July 1994: A Report of the Investigation by the Health and Safety Executive Into the Explosion and Fires on the Pembroke Cracking Company Plant at the Texaco Refinery, Milford Haven on 24 J, Incident Report Series. HSE Books.

Katzel, J., 2007. Control Engineering | Managing Alarms [WWW Document]. URL www.controleng.com/articles/managing-alarms (accessed 1.23.20).

Kondaveeti, S.R., Izadi, I., Shah, S.L., Shook, D.S., Kadali, R., Chen, T., 2013. Quantification of alarm chatter based on run length distributions. Chemical Engineering Research and Design 91, 2550–2558. https://doi.org/10.1016/j.cherd.2013.02.028

Paltrinieri, N., Comfort, L., Reniers, G., 2019. Learning about risk: Machine learning for risk assessment. Safety Science 118, 475–486. https://doi.org/10.1016/j.ssci.2019.06.001

Reis, M.S., Kenett, R., 2018. Assessing the value of information of data-centric activities in the chemical processing industry 4.0. AIChE Journal 64, 3868–3881. https://doi.org/10.1002/aic.16203

Stauffer, T., Clarke, P., 2016. Using alarms as a layer of protection. Process Safety Progress 35, 76–83. https://doi.org/10.1002/prs.11739

Tamascelli, N., Arslan, T., Shah, S.L., Paltrinieri, N., Cozzani, V., 2020a. A Machine Learning Approach to Predict Chattering Alarms. Chemical Engineering Transactions 82. https://doi.org/10.3303/CET2082032

Tamascelli, N., Paltrinieri, N., Cozzani, V., 2020b. Predicting Chattering Alarms: a Machine Learning Approach. Computers & Chemical Engineering 107122. https://doi.org/https://doi.org/10.1016/j.compchemeng.2020.107122

Tian, Y., Fu, M., Wu, F., 2015. Steel plates fault diagnosis on the basis of support vector machines. Neurocomputing 151, 296–303. https://doi.org/10.1016/j.neucom.2014.09.036

United States Nuclear Regulatory Commission, 2018. Backgrounder on the Three Mile Island Accident, United States Nuclear Regulatory Commission Library.

Wall, K., 2009. Complexity of chemical products, plants, processes and control systems. Chemical Engineering Research and Design 87, 1430–1437. https://doi.org/10.1016/j.cherd.2009.03.007

Article IX.

Tamascelli, N., Rao, H. R. M., Cozzani, V., Paltrinieri, N., Chen, T. (2023). **Online Classification of Alarm Floods Using a Word2vec Algorithm**. IECON 2023 – 49th Annual Conference of the IEEE Industrial Electronics Society, Singapore, 2023. https://doi.org/10.1109/IECON51785.2023.10312435.

Online Classification of Alarm Floods Using a Word2vec Algorithm

Nicola Tamascelli*‡[1], Harikrishna Rao Mohan Rao†[1], Valerio Cozzani‡[2], Nicola Paltrinieri*[2], Tongwen Chen†[2]
*Department of Mechanical and Industrial Engineering,
Norwegian University of Science and Technology, Trondheim, Norway
Email: [1]nicola.tamascelli@ntnu.no, [2]nicola.paltrinieri@ntnu.no
†Department of Electrical & Computer Engineering,
University of Alberta, Edmonton, Alberta T6G 1H9, Canada
Email: [1]mohanrao@ualberta.ca, [2]tchen@ualberta.ca
‡Department of Civil, Chemical, Environmental, and Materials Engineering,
University of Bologna, Bologna, Italy
Email: [1]nicola.tamascelli2@unibo.it, [2]valerio.cozzani@unibo.it

IECON 2023- 49th Annual Conference of the IEEE Industrial Electronics Society | 979-8-3503-3182-0/23/$31.00 ©2023 IEEE | DOI: 10.1109/IECON51785.2023.10312435

***Abstract*—Alarm floods are periods of intense alarm activity that may hinder control room operators' ability to diagnose and respond to process abnormalities. In this context, a method to guide and assist operators during alarm floods would provide critical support in preventing abnormalities from escalating into serious accidents. Therefore, this study introduces a novel approach for the online classification of alarm floods based on their fault categories. Historical alarm data are used to train an ensemble of Natural Language Processing models, specifically word2vec, which learn contextual relationships between alarms under different fault conditions. As a new alarm flood appears, the models predict the most probable context alarms by exploiting the knowledge gained during training. Finally, a scoring system is proposed to reward the models that make correct predictions and eventually identify the most probable fault category. The efficacy of the method has been tested on simulated alarm data from the Tennessee Eastman Process benchmark. The results are encouraging, as the models achieved relatively high accuracy in most fault categories.**

***Index Terms*—Alarm Floods, Online Classification, Word2vec.**

I. Introduction

Alarm systems are integral to modern process plants ensuring their safe and efficient operation, necessitated by their increasing complexity and the demanding production requirements [1], [2]. The advances in digital technology have introduced complex monitoring and alerting capabilities, making it convenient to design and configure alarms. However, the ease of adding alarm points has resulted in numerous alarm management problems, including alarm floods - the presence of a large number of alarms beyond what a plant operator can efficiently handle at a time. The industrial standards, ISA [3] and EEMUA [4], define an Alarm Flood (AF) as a period having 10 or more annunciated alarms per 10 minutes per operator and recommend that an operator shall receive no more than 6 alarms/hour.

During AFs, operators may be overwhelmed by the numerous alarms distracting them from addressing critical alarms and ongoing abnormalities, resulting in potentially dangerous situations. AFs have contributed to catastrophic incidents, including the Three Mile Island (1979), Chernobyl disaster (1986), Texaco Refinery (1994), among others. In addition to compromising safety, the presence of AFs can significantly reduce the efficiency and performance of alarm systems. Due to the complex connectivity and interactions, the fault originating at one point can lead to a cascade of alarms. Furthermore, alarm sequences originating from the same fault category are expected to be similar, and analyzing AFs based on the alarms and their sequential order can provide insights into the root causes of the associated abnormalities. However, this task is challenging due to the presence of noise and varying fault conditions, which can lead to mismatches in alarm sequences. Therefore, advanced techniques are needed to effectively analyze AFs through accurate pattern matching and similarity calculation.

Research interest in AF analysis, classification, and prediction has increased over the recent years [5]. Based on the implementation, these methods can be broadly classified into offline and online techniques. Offline techniques identify similar similar AF sequences based on various similarity metrics to provide decision support for operators. Cheng *et al.* modified the Smith-Waterman algorithm to identify similar AF patterns [6]. The computational complexity of the approach in [6] was addressed through a local alignment approach based on the basic local alignment search tool (BLAST) in [7]. The order-ambiguity of alarms in AF sequences was addressed using extended term frequency-inverse document frequency (TF-IDF)-based clustering approaches [8] and a modified PrefixSpan algorithm considering AFs as time-stamped sequences in [9]. Manca *et al.* used dynamic causal dependencies of highly affected process variables for early detection of AFs.

Online alarm flood analysis uses advanced machine learning techniques to identify and classify ongoing alarm floods, enabling early detection of potential root causes of abnormal conditions, and enabling plant operators to take corrective actions before the situation escalates. Various approaches have been

This work was partially supported by the Natural Sciences and Engineering Research Council of Canada.

proposed, such as incremental dynamic programming [10], a binary series classification using Support Vector Machines and k-Nearest Neighbors [11], and Exponentially Attenuated Component Analysis that prioritizes alarms triggering earlier in the alarm flood [12]. Furthermore, operator assistance systems were developed using a Natural Language Processing (NLP) technique, namely, bag-of-words in [13] and real-time pattern matching and alarm ranking approach in [14]. Finally, Wang *et al.* utilized HAZOP analysis to identify abnormal scenarios and built an online model for process monitoring using a Bayesian network of process variables in [15].

Despite the advances in the field of data mining and computational technology, online alarm flood classification remain under-explored in the literature. This can be attributed to the computational complexity of advanced algorithms, which limits their implementation in online settings. Recently, the field of machine learning has made progress in proposing simple and robust Natural Language Processing (NLP) techniques, which are applied to various tasks such as chatbot development, language translation, sentiment analysis, text generation, question answering, and more. For example, the latest release of the GPT (Generative Pre-trained Transformer) series by OpenAI [16], GPT-4 brings a new approach to language models that can provide better results for NLP tasks. Nevertheless, there are still few studies that utilize NLP techniques for the online classification of alarm floods. Motivated by the above problem and the gap in the literature, we propose a novel and computationally efficient approach for the online classification of alarm floods using word2vec, an advanced NLP technique. The main contributions are:

1) The most probable alarms in the ongoing alarm flood are predicted by capturing the contextual relationships between the alarms in different fault conditions.
2) To reduce the computational complexity, a scoring system is utilized to classify the ongoing alarm floods, thereby removing the need for an additional classification or clustering algorithm.

The rest of the paper is organized as follows. Section II presents the detailed steps involved in the online classification of alarm floods. The effectiveness of the proposed method is demonstrated via a case study in Section III, followed by concluding remarks in Section IV.

II. Methodology

Details of the proposed method for the online classification of AFs using the word2vec algorithm are presented in this section, where the approach has two main stages, namely, offline training of the models and online AF predictions.

A framework of the method is provided in Fig. 1, where the steps in offline training of the models are shown in blue and the steps in online AF prediction are shown in green. Specifically, the offline stage involves the preprocessing of alarm data and training of an ensemble of word2vec models using a cluster of similar AFs; whereas, in the online stage, the ongoing AF is analyzed using trained models to predict the most probable

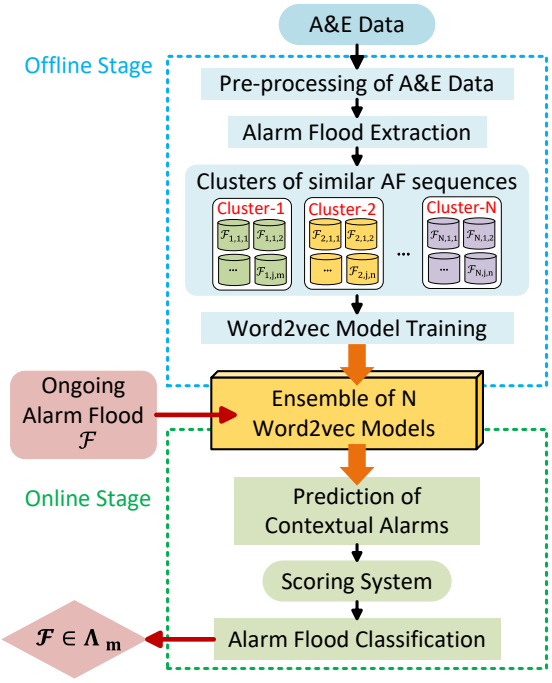


Fig. 1: Framework of the proposed method, consisting of two main stages, namely, offline training of the model (highlighted in blue) and online alarm flood classification (highlighted in green). The algorithm classifies the ongoing alarm flood $\mathcal{F}$ as belonging to a known fault category Λ_m.

alarm and a scoring system is introduced to classify the AF into specific fault categories.

A. Offline Stage - I: Data Pre-processing & AF Extraction

In the offline stage, an ensemble of word2vec models is trained using alarm floods sequences extracted from historical Alarm & Event (A&E) data, where the calculations are performed in three steps, including the pre-processing of A&E data, alarm flood extraction, and model training.

1) Pre-processing of A&E Data: To systematically extract the contextual relationships between alarms, the historical data is pre-processed to obtain an ensemble of word2vec models. An A&E log is a chronologically ordered series of alarm events in textual form, where an alarm event is defined as

$$\mathcal{E} = (a, m, t), \tag{1}$$

where $a \in \mathcal{A}$ is the alarm, $m = \{0, 1\}$ is the status of the alarm at time $t \in \mathcal{T}$. Here, $\mathcal{A}$ represents the set of alarms configured in the plant, and $\mathcal{T}$ is the time duration for which the A&E data was collected. Furthermore, an alarm a is characterized by an alarm tag and identifier as $a = (\alpha, \nu)$, where α is the alarm tag, which contains the information about the area and component to which the alarm is configured, and ν provides the details about the type of alarm. For instance, the alarm "PI100.LL" is a combination of the alarm tag PI100 (indicating Pressure Indicator belonging to control loop number 100) and the identifier "LL" (indicating an analog alarm LowLow). Thereafter, chattering alarms are identified

and discarded because such alarms are typically a result of noise or disturbance in the process.

2) Alarm Flood Extraction: An A&E log may contain sequences of alarms due to multiple process faults and disturbances. As mentioned in Section I, the alarm sequences are considered to be similar if they originated from the same fault category. Therefore, labeled clusters of similar alarm sequences, with pre-identified fault categories are taken as the input to the offline stage. Define an alarm sequence from the A&E log as

$$\mathcal{S} = \langle \mathcal{E}_1, \mathcal{E}_2, \cdots, \mathcal{E}_{|\mathcal{S}|} \rangle, \tag{2}$$

where, $\mathcal{E}_k$ represents the kth alarm event, $k = 1, 2, \cdots |\mathcal{S}|$; the operator $\langle \cdot \rangle$ indicates a sequence; and the operator $|\cdot|$ gives the size of the alarm sequence. If the analysis focuses only on the alarms triggered ($m = 1$) and the time of occurrence is not considered, the alarm sequence can be represented as

$$\mathcal{S} = \langle a_1, a_2, \cdots, a_{|\mathcal{S}|} \rangle, \tag{3}$$

where $\mathcal{S}$ is in the form of strings (textual data). Consider the A&E log consists of alarm sequences from N fault categories, $\Lambda = \{\Lambda_1, \Lambda_2, \cdots, \Lambda_N\}$, and it is assumed that there exists at least one alarm sequence in each fault category, Λ_i. The alarm sequences associated with Λ_i can be grouped into a cluster $\mathcal{C}_i$

$$\mathcal{C}_i = \{\mathcal{S}_{i,1}, \mathcal{S}_{i,2}, \cdots, \mathcal{S}_{i,|\mathcal{C}_i|}\}, \tag{4}$$

where, $\mathcal{S}_{i,j}$ represents the jth alarm sequence in the cluster associated with the Λ_i. Here, $|\mathcal{C}_i| \geq 1$ or $\mathcal{C}_i \neq \emptyset$. Finally, the clusters of alarm sequences from the N fault categories are collected into a set as

$$\mathbb{C} = \{\mathcal{C}_1, \mathcal{C}_2, \cdots, \mathcal{C}_N\}. \tag{5}$$

Thereafter, AF sequences are extracted from these clusters following the definition in [3]. An AF starts when the alarm rate (namely, the number of alarms within a time window Δt) exceeds an upper threshold τ_{max} and ends when the alarm rate drops below a lower threshold τ_{min}. Therefore, a binary indicator σ is defined to differentiate between alarms that belong (σ=1) and do not belong (σ=0) to a flood. The indicator σ of an alarm event $\mathcal{E}_k$ can be defined as

$$\sigma_{(\mathcal{E}_k)} = \begin{cases} 1, & \text{if } \Gamma \geq \tau_{max}, \\ 1, & \text{if } \Gamma \geq \tau_{min} \text{ and } \sigma_{(\mathcal{E}_{k-1})} = 1, \\ 0, & \text{if } \Gamma < \tau_{min}, \end{cases} \tag{6}$$

where Γ denotes the alarm count within $t_k + \Delta t$. The indicator σ is used to extract AFs from each alarm sequence. As a result, clusters of alarm sequences in (5) are converted into clusters of alarm floods as

$$\mathbb{F}_{i,j} = \{\mathcal{F}_{i,j,1}, \mathcal{F}_{i,j,2}, \cdots, \mathcal{F}_{i,j,|\mathbb{F}_{i,j}|}\}, \tag{7}$$

where, $\mathcal{F}_{i,j,k}$ represents the k^{th} flood extracted from $\mathcal{S}_{i,j}$. It is worth noting that an alarm sequence $\mathcal{S}$ may contain more than one AF depending on process dynamics. Afterward, the AF sequences are represented in the form of a list of strings as in (3), by removing the time stamps associated with the alarms to be compatible with the requirements of the word2vec model.

B. Offline Stage - II: Preliminaries & Model Training

Some preliminaries of the word2vec algorithm and the detailed steps for training the model are provided.

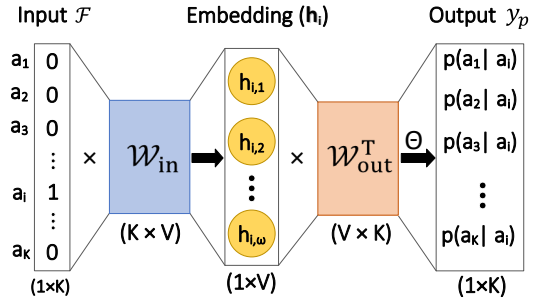


Fig. 2: The framework of Skipgram model, where the input layer is a K dimensional one-hot encoded representation of a_i, the embedding layer $\mathbf{h}_i$ is the vector representation of a_i, and the output layer is the conditional probability $p(a_k|a_i)$, $k = 1, 2, \cdots, K$. $\mathcal{W}_{in}$ and $\mathcal{W}_{out}$ are the internal weights of the model. Θ is the softmax transformation function.

1) Word2vec Algorithm: The proposed work utilizes a widely used NLP model "word2vec", that transforms words in a text into fixed-length vectors (word embeddings) capturing their semantic and syntactic relationships in a high-dimensional vector space [17]. In semantic analysis, this model is used to predict contextual words in textual data, where the context of a word in a sentence is described by the words preceding and succeeding it. In this study, we consider that a cluster of similar AFs is analogous to a collection of topic-specific texts, where alarms are analogous to the words composing sentences. Therefore, the word2vec model is adapted to suit alarm flood applications to predict context alarms, where the context of an alarm in an alarm sequence is defined by the abnormal situation that triggered the alarm (namely, the fault category). Specifically, the context of an alarm a_i in an alarm sequence $\mathcal{S}$ is featured by the alarms occurring within a short temporal vicinity of a_i or in other words, the alarms preceding and succeeding a_i. This study uses the Skipgram architecture of the model to predict context alarms based on an input set of alarms, defined as target alarms [17]. Fig. 2 provides the framework of the Skipgram architecture, where the input to the model is a one-hot encoded representation of an alarm a_i, and the model is a single-layer Neural Network that converts the target alarm a_i into a vector $\mathbf{h}_i$ of cardinality V and generates the output layer of conditional probabilities $p(a_k|a_i)$, $k = 1, 2, \cdots, K$. Here, $K = |\mathcal{A}|$, the number of unique alarms configured in the system. Thereafter, the model is trained to be used for online alarm flood classification.

2) Model Training: Each cluster of AFs is utilized to train a word2vec model as in [17], such that each model learns contextual similarities between alarms based on the fault category associated with the AF. The model is trained using the cluster of AFs to obtain two matrices $\mathcal{W}_{in}, \mathcal{W}_{out} \in \mathbb{R}^{\mathrm{K \times V}}$, representing internal weights. Here, K is the number of unique alarms configured and V is a user-defined parameter representing the cardinality of the word embedding. Each row of $\mathcal{W}_{in}$ contains word embedding of a specific alarm, whereas rows of $\mathcal{W}_{out}$ represent contextual information between alarms.

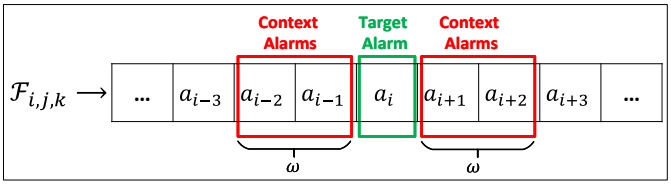


Fig. 3: An example of target and context alarms in a flood. Here, a_i represents the ith alarm, $\mathcal{F}_{i,j,k}$ represents the incoming alarm flood, and ω is the user-defined parameter (window size) to determine the number of context alarms.

Thereafter, $\mathcal{W}_{in}$ and $\mathcal{W}_{out}$ are tuned to learn the contextual relationship between alarms.

One model is trained on each AF cluster, where it iterates over the alarms in each flood sequence. The concept of "context alarm" and "target alarm" is better explained using Fig. 3, where an example of an alarm sequence of an ongoing AF is analyzed. As the model iterates over each alarm in the flood, the alarm currently being analyzed (a_i) is referred to as the "target alarm" and the alarms in the vicinity of a_i, namely, a_{i-2}, a_{i-1}, a_{i+1}, and a_{i+2}, are referred to as the "context alarms". The user-specified parameter $\omega \in \mathbb{N}^+$ (window size) determines the number of "context alarms". The conditional probability of an alarm a_k being a context alarm for the target alarm a_i is obtained from the output y_p as the softmax function transformation $\Theta(\cdot)$ of $\mathbf{h}_i \cdot \mathcal{W}_{out}^T$ given by [18],

$$y_p = \Theta\left(\mathbf{h}_i \cdot \mathcal{W}_{out}^T\right) = \begin{bmatrix} p(a_1 \mid a_i) \\ p(a_2 \mid a_i) \\ \cdots \\ p(a_K \mid a_i) \end{bmatrix}, \tag{8}$$

where, y_p is a vector of dimension K, and satisfies that $\sum_{k=1}^{K} p(a_k|a_i) = 1$. Afterward, the weights $\theta = [\mathcal{W}_{in}, \mathcal{W}_{out}]$ of the model are tuned to minimize the prediction error $(y_p - y_{true})$. Here, y_{true} is a one-hot encoded vector, with the conditional probability of the true context alarm, $p(a_k \mid a_i) = 1$, and of the rest alarms is 0. The internal parameters of the model are tuned to maximize the probability of predicting all the correct context alarms based on

$$\hat{\theta} = \arg\min_{\theta} \left(-log \prod_{c} p(a_c \mid a_i) \right), \tag{9}$$

where $\hat{\theta}$ represents the updated model weights, and c indicates true context alarms. One word2vec model is trained for each cluster $\mathcal{C}_i$ of similar AF, where the model embeds the contextual relationships between alarms in a specific fault, resulting in an ensemble of N models, $\Phi = \{\Phi_1, \Phi_2, \cdots, \Phi_N\}$. Therefore, it can be seen that a cluster $\mathcal{C}_i$ would be associated with a fault Λ_i and a word2vec model Φ_i.

C. Online Stage: AF Prediction and Classification

In the online stage, the alarms in an ongoing AF are fed to the ensemble of word2vec models obtained in the offline stage. Each of the alarms is considered a target alarm and the most probable context alarm is determined as the alarm with the highest conditional probability in y_p. Consider the ongoing AF $\mathcal{F}$,

$$\mathcal{F} = \{a_1, a_2, ..., a_{|\mathcal{F}|}\}, \tag{10}$$

where, a_i represents the ith alarm in $\mathcal{F}$, $i = 1, 2, \cdots, |\mathcal{F}|$. The N ensemble models predict the most probable context alarms based on contextual relationships captured in the offline stage. As the AF proceeds, the predictions are updated as

$$\mathcal{Y}_p = [y_{p,1}, y_{p,2}, \cdots, y_{p,|\mathcal{F}|}], \tag{11}$$

where, $\mathcal{Y}_p$ is the matrix of predictions, obtained by the concatenation of rank-ordered predictions $y_{p,i}$, and $i = 1, 2, \cdots, |\mathcal{F}|$. Specifically, elements in $y_{p,i}$ are rearranged (sorted) in descending order before concatenation. Thus, each model returns a matrix of predictions $\mathcal{Y}_p \in \mathbb{R}^{\mathrm{K} \times |\mathcal{F}|}$, where each column represents the predictions corresponding to alarm $a_i \in \mathcal{F}$. The top n predictions are selected from $\mathcal{Y}_p$ as

$$\overline{\mathcal{Y}}_{p,n} = [\overline{y}_{p,1}, \overline{y}_{p,2}, \cdots, \overline{y}_{p,|\mathcal{F}|}] = [\pi_1, \pi_2, \cdots, \pi_n]^T, \tag{12}$$

where, $\overline{y}_{p,i}$ is the ith column of $\overline{\mathcal{Y}}_{p,n}$, and π_1 (π_n) is the first (nth) row in $\mathcal{Y}_p$, and it satisfies that $\pi_{1,i} \geq \pi_{2,i} \geq \cdots \geq \pi_{K,i}$, where $\pi_{i,j}$ represents the (i,j)th value of $\mathcal{Y}_p$. Thereafter, the alarms corresponding to each prediction are obtained in the form of a matrix as

$$\mathcal{M} = [\mu_1, \mu_2, \cdots, \mu_{|\mathcal{F}|}] = \{a_{i,j} | a_{i,j} \succ \pi_{i,j}, a \in \mathcal{A}\}, \tag{13}$$

where $\mathcal{M}$ is the matrix of alarms, μ_i is the ith column of $\mathcal{M}$, and the operator $\succ$ indicates that the alarm $a_{i,j}$ is corresponding to the prediction value $\pi_{i,j}$. Thus, the output of

Algorithm 1: Online AF Prediction and Classification

Input: $\mathcal{F}$, Λ, Φ, ω_t, n
Output: Λ_{out}

```
S = {s_1 = 0, s_2 = 0, ..., s_|Φ| = 0}   ▷ Initialize scores
for a_i in F do
    Obtain the index i of a_i in F
    for Φ_j in Φ do
        Find the index j of Φ_j in Φ
        Calculate y_p for a_i from Φ_j by (8)
        Obtain ȳ_{p,i} the top n predictions by (12)
        Obtain the corresponding alarms μ_i by (13)
    end
    for a_past in F[i − ω_t : i] do
        if a_past ∈ μ_i then
            s_j = s_j + 1                ▷ Increase the score
        end
        Calculate y_p for a_past from Φ_j by (8)
        Obtain ȳ_{p,past} for a_past by (12)
        Obtain μ_past by (13)
        if a_i in μ_past then
            s_j = s_j + 1                ▷ Increase the score
        end
    end
end
Find the index k of the highest score in S
The fault category Λ_out of F is the kth element of Λ
return Λ_out
```

the model can be interpreted as a list of alarms rank-ordered based on their likelihood of occurrence as a context alarm. In other words, the first column of $\overline{\mathcal{Y}}_{p,n}$ represents the n most probable context alarms of $a_1 \in \mathcal{F}$.

Furthermore, an incremental scoring system is introduced to reward models for generating correct predictions. At the beginning of the predictions, the scores are initialized to 0 and are incremented by one unit for each correct prediction. The algorithm requires two user-specified parameters, namely, ω_t and n. Here, $\omega_t \in \mathbb{N}^+$ determines the number of context alarms to be considered in the ongoing AF sequence. It has to be noted that ω_t has the same purpose as that of ω utilized in the model training, as described in Section II-B, and it is not necessary that $\omega_t = \omega$. The parameter $n \in \mathbb{N}^+$ determines the number of top predictions to be selected in (12) to identify the most similar AF sequence.

If an AF is triggered due to a fault Λ_i, the model $\Phi_i \in \Phi$ trained on the cluster $\mathcal{C}_i$ corresponding to Λ_i would give the most accurate predictions and hence would result in the highest number of similar context alarms. This principle is utilized to classify the ongoing AF, i.e., the AF is classified into the fault category of the model with the highest score. Algorithm 1 summarizes the online AF Prediction and Classification.

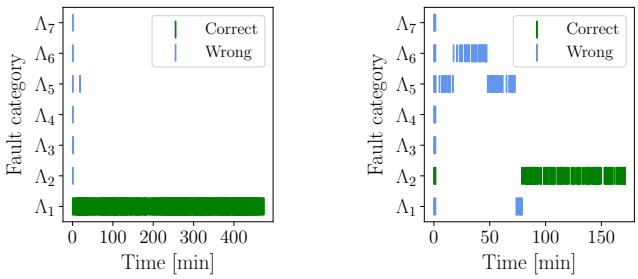
III. Case Study

The applicability and effectiveness of the proposed method are demonstrated through a case study using simulated alarm data from the benchmark Tennessee Eastman Process (TEP).

A. Description of the Simulated Data

The closed-loop simulator of the benchmark Tennessee Eastman Process, developed by Bathelt *et al*. [19] was utilized in this study. The alarm data was prepared following the procedure in [12]. Specifically, seven faults $\Lambda = \{\Lambda_1, \Lambda_2, \cdots, \Lambda_7\}$ were simulated by introducing disturbances as step inputs or valve stiction. Four types of alarms were configured on 52 process variables (PV), namely, "PV.HH" (High-High), "PV.H" (High), "PV.L" (Low), and "PV.LL" (Low-Low), resulting in 208 unique alarms. For each fault category, a set of 40 independent simulations were performed, where the duration of each simulation was 10h and the faults were introduced after 2h of steady-state operation. The chattering alarms were identified and discarded. Alarm floods were extracted using (6), and 7 clusters with 40 AF sequences each were obtained. Here, the alarm sequences generated by a specific fault were regarded as a cluster of similar alarm sequences. The number of AF sequences in each fault category is as follows: $|\mathcal{C}_1| = 41$, $|\mathcal{C}_2| = 71$, $|\mathcal{C}_3| = 25$, $|\mathcal{C}_4| = 1$, $|\mathcal{C}_5| = 42$, $|\mathcal{C}_6| = 66$, and $|\mathcal{C}_7| = 40$. Due to insufficient AF sequences, cluster $\mathcal{C}_4$ associated with fault Λ_4 was excluded from the analysis.

Thereafter, an ensemble of 6 word2vec models was trained using a Leave-one-out cross-validation approach. Specifically, the first 20 flood sequences of each fault category were selected to evaluate the performance of the models. Therefore, 20 independent simulations have been performed such that in each simulation, models were trained on all the AF sequences except one (test sequence) and the last AF sequence was used to evaluate the model performance and tuning. Subsequently, the online alarm flood classification is performed using this ensemble of word2vec models.



a) True category → Λ_1 b) True category → Λ_2

Fig. 4: Online classification of two floods that originated from fault category (a) Λ_1 and (b) Λ_2. The vertical lines represent the result of the classification, where the correct (incorrect) predictions are shown in green (blue).

B. Results and Discussion

The models were trained using the open-source Python library Gensim [20] v4.2.0 running on Python v3.9.12. For the offline stage, the parameters used were $\omega = 5$, $V = 15$, and the number of epochs = $100,000$, which indicates the number of iterations made by the model over the dataset. For the online predictions and classification, the parameters used were $\omega_t = 3$ and $n = 10$. Fig. 4 provides an example of the output from the online prediction and classification stage. The vertical lines represent the result of the classification, where the correct (incorrect) predictions are shown in green (blue). Fig. 4(a) shows the accurate prediction and classification of the AF sequence to be associated with fault Λ_1, without any significant delay. However, the AF sequence in Fig. 4(b) was classified into fault Λ_2 after about 70 minutes.

To provide a comprehensive overview of the model performance, the class-wise accuracy has been calculated by considering all predictions obtained for each fault category. In addition, to evaluate if the model was able to accurately classify the AF sequence using the complete AF sequence, the class-wise accuracy was determined by considering only the last prediction (i.e., the last alarm of a flood sequence). These performance metrics of the model using class-wise accuracy are summarized in Table I. It can be seen that the models achieve prediction accuracy above 0.85 (all predictions) and 0.90 (last predictions), respectively, in the identification of fault categories Λ_1, Λ_3, and Λ_5. The model performance is satisfactory for category Λ_6, which shows an accuracy of 0.77 considering all the predictions. However, the identification of fault category Λ_2 is particularly challenging (accuracy = 0.37).

Furthermore, the results indicate that the model performance increases as the flood proceeds because the accuracy based on the last prediction is always greater than the accuracy based on all predictions. This behavior is especially evident for categories Λ_2 and Λ_3 and it may indicate that most errors are made during the early stages of AFs (see Fig. 4.b). This

TABLE I: Class-wise accuracy

Accuracy Type	Fault Categories						Mean Accuracy
	Λ_1	Λ_2	Λ_3	Λ_5	Λ_6	Λ_7	
All predictions	0.98	0.37	0.89	1.0	0.77	0.001	0.67
Last prediction	1.0	0.55	0.9	1.0	0.85	0.00	0.72

could be explained by the lack of sufficient information at the beginning of the AF. Finally, it is to be noted that the models could not classify the sequences associated with the fault Λ_7 (accuracy=0.01), and those sequences (belonging to Λ_7) were always incorrectly classified into the fault Λ_2. Further investigation using process knowledge is recommended to identify the reason for such a performance with fault Λ_7.

In summary, the performance of the models is satisfactory, which indicates that the proposed method can support the operator in the real-time monitoring of AFs. The challenges with lower accuracy values for two fault categories could be attributed to the fact that model hyper-parameters were chosen based on the best practices and were not tuned for this specific process or application. The model performance is expected to improve from an exhaustive hyper-parameter tuning using a grid-search algorithm. Further research is recommended to improve the detection performance during the early stage of the AF, employing different NLP algorithms, such as BERT [21] and XLNET [22].

IV. Conclusions

This study presents a novel approach for online alarm flood classification using the word2vec algorithm and historical A&E data. The method articulates in two phases, namely, offline training and online predictions. In the offline stage, the contextual relationships between alarms are captured to train an ensemble of word2vec models using clusters of labeled AF sequences. In the online stage, the most probable context alarms are predicted using the trained ensemble of models, and the AF is classified into appropriate fault categories using a scoring system. Unlike other methods in AF classification, this study utilizes the NLP algorithms not only to learn hidden relationships between alarms but also to perform the classification of ongoing AFs without any classifiers or clustering algorithms, thereby reducing the computational complexity. The approach has been tested on simulated alarm data obtained from the benchmark TEP. The models achieved accuracy in the range of 0.77 to 1.0 in four out of six categories. Additionally, the results indicate that the model performance improves as the AF proceeds, leading to more accurate predictions as more information is available.

Further research is recommended to optimize the model parameters and improve the accuracy during the early stage of AFs. Additional investigation using process knowledge is required in the two fault categories resulting in poor prediction accuracy. Notwithstanding these limitations, the approach shows the potential of NLP algorithms in alarm flood analysis and makes a significant contribution to the novel line of research using NLP models for online AF classification.

References

[1] G. Manca and A. Fay, "Detection of historical alarm subsequences using alarm events and a coactivation constraint," *IEEE Access*, vol. 9, pp. 46 851–46 873, 2021.

[2] J. Wang, F. Yang, T. Chen, and S. L. Shah, "An overview of industrial alarm systems: Main causes for alarm overloading, research status, and open problems," *IEEE Transactions on Automation Science and Engineering*, vol. 13, no. 2, pp. 1045–1061, 2016.

[3] *ANSI/ISA-18.2: Management of Alarm Systems for the Process Industries*, ISA (International Society of Automation), Durham, NC USA, 2016.

[4] *Alarm Systems: A Guide to Design, Management and Procurement*, EEMUA (Engineering Equipment and Materials Users' Association), London, 2013.

[5] G. Dorgo, F. Tandari, T. Szabó, A. Palazoglu, and J. Abonyi, "Quality vs. quantity of alarm messages-how to measure the performance of an alarm system," *Chemical Engineering Research and Design*, vol. 173, pp. 63–80, 2021.

[6] Y. Cheng, I. Izadi, and T. Chen, "Pattern matching of alarm flood sequences by a modified Smith–Waterman algorithm," *Chemical Engineering Research and Design*, vol. 91, no. 6, pp. 1085–1094, 2013.

[7] W. Hu, J. Wang, and T. Chen, "A local alignment approach to similarity analysis of industrial alarm flood sequences," *Control Engineering Practice*, vol. 55, pp. 13–25, 2016.

[8] G. Manca, M. Dix, and A. Fay, "Clustering of similar historical alarm subsequences in industrial control systems using alarm series and characteristic coactivations," *IEEE Access*, vol. 9, pp. 154 965–154 974, 2021.

[9] Q.-X. Zhu, C. Jin, Y.-L. He, and Y. Xu, "Pattern mining of alarm flood sequences using an improved prefixspan algorithm with tolerance to short-term order ambiguity," *Industrial & Engineering Chemistry Research*, vol. 60, no. 11, pp. 4375–4384, 2021.

[10] S. Lai, F. Yang, and T. Chen, "Online pattern matching and prediction of incoming alarm floods," *Journal of Process Control*, vol. 56, pp. 69–78, 2017.

[11] M. Lucke, M. Chioua, C. Grimholt, M. Hollender, and N. F. Thornhill, "Advances in alarm data analysis with a practical application to online alarm flood classification," *Journal of Process Control*, vol. 79, pp. 56–71, 2019.

[12] J. Shang and T. Chen, "Early classification of alarm floods via exponentially attenuated component analysis," *IEEE Transactions on Industrial Electronics*, vol. 67, no. 10, pp. 8702–8712, 2019.

[13] H. S. Alinezhad, J. Shang, and T. Chen, "Early classification of industrial alarm floods based on semisupervised learning," *IEEE Transactions on Industrial Informatics*, vol. 18, no. 3, pp. 1845–1853, 2021.

[14] M. R. Parvez, W. Hu, and T. Chen, "Real-time pattern matching and ranking for early prediction of industrial alarm floods," *Control Engineering Practice*, vol. 120, p. 105004, 2022.

[15] H. Wang, F. Khan, and S. Ahmed, "Design of scenario-based early warning system for process operations," *Industrial & Engineering Chemistry Research*, vol. 54, no. 33, pp. 8255–8265, 2015.

[16] OpenAI, "Gpt-4 technical report," 2023.

[17] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *1st International Conference on Learning Representations - Workshop Track Proceedings*, 2013.

[18] J. Bridle, "Training stochastic model recognition algorithms as networks can lead to maximum mutual information estimation of parameters," *Advances in Neural Information Processing Systems*, vol. 2, 1989.

[19] A. Bathelt, N. L. Ricker, and M. Jelali, "Revision of the Tennessee Eastman Process model," *IFAC-PapersOnLine*, vol. 48, no. 8, pp. 309–314, 2015.

[20] R. Řehůřek and P. Sojka, "Software framework for topic modelling with large corpora," *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pp. 45–50, 2010.

[21] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "BERT: Pre-training of deep bidirectional transformers for language understanding," in *2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 2019, pp. 4171 – 4186.

[22] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. R. Salakhutdinov, and Q. V. Le, "XLNet: Generalized autoregressive pretraining for language understanding," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

Article X.

Tamascelli, N., Dal Pozzo, A., Liu, Y., Cozzani, V., Paltrinieri, N. (2022). **Integration between data-driven process simulation models and resilience analysis to improve environmental risk management in the Waste-to-Energy industry**. Proceedings of the 32nd European Safety and Reliability Conference (ESREL 2022), Dublin, Ireland, 2022. 1409–1416. https://doi.org/10.3850/978-981-18-5183-4_R23-03-206-cd.

Proceedings of the 32nd European Safety and Reliability Conference (ESREL 2022)
Edited by Maria Chiara Leva, Edoardo Patelli, Luca Podofillini, and Simon Wilson
 Published by Research Publishing, Singapore.
doi: 10.3850/978-981-18-5183-4_R23-03-206-cd

Integration between data-driven process simulation models and resilience analysis to improve environmental risk management in the Waste-to-Energy industry

Nicola Tamascelli

Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology, Norway. E-mail: nicola.tamascelli@ntnu.no

Alessandro Dal Pozzo

Department of Civil, Chemical, Environmental, and Materials Engineering, University of Bologna, Italy. E-mail: alessandro.dalpozzo3@unibo.it

Yiliu Liu

Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology, Norway. E-mail: yiliu.liu@ntnu.no

Valerio Cozzani

Department of Civil, Chemical, Environmental, and Materials Engineering, University of Bologna, Italy. E-mail: valerio.cozzani@unibo.it

Nicola Paltrinieri

Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology, Norway. E-mail: nicola.paltrinieri@ntnu.no

Municipal Solid Waste Incineration plants must comply with stringent emissions standards. Flue gas treatment technologies are essential to ensure compliance and protect human health and the environment. Although the most recent research has focused on estimating the risk for human health and comparing different gas treatment strategies, few efforts have been directed toward the definition of a thorough methodology for identifying critical scenarios and evaluating safety barriers. In this context, this study aims at filling this knowledge gap and investigating how traditional hazard identification techniques and novel approaches (data-driven process simulation models and Resilience analysis) may be used to (i) identify critical events that may lead to an overrun of emission limits, (ii) identify additional safety barriers that may prevent/mitigate such events, (iii) simulate the system behavior with and without additional safety barriers, and (iv) quantify the gain in performance and resilience and support decision-making. The methodology has been tested on a single-stage Dry Sorbent Injection (DSI) system. Actual data from a waste incineration plant have been used to develop the data-driven model. The results suggest that the method is particularly suited for evaluating and comparing design alternatives in industrial facilities where field tests are impractical or dangerous due to strict regulations and the inherent dangerousness of operations.

Keywords: Environmental Risk Management, Waste-to-Energy, Data-Driven Process Simulation, Resilience analysis

1. Introduction

The rising global population and growing urbanization have driven a continuous rise in Municipal Solid Waste generation (Nanda and Berruti 2021). In addition, the increased awareness of sustainability and environmental issues has raised new challenges and concerns about waste management (Sabbas et al. 2003).

Waste incineration is one of the most efficient and well-established methods to reduce the volume of non-recyclable wastes and recover energy from the process (Das et al. 2019). However, combustion gases may contain a significant amount of acidic compounds (e.g., HCl, SO_2, HF), which must be removed before releasing the gas into the atmosphere. The growing concern for climate change and environmental degradation has led to the adoption of increasingly stringent and ambitious emission limits, which eventually pushed the industry to the development of more advanced and efficient gas treatment systems (Dal Pozzo, Guglielmi, et al. 2018). In this context, one of the most effective techniques to neutralize acidic compounds is the injection of dry alkaline sorbents, such as $Ca(OH)_2$ and $NaHCO_3$, into the flue gas (Dal Pozzo et al. 2017).

Although considerable progress has been made in the area of pollution control, situations may arise where the gas treatment system cannot manage unexpected internal or external disturbances, which may lead to exceeding the emission limits. In these situations, non-compliant plants may incur fines, loss of reputation, and environmental damage. Therefore, it is critical to ensure that potential sources of deviation are correctly identified, and strategies to improve the performance and robustness of the system are evaluated and eventually implemented.

Nevertheless, most studies in the literature propose methodologies to assess the impact of MSWI emissions on human health (Meneses, Schuhmacher, and Domingo 2004; Morselli et al. 2011; Scungio et al. 2016), or to compare the effectiveness and efficiency of different gas treatment technologies (Bodénan and Deniard 2003; Dal Pozzo et al. 2016). To the best of the authors' knowledge, no past study presented a thorough methodology to identify critical scenarios that may lead to an overrun of emission limits and evaluate the effectiveness of additional safety barriers to prevent or mitigate such events.

In order to fill this knowledge gap, this study proposes a novel method based on the integration between traditional hazard identification approaches and novel techniques, such as data-driven models and resilience analysis. This new methodology may be used to (i) identify critical events that may lead to an overrun of emission limits, (ii) identify additional safety barriers that may prevent/mitigate such events, (iii) simulate the system behavior with and without additional safety barriers, and (iv) quantify the gain in performance and resilience

The paper is organized into five sections. Section 2 describes the reference gas treatment system used to support and illustrate the analyses. The methodology is described in section 3, and the results are presented and discussed in section 4. Finally, conclusions are drawn in section 5.

2. Gas Treatment System Description

In order to support and describe the approach, a single-stage dry sorbent injection system is taken as a reference. The main equipment, controllers, and actuators are shown in Fig. 1.

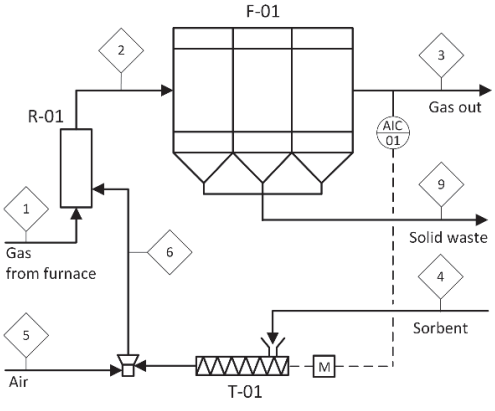


Fig. 1. Process scheme of the gas treatment system

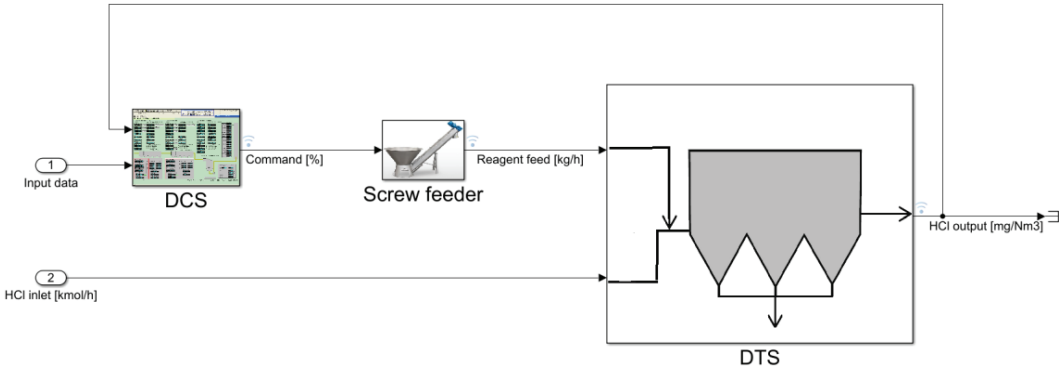


Fig. 2. The digital model of the gas treatment plant described in Fig. 1. "DCS" indicates the Distributed Control System, and "DTS" indicates the Dry Treatment System.

The combustion gas (stream 1 in Fig. 1) flows from the furnace into the in-line reactor (R-01). Here, the gas is mixed with the hydrated lime (stream 6), and acidic compounds are neutralized according to the following reactions:

$$2HCl + Ca(OH)_2 \leftrightarrow CaCl_2 + 2H_2O \quad (1)$$

$$SO_2 + Ca(OH)_2 + 1/2\, H_2O \leftrightarrow CaSO_4 + H_2O \quad (2)$$

$$HF + 1/2\, Ca(OH)_2 \leftrightarrow 1/2\, CaF_2 + H_2O \quad (3)$$

Next, the gas stream proceeds through the filter (F-01), where the spent sorbent (stream 9) is separated from the gas (stream 3). A feedback control loop and a PI controller are used to manipulate the sorbent flow rate by varying the speed of the screw feeder motor drive (M in Fig. 1). The controlled variable is the acidic compound concentration downstream of the bag filter (AIC-01).

The most critical and abundant acidic compound is HCl (Dal Pozzo et al. 2016). For this reason, this study focuses on HCl emissions, and the other compounds are not considered further. In Europe, MSWI plants fall under the Directive 2010/75/EU of the European Parliament and of the Council on industrial emissions (European Council 2010), which sets the half-hourly average HCl emission limit at 10 mg/Nm^3. That is, the average HCl concentration calculated over a 30-minutes time window must be lower than 10 mg/Nm^3.

3. Method

The methodology comprises three steps: (i) identification of critical scenarios and additional safety measures, (ii) scenarios simulation, and (iii) performance and resilience analysis. The system in Fig. 1 is used to support the analysis. Also, the analyses focus on HCl only. However, the methodology has general validity and can be promptly adapted and extended to consider different chemical compounds and different gas treatment systems.

3.1. Identification of critical scenarios and additional safety measures

Historical plant data was analyzed to identify process deviations that caused a significant increase in the HCl concentration downstream of the gas treatment section. These unwanted events, namely "critical scenarios", were further examined to identify their causes and consequences. As a result, a list of critical scenarios was obtained.

Later, a semi-structured interview was conducted with the plant manager and other experienced engineers working in the same sector. The objective of the interview was twofold. Firstly, to exclude improbable or less critical scenarios. Secondly, to discuss and propose a list of additional safety measures (e.g., safety barriers) that may prevent or mitigate the effects of such critical scenarios.

As a result, a list of credible critical scenarios and recommendations for additional safety measures is obtained.

3.2. Scenario simulation

A digital model of the gas treatment section was built to simulate (i) the system behavior during critical scenarios and (ii) the effect of the additional safety measures. One of the major advantages of this approach is the opportunity to evaluate different process configurations without the need for field tests, which are impractical due to strict regulations.

The plant model depicted in Fig. 2 was developed using Matlab and Simulink, and comprises three main blocks:

(i) Distributed Control System (DCS);
(ii) Screw feeder;
(iii) Dry Treatment System (DTS).

The block "DCS" is designed to mimic the plant control logic. It takes as input the HCl concentration in the gas stream entering the system and the HCl concentration in the stream leaving the treatment section. A PI controller compares the two measures and delivers a signal to the "Screw Feeder block", which calculates the corresponding sorbent mass flow rate.

The block "DTS" represents the core of the model. It converts the molar flow rates of HCl and sorbent entering the system into the HCl concentration in the gas leaving the system. In other words, the block mimics the neutralization reaction described by Eq. (1).

The modeling of the neutralization process through first principles would pose significant difficulties because of the complexity of the phenomena involved (e.g., convection, diffusion in a solid porous media, reaction kinetics, thermodynamic equilibria) and other external and internal factors that are difficult to control and monitor (e.g., thickness and reactivity of the filter cake, the composition of the flue gas, changes in the sorbent structure) (Giacomo Antonioni et al. 2016; Dal Pozzo, Moricone, et al. 2018).

This study proposes a data-driven model to derive the process dynamic directly from historical data and avoid the limitations of a first principle approach. Specifically, the HCl molar flow rate in the outlet flue gas $\left(\dot{n}_{HCl,out}\right)$ is calculated as follows:

$$\dot{n}_{HCl,out}(t) = \sum_{i=1}^{3} a_i \cdot \dot{n}_{HCl,out}(t-i) + \\ + \sum_{j=0}^{2} b_j \cdot \dot{n}_{HCl,in}(t-j) \cdot \left(1 - \chi(t-j)\right) \quad (4)$$

Where:

- $\dot{n}_{HCl,out}(t-i)$ = outlet HCl molar flow rate at time $t-i$;
- $\dot{n}_{HCl,in}(t-j)$ = inlet HCl molar flow rate at time $t-j$;
- $\chi(t-j)$ = HCl conversion at time $t-j$;
- $a_i, b_j \in [0,1]$ = model parameters.

The conversion χ at a given time instant t is calculated according to the empirical model proposed by Antonioni et al. (2011):

$$\chi(t) = \frac{SR(t)^n - SR(t)}{SR_{lime}(t)^n - 1} \quad (5)$$

Where $SR(t)$ represents the ratio between the sorbent flow rate injected at time t and the stoichiometric sorbent demand required to neutralize the HCl entering the system at time t. The exponent n is an adjustable parameter obtained from fitting actual plant data. In this study, n = 5.

The model calculates $\dot{n}_{HCl,out}(t)$ as the sum of two contributions: autoregressive and exogenous. The first contribution takes into account the value of $\dot{n}_{HCl,out}$ up to three timesteps in the past. The second summation takes into account the fraction of unreacted $\dot{n}_{HCl,in}$ at the current time and up to two timesteps in the past. The regression parameters a_i and b_j can be learned from historical data by, e.g., Least Squares minimization; they represent the weights of the different contributions in Eq. (4). In this study, these parameters were set manually for demonstration purposes. Equal weights were assigned to the autoregressive and exogenous parts. Specifically, the weights are: $a_1=b_0=0.3$, $a_2=b_1=0.15$, and $a_3=b_2=0.05$.

3.3. Performance and resilience analysis

The results of the simulations were used to calculate performance and resilience metrics, which ultimately allow the evaluation of the safety measures.

In this study, the resilience of the gas treatment system may be interpreted as the capacity to perform its purpose in a variety of adverse conditions. Resilience analysis was used because it represents an interesting attempt to go beyond the canonical safety approach – which focuses primarily on failures and probabilities of failures – and evaluate the system based on the ability to "function under both expected and unexpected conditions rather than just to avoid failures" (Hollnagel et al. 2010).

Over the past two decades, several metrics have been proposed to quantify the resilience of engineered systems (Yodo and Wang 2016). Many resilience metrics rely on a time-dependent function representing the system performance (Hosseini, Barker, and Ramirez-Marquez 2016). The typical trend of the performance metric φ(t) after a disruptive event *e* is represented in Fig. 3.

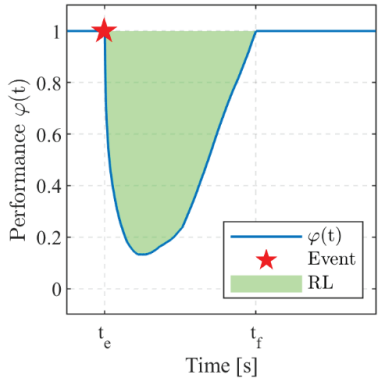


Fig. 3. Time trend of system performance after a disruptive event. "RL" indicates the Resilience Loss (Eq. (7)), t_e and t_f indicate the event occurrence and the recovery time.

In this study, the system performance is formulated as follows:

$$\varphi(t) = \begin{cases} 1, & if\ \ \bar{C}_{HCl}(t) \le T \\ 1 - \dfrac{\bar{C}_{HCl}(t) - T}{L - T}, & if\ \ T < \bar{C}_{HCl}(t) < L \\ 0, & if\ \ \bar{C}_{HCl}(t) \ge L \end{cases} \quad (6)$$

Where $\bar{C}_{HCl}(t)$ is the half-hourly HCl concentration at time t, L = 10 mg/Nm^3 is the law limit, and T = 7.5 mg/Nm^3 is the controller setpoint increased by 10 %. In words, $\varphi(t)$ = 1 if $\bar{C}_{HCl}(t)$ stays within ±10 % of the controller setpoint. When $\bar{C}_{HCl}(t)$ deviates more, the system is degrading and $\varphi(t)$ decreases. If $\bar{C}_{HCl}(t)$ increase further and exceeds the law limit, the system has failed to fulfil its purpose, and therefore $\varphi(t)$ = 0.

The Resilience Loss metric (RL) is used to quantify the (non)resilience of the system (Bruneau et al. 2003). RL represents the loss in performance after event *e*. Geometrically, it can be interpreted as the area between the curves φ = 1 and φ = φ(t) – i.e., the shaded area in Fig. 3; that is:

$$RL = \int_{t_e}^{t_f} [1 - \varphi(t)]dt \quad (7)$$

Where t_e and t_f indicate the time occurrence of event *e* and the recovery time, as indicated in Fig. 3.

One RL is calculated for each simulated scenario and additional safety measure. The comparison between RL measures permits to quantify the effectiveness of additional safety measures. Intuitively, design solutions that exhibit smaller Resilience Loss are preferred.

4. Results and Discussion

The analysis of past events and the interview with the plant manager highlighted that an abrupt increase in the HCl concentration upstream of the treatment section is one of the most probable critical events. This scenario could be caused by, e.g., the delivery of a significant amount of high-chloride waste. In addition, the installation of a furnace sorbent injection system was proposed as an additional safety measure that may prevent the critical scenario from escalating and exceeding the HCl emission limits.

The functioning of the furnace sorbent injection was described by Biganzoli et al. (2015; 2015). In essence, this technology involves the injection of dolomitic powder directly into the furnace. It has been demonstrated that the system effectively reduces the acidic compound content and decreases the

workload for the subsequent gas treatment system (Dal Pozzo et al. 2020). In this study, furnace sorbent injection is proposed as an additional safety barrier that activates when a peak in the HCl concentration is detected.

The digital model of the plant was modified in order to account for the effect of the furnace sorbent injection system. Firstly, the behavior of the system without additional safety barriers was simulated. The HCl peak was modeled as a pulse disturbance of duration 15 minutes and amplitude equal to 1, 3, 5.5, and 6 times the normal HCl concentration. Secondly, a new set of simulations was performed considering the response of the additional safety barrier. To this end, the furnace sorbent injection system was activated 2 minutes after the detection of the peak and deactivated when the peak ended. Different dolomitic sorbent feed rates were simulated. Specifically, the ratio (SR) between the actual sorbent flow rate and the stoichiometric demand was set to 1, 1.7, and 2.5.

Fig. 4 and Fig. 5 show the most significant results obtained from the performance analysis. Fig. 4 displays the performance measure calculated for an HCl peak equal to 3 times the normal HCl concentration. Fig. 5 refers to a peak amplitude equal to 5.5 times the normal concentration. In each figure, the system performance obtained without safety barrier (i.e., SR = 0) and with the barrier (SR = 2.5) is displayed.

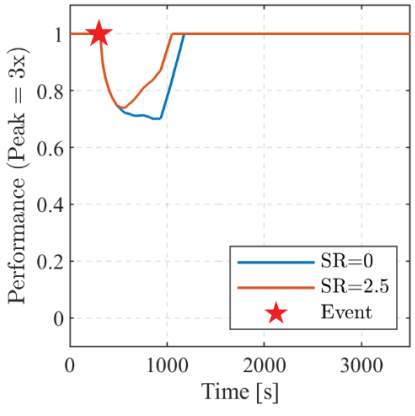


Fig. 4. Performance analysis for peak amplitude 3x.

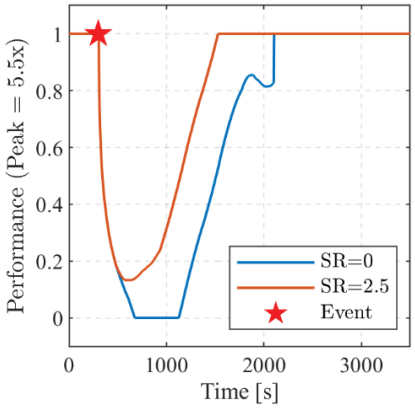


Fig. 5. Performance analysis for peak amplitude 5.5x.

Fig. 4 suggests that the system can manage HCl peaks equal to 3 times the normal concentration levels. In fact, the performance metric is always larger than zero, indicating that the half-hourly HCl concentration is always below the law limit. However, the system performance shows a slight degradation, especially when no additional safety barrier is installed (blue curve), which indicates that the disturbance can cause a significant deviation of the outlet HCl. The use of furnace sorbent injection with SR = 2.5 (orange curve) reduces the maximum degradation of the system and ensures faster recovery.

Fig. 5 reveals that a peak of 5.5 times the normal HCl concentration causes the complete degradation of the system if no additional safety barrier is installed (blue curve). Specifically, the performance degrades rapidly after the event and eventually reaches 0 after 6 minutes, indicating that the emission limit has been crossed. On the other hand, the performance of the system with furnace sorbent injection (orange curve) is always larger than zero, indicating that the additional safety barrier would have avoided crossing law limits and ensured faster recovery.

Performance curves were used to calculate the resilience loss according to Eq. (7). Table 1 presents the results obtained from the RL analysis.

Table 1. Resilience Loss for different dolomitic sorbent feed rates (SR). Parentheses indicate the gain in RL with respect to the simulation without the additional safety barrier.

RS	RL (3x)	RL (5.5x)
0	196	1166
1	187 (- 5 %)	1081.6 (- 7 %)
1.7	170 (- 13 %)	943.4 (- 19 %)
2.5	131 (- 33 %)	714.2 (- 39 %)

The table above shows resilience losses for the two critical scenarios considered (i.e., HCl peak equal to 3x and 5.5x). Each row represents a specific feed rate of dolomitic sorbent in the furnace injection system, from 0 (no injection) to 2.5. The data suggest that the furnace injection system decreases the resilience loss in every scenario. Therefore, the additional safety barrier can effectively relieve the effects of an abrupt increase in the inlet HCl concentration. As expected, the best results are obtained with larger dolomitic sorbent feed rates – i.e., 2.5 times the stoichiometric demand. Moreover, the improvements in system resilience are more than linear with increasing SR.

The results suggest that the proposed methodology can be used to (i) identify critical scenarios and additional safety measures and (ii) evaluate and compare the effectiveness of additional safety barriers without the need for field tests.

In spite of the promising results, it is worth acknowledging a few limitations. Firstly, only one safety barrier has been explicitly modeled. Secondly, the data-driven model needs to be further refined to ensure high accuracy for long simulation times. Thirdly, the performance measure proposed in this study does not consider the cost associated with installing and operating additional safety barriers and different design configurations. Finally, it must be acknowledged that the current methodology does not consider the possible side effects of installing and maintaining an additional barrier, such as the increase in system complexity. Further studies need to be carried out to address these issues and take into account economic aspects (e.g., the purchasing and installation of the barrier, the cost for the dolomitic sorbent, maintenance costs) in order to provide a more comprehensive framework for the evaluation and comparison of safety barriers.

5. Conclusions

This study proposes a novel approach based on traditional hazard identification approaches, data-driven models, and resilience analysis to enhance environmental risk management in the Waste to Energy industry. The attention has been directed toward the flue gas treatment section of municipal waste incineration plants. HCl emissions have been considered as a critical parameter. The results show that the proposed method enables (i) the identification of critical scenarios and additional safety measures, and (ii) qualitative and quantitative comparison between different design alternatives. The effectiveness of different safety barriers may be compared, and the attention of safety practitioners could be directed towards the most effective configurations. The proposed methodology appears particularly suited for evaluating and comparing design alternatives in industrial facilities where field tests are impractical or dangerous due to strict regulations and the inherent dangerousness of operations

References

Antonioni, G, F Sarno, D Guglielmi, P Morra, and V Cozzani. 2011. “Simulation of a Two-Stage Dry Process for the Removal of Acid Gases in a MSWI.” *Chemical Engineering Transactions* 24: 1063–68. https://doi.org/10.3303/CET1124178.

Antonioni, Giacomo, Alessandro Dal Pozzo, Daniele Guglielmi, Alessandro Tugnoli, and Valerio Cozzani. 2016. “Enhanced Modelling of Heterogeneous Gas-Solid Reactions in Acid Gas Removal Dry Processes.” *Chemical Engineering Science* 148: 140–54. https://doi.org/10.1016/j.ces.2016.03.009.

Biganzoli, Laura, Gaia Racanella, Roberto Marras, and Lucia Rigamonti. 2015. “High Temperature Abatement of Acid Gases from Waste Incineration. Part II: Comparative Life Cycle Assessment Study.” *Waste Management* 35: 127–34. https://doi.org/10.1016/j.wasman.2014.10.021.

Biganzoli, Laura, Gaia Racanella, Lucia Rigamonti, Roberto Marras, and Mario Grosso. 2015. “High Temperature Abatement of Acid Gases from Waste Incineration. Part I: Experimental

Tests in Full Scale Plants." *Waste Management* 36: 98–105. https://doi.org/10.1016/j.wasman.2014.10.019.

Bodénan, F., and Ph Deniard. 2003. "Characterization of Flue Gas Cleaning Residues from European Solid Waste Incinerators: Assessment of Various Ca-Based Sorbent Processes." *Chemosphere* 51 (5): 335–47. https://doi.org/10.1016/S0045-6535(02)00838-X.

Bruneau, Michel, Stephanie E. Chang, Ronald T. Eguchi, George C. Lee, Thomas D. O'Rourke, Andrei M. Reinhorn, Masanobu Shinozuka, Kathleen Tierney, William A. Wallace, and Detlof Von Winterfeldt. 2003. "A Framework to Quantitatively Assess and Enhance the Seismic Resilience of Communities." *Earthquake Spectra* 19 (4): 733–52. https://doi.org/10.1193/1.1623497.

Dal Pozzo, Alessandro, Giacomo Antonioni, Daniele Guglielmi, Carlo Stramigioli, and Valerio Cozzani. 2016. "Comparison of Alternative Flue Gas Dry Treatment Technologies in Waste-to-Energy Processes." *Waste Management* 51: 81–90. https://doi.org/10.1016/j.wasman.2016.02.029.

Dal Pozzo, Alessandro, Daniele Guglielmi, Giacomo Antonioni, and Alessandro Tugnoli. 2017. "Sustainability Analysis of Dry Treatment Technologies for Acid Gas Removal in Waste-to-Energy Plants." *Journal of Cleaner Production* 162: 1061–74. https://doi.org/https://doi.org/10.1016/j.jclepro.2017.05.203.

———. 2018. "Environmental and Economic Performance Assessment of Alternative Acid Gas Removal Technologies for Waste-to-Energy Plants." *Sustainable Production and Consumption* 16: 202–15. https://doi.org/10.1016/j.spc.2018.08.004.

Dal Pozzo, Alessandro, Lorenzo Lazazzara, Giacomo Antonioni, and Valerio Cozzani. 2020. "Techno-Economic Performance of HCl and SO2 Removal in Waste-to-Energy Plants by Furnace Direct Sorbent Injection." *Journal of Hazardous Materials* 394 (February): 122518. https://doi.org/10.1016/j.jhazmat.2020.122518.

Dal Pozzo, Alessandro, Raffaela Moricone, Giacomo Antonioni, Alessandro Tugnoli, and Valerio Cozzani. 2018. "Hydrogen Chloride Removal from Flue Gas by Low-Temperature Reaction with Calcium Hydroxide." *Energy and Fuels* 32 (1): 747–56. https://doi.org/10.1021/acs.energyfuels.7b03292.

Das, Subhasish, S. H. Lee, Pawan Kumar, Ki Hyun Kim, Sang Soo Lee, and Satya Sundar Bhattacharya. 2019. "Solid Waste Management: Scope and the Challenge of Sustainability." *Journal of Cleaner Production* 228: 658–78. https://doi.org/10.1016/j.jclepro.2019.04.323.

European Council. 2010. "Directive 2010/75/EU Industrial Emissions." *Official Journal of the European Union* L334: 17–119. https://doi.org/10.3000/17252555.L_2010.334.eng.

Hollnagel, Erik, Jean Paries, D Woods David, and John Wreathall. 2010. *Resilience Engineering in Practice: A Guidebook*. Ashgate Studies in Resilience Engineering. Ashgate Publishing. https://hal-mines-paristech.archives-ouvertes.fr/hal-00613345.

Hosseini, Seyedmohsen, Kash Barker, and Jose E. Ramirez-Marquez. 2016. "A Review of Definitions and Measures of System Resilience." *Reliability Engineering and System Safety* 145: 47–61. https://doi.org/10.1016/j.ress.2015.08.006.

Meneses, Montse, Marta Schuhmacher, and José L. Domingo. 2004. "Health Risk Assessment of Emissions of Dioxins and Furans from a Municipal Waste Incinerator: Comparison with Other Emission Sources." *Environment International* 30 (4): 481–89. https://doi.org/10.1016/j.envint.2003.10.001.

Morselli, Luciano, Fabrizio Passarini, Laura Piccari, Ivano Vassura, and Elena Bernardi. 2011. "Risk Assessment Applied to Air Emissions from a Medium-Sized Italian MSW Incinerator." *Waste Management and Research* 29 (10 SUPPL.): 48–56. https://doi.org/10.1177/0734242X10380115.

Nanda, Sonil, and Franco Berruti. 2021. "Municipal Solid Waste Management and Landfilling Technologies: A Review." *Environmental Chemistry Letters* 19 (2): 1433–56. https://doi.org/10.1007/s10311-020-01100-y.

Sabbas, T., A. Polettini, R. Pomi, T. Astrup, O. Hjelmar, P. Mostbauer, G. Cappai, et al. 2003. "Management of Municipal Solid Waste Incineration Residues." *Waste Management* 23 (1): 61–88. https://doi.org/10.1016/S0956-053X(02)00161-7.

Scungio, Mauro, Giorgio Buonanno, Luca Stabile, and Giorgio Ficco. 2016. "Lung Cancer Risk Assessment at Receptor Site of a Waste-to-Energy Plant." *Waste Management* 56: 207–15. https://doi.org/10.1016/j.wasman.2016.07.027.

Yodo, Nita, and Pingfeng Wang. 2016. "Engineering Resilience Quantification and System Design Implications: A Literature Survey." *Journal of Mechanical Design, Transactions of the ASME* 138 (11). https://doi.org/10.1115/1.4034223.

Article XI.

Tamascelli, N., Dal Pozzo, A., Scarponi, G.E., Paltrinieri, N., Cozzani, V. (2023) **Assessment of Safety Barrier Performance in Environmentally Critical Facilities: Bridging Conventional Risk Assessment Techniques with Data-Driven Modelling**. Process Safety and Environmental Protection. https://doi.org/10.1016/j.psep.2023.11.021.

Process Safety and Environmental Protection 181 (2024) 294–311

Contents lists available at ScienceDirect

Process Safety and Environmental Protection

journal homepage: www.journals.elsevier.com/process-safety-and-environmental-protection

Assessment of Safety Barrier Performance in Environmentally Critical Facilities: Bridging Conventional Risk Assessment Techniques with Data-Driven Modelling

Nicola Tamascelli [a,b], Alessandro Dal Pozzo [a], Giordano Emrys Scarponi [a], Nicola Paltrinieri [b], Valerio Cozzani [a,*]

[a] LISES - Laboratory of Industrial Safety and Environmental Sustainability, DICAM - Department of Civil, Chemical, Environmental and Materials Engineering, Alma Mater Studiorum – University of Bologna, via Terracini n.28, 40131 Bologna, Italy
[b] Department of Mechanical and Industrial Engineering, NTNU, Trondheim, Norway

ARTICLE INFO

Keywords:
Hazard identification
Safety barriers
Digital model
Dynamic performance assessment
Flue gas treatment
Waste-to-energy

ABSTRACT

The failure of emission control systems in industrial processes undergoing emission regulations can cause severe harm to the environment. In this context, safety engineering principles can be applied to analyze process deviations and identify suitable safety barriers to mitigate harmful emissions during critical events. However, the selection, design, and assessment of proper safety barriers may be complex due to several contingencies such as the inability to perform extensive field tests on systems under strict emission regulations. In this study, an approach is proposed to couple conventional hazard identification techniques with a digital model of a flue gas treatment system to support the identification and performance assessment of safety barriers for emission control. Resilience analysis is used to evaluate the behavior of the most relevant safety barrier options, selected through a screening with conventional hazard identification tools. Barriers are simulated using the digital model of the system, gathering key information for their design and evaluation, and overcoming the limitations to field tests at the real plant. The methodology is illustrated with reference to acid gas removal in waste-to-energy facilities, a relevant example of an emission control system that is typically exposed to significant process deviations.

1. Introduction

Several industrial processes have the potential to cause significant harm to the environment if their routine emissions to air and water are not minimized thanks to the application of proper treatment systems. In analogy with the definition of safety-critical systems in the field of safety engineering (Daintith and Wright, 2008; Knight, 2002; Maurya and Kumar, 2020), these systems can be defined environmentally critical systems, as their failure or malfunction may result in an unacceptable environmental damage.

Environmentally critical systems in the field of emission reduction need to exhibit: i) high performance, often corresponding to > 90% pollutant removal efficiency (e.g. see the Best Available Techniques reference documents of the European IPPC Bureau (European Commission, 2020)), and ii) high availability, according to the continuous operation of the plants on which they are installed.

Flue gas cleaning in waste-to-energy (WtE) plants represents a relevant example of such systems. WtE facilities are subject to some of the more stringent emission standards among industrial sectors (Dal Pozzo et al., 2023c; Van Caneghem et al., 2019) for a variety of pollutants, including nitrogen oxides (NO_X), acid pollutants such as hydrogen chloride (HCl) and sulfur dioxide (SO_2), and trace elements such as mercury (Hg). In Europe continuous emission measurement at stack is prescribed for these pollutants (European Commission, 2020). Therefore, WtE flue gas treatment (FGT) systems have to meet low emission levels in continuous operation, typically in presence of high fluctuations of the pollutant concentrations in the raw flue gas, as a consequence of the wide variety over time of the composition of the waste fed to the plant (Dal Pozzo et al., 2016).

In this context, the system is required to perform adequately during normal operating conditions and/or in the presence of external and internal disturbances. Actually, deviations caused by sudden variations in

* Corresponding author.
E-mail address: valerio.cozzani@unibo.it (V. Cozzani).

https://doi.org/10.1016/j.psep.2023.11.021
Received 20 July 2023; Received in revised form 4 October 2023; Accepted 10 November 2023
Available online 15 November 2023

the composition of the waste feed or by malfunctions in FGT components can lead to a loss of control of pollutant emissions, which may result in exceeding the emission limit values (ELV). Therefore, it is critical to ensure that FGT systems are robust against unwanted events, thus safeguarding WtE systems with respect to the risk of environmental damage deriving from ELVs exceedance. However, specific methodologies aimed at assessing and managing such risk are still missing.

The chemical and process industry has developed and consolidated risk management techniques based on extensive experience in managing hazardous substances and safety-critical operating conditions (Khan et al., 2015). Many of these techniques have become routine in risk management and have been included in standards and guidelines (Delvosalle et al., 2006; International Organization for Standardization, 2019, 2018). Nevertheless, these methods are not specifically conceived to evaluate and improve environmentally critical systems, and their application to such systems is not straightforward. As an example, the conventional approach towards the analysis and assessment of the environmental performance of FGT systems in current industrial practice is highly empirical and is based on extensive test run campaigns at the plant, which have a critical limitation in the aforementioned need for continuous compliance with strict emission limits. Thus, an alternative perspective is required to address the systematic assessment of critical events that may undermine the performance of FGT in WtE systems.

Approaching the study of environmentally critical systems from a process safety standpoint, the loss of control of pollutant emissions may be considered a top event leading to the exceedance of the ELVs, caused by a set of initiating events (e.g., failures of technical systems). A Bow-Tie diagram may be used to represent such critical scenarios (CCPS and Energy Institute, 2018). Bow-tie diagrams are graphical tools including the causes (i.e., initiating events, on the left side of the diagram), the top event (in the center), and the consequences (on the right side of the diagram) of critical scenarios. Physical and non-physical measures intended to mitigate, prevent, or control such critical scenarios may be considered safety barriers (Sklet, 2006) and are usually represented in Bow-Tie Diagrams. A schematic representation of a Bow-Tie diagram is shown in Fig. 1.

Safety barriers play a key role in ensuring the safety of process operations in safety-critical systems (Liu, 2020), thus may have an important role as well in the safe operation of environmentally critical systems. Several studies address the role and performance assessment of safety barriers in safety-critical systems (e.g. see Landucci et al., 2015 and Misuri et al., 2021). However, to the best of the authors' knowledge, there is no attempt to specifically address the estimation of safety barrier performance in environmentally critical systems, such as FGT plants. Actually, the analysis of the relevant literature, further discussed in the following (see Section 2), highlights two substantial gaps concerning safety barrier evaluation in environmentally critical systems. Firstly, there is little (if any) use of well-established risk management techniques derived from other industrial sectors with extensive experience in risk management (e.g., the chemical and process industry). Secondly, the advent of digitalization and digital technologies allows the development of dynamic and inherently updatable models that may be used for assessing the performance of safety barriers. Yet, such models are hardly used in the field of environmental risk management.

In order to address the gaps evidenced above, the present study aims at presenting a specific innovative methodology combining conventional hazard identification techniques with a digital model of a FGT process in order to identify, simulate, and evaluate safety barriers that may prevent or mitigate excessive emissions in case of process deviations. In the proposed methodology, hazard identification approaches are used to screen possible process deviations and identify the most critical scenarios, which are then simulated using the digital model of the system with or without the application of safety barriers considered for installation. Resilience analysis is then performed to obtain a dynamic measure of the barrier performance under different conditions and barrier configurations. The methodology is demonstrated by its application to a representative case study, addressing the acid gas removal in a WtE facility. Although the detailed procedure required for the application of some steps of the methodology is governed by the specific features of the case-study considered, the overall approach and the structure of the methodology have a general validity, allowing its application to other environmentally critical systems, aiming at the assessment of the effectiveness of safety barriers and the performance tuning of scalable safety barriers.

The remainder of this paper is organized as follows. Section 2 outlines the state of the art of safety barrier performance assessment in relation to dynamic risk assessment (DRA) and resilience engineering, which is the starting point of the developed methodology. Section 3 presents the innovative methodology developed, in combination with its application to the case-study. The reference FGT facility used to test the approach is described in Section 4. Results are presented in Section 5 and discussed in Section 6, which also highlights the limitations of the study and provides suggestions for future developments. Finally, conclusions are drawn in Section 7.

2. Safety barrier assessment in the perspective of DRA and resilience engineering

Regardless of the specific field of application, most contributions estimate the performance of a safety barrier based on a set of indicators, such as barrier effectiveness and availability (Sklet, 2006). The effectiveness of a safety barrier represents its "ability to perform a safety function for a duration, in a non-degraded mode and in specified conditions" (De Dianous et al., 2004), while the availability represents the ability to perform its function while needed. Several studies focused on

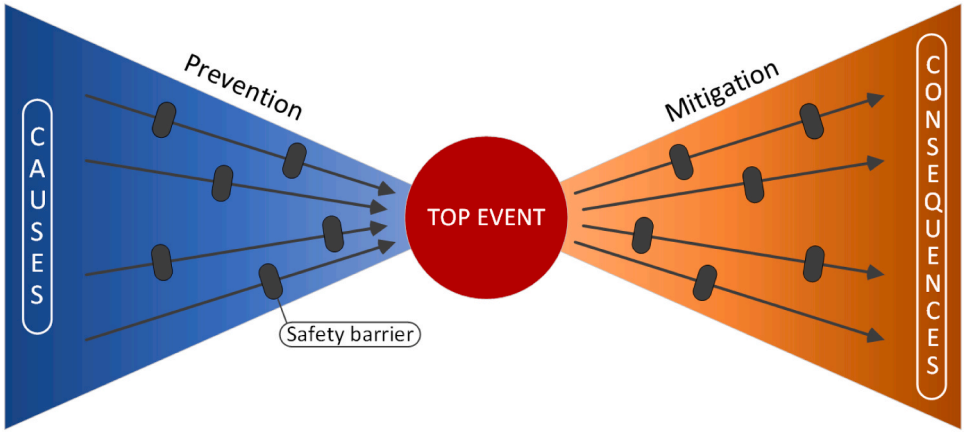


Fig. 1. A Bow-Tie diagram.

estimating the performance of safety barriers (Liu, 2020). For example, Landucci et al. (2015) proposed a method for the quantification of the effectiveness and availability of safety barriers during domino scenarios triggered by fire. The study was further refined by Bucelli et al. (2018) to consider the influence of harsh climate conditions in offshore facilities. Misuri et al. (2021) investigated the impact of Natural Hazards Triggering Technological Accidents (Natech) on barrier performance and proposed a method to modify the Probability of Failure on Demand of safety barriers to account for the effect of natural disasters. However, formal techniques treat safety barriers as static objects, with constant effectiveness and availability values. Such a static perspective cannot capture the dynamic features of the processes involved (e.g., degradation, aging, overlooked hazards). In fact, most canonical methods in risk management are not designed to be easily updatable (Paltrinieri and Khan, 2020), which implies their inability to reflect the evolving real-world risk. The inherently static nature of conventional Risk Assessment methods has been often criticized by academics and practitioners, and is of specific concern in some environmentally critical systems, as WtE, which are inherently exposed to relevant modifications of operating conditions in time. Moreover, the advancements in industrial automation and robotics have increased the complexity and interconnectedness of industrial plants (Villa et al., 2016). In an attempt to overcome these limitations, methods for safety barrier assessment have been directed towards the so-called Dynamic Risk Analysis (DRA), which deals with the development of methods that can provide the update of risk figures considering the variations in the performance of safety-critical systems, such as the control and alarm systems, safety barriers, and maintenance activities (Khan et al., 2016). In the context of DRA, safety barriers are no longer considered static units but dynamic entities, that interact with and are affected by a dynamic environment, and whose performance varies over time due to changes in internal and external conditions (Bubbico et al., 2020). Therefore, DRA aims to define methods and frameworks that are inherently updatable in order to consider new information and capture unsafe operating conditions among highly connected systems. A survey of existing literature indicates that there are only a few DRA methodologies that specifically address the dynamics of safety barriers. For instance, Han et al. (2019) employed Bayesian Networks to model the failure rate of safety barriers. They utilized historical failure data to establish a prior distribution for the barrier failure rate, which was eventually updated as new data emerged. Similarly, Sarvestani et al. (2021) applied Bayesian reasoning to assess the risks associated with LPG accidents. Also, Zeng et al. (2020) employed Bayesian Networks to trace the spatial-temporal progression of fire-related domino effects, integrating the influence of safety barriers directly into the network structure. However, such approaches often necessitate a significant amount of data, in particular concering system failures, for network calibration. Given the infrequent occurrence of such events, obtaining these data is challenging. Furthermore, expert elicitation is commonly used to determine probability distributions, introducing an additional layer of uncertainty.

To the best of the authors' knowledge, there has not been a dedicated study addressing the dynamic evaluation of barrier effectiveness within environmentally critical facilities. In this context, resilience engineering has gained significant importance among safety scientists, motivated by the need to manage risk in increasingly complex systems (Bergström et al., 2015). Similarly to DRA, resilience analysis focuses on capturing risk variability due to component failures, external disturbances, and/or dysfunctional interactions among system components (Leveson et al., 2006). However, resilience puts more emphasis on the intrinsic ability of a system to "adjust its functioning prior to, during, or following changes and disturbances, so that it can sustain required operations under both expected and unexpected conditions" (Hollnagel et al., 2011). That is, resilience engineering approaches system safety from a slightly different perspective, which focuses on how systems absorb sudden disturbances, recover after disruptive events, and adapt to new conditions while maintaining acceptable performance (Yarveisy et al., 2020). Several studies have focused on resilience analysis to address safety of complex socio-technical systems (Patriarca et al., 2018). However, only a few contributions leverage resilience engineering to evaluate the performance of safety barriers (Bai et al., 2022; Sun et al., 2021; Thieme and Utne, 2017). In addition, no study has been proposed to address the safety of environmentally critical systems from a resilience perspective.

DRA and resilience analysis rely on updatable models that can (i) grasp the system dynamics and (ii) consider the effects of unsafe interactions. However, the increasing complexity and interconnectedness of industrial plants prevent the development of rigorous modeling. For example, it is challenging to describe the dynamics of the acid gas neutralization mechanism occurring in an FGT plant through first principles due to the complexity of the phenomena involved (e.g., convection, diffusion in a solid porous media, reaction kinetics, thermodynamic equilibria) and other external and internal factors that are difficult to control and monitor (e.g., thickness and reactivity of the filter cake, the composition of the flue gas, changes in the sorbent structure) (Antonioni et al., 2016; Dal Pozzo et al., 2018b). In this context, the emergence of digitalization in process industry can provide tools and methods to overcome such limitation (Kockmann, 2019). Thanks to the wealth of data typically available from plant sensors and measurement devices, it is possible to derive digital models of the pollutant removal processes of varying degree of complexity that can be used for process optimization purposes. Reliable data-driven models of different operations in the WtE flue gas cleaning can be developed from representative datasets of past performance of the plant (Magnanelli et al., 2020; Pozzo et al., 2018) or compact test protocols (Bacci Di Capaci et al., 2022). Dal Pozzo et al. (2021) demonstrated how the use of a properly calibrated digital model reproducing the behavior of a real FGT system enables an extensive testing of alternative control strategies in a virtual environment. The final application of the optimized control strategy tuned via the digital model to the real plant showed a significant reduction compared to the default control logic of the plant. The approach allowed to achieve such process control optimization with a minimal need for test runs at the real plant.

Therefore, based on the above analysis of the relevant literature, the method developed in the present study introduces a dynamic evaluation of the barrier performance, allowing its update based on new data and knowledge becoming available during process operation. Moreover, a specific approach based on a digital model of the FGT system is developed to allow testing the limits of system performance in a virtual environment, limiting the use of full-scale test-runs that may lead to hazardous conditions when approaching critical emission values. Finally, resilience analysis is applied to obtain a dynamic measure of the barrier performance.

3. Methodology

The approach proposed in this study is composed of six steps, which are outlined in Fig. 2 together with their inputs and outputs. The methodology relies on the integration between advanced risk management tools (e.g., hazard identification techniques) and innovative modeling methods (e.g., data-driven regression models). The former are used to define a set of critical scenarios and additional safety barriers that may prevent or mitigate such critical events. The latter allow the simulation of critical scenarios and of safety barrier performance without the need for field tests or first principles models. A detailed description of each step included in the methodology is given in the following. For the sake of clarity, the specific steps of the methodology addressing digital model development and safety barrier modeling are developed addressing the features of FGT systems in WtE, for which a case-study will be discussed in the following.

3.1. Process layout definition and data collection

In this step, relevant information on the process considered must be

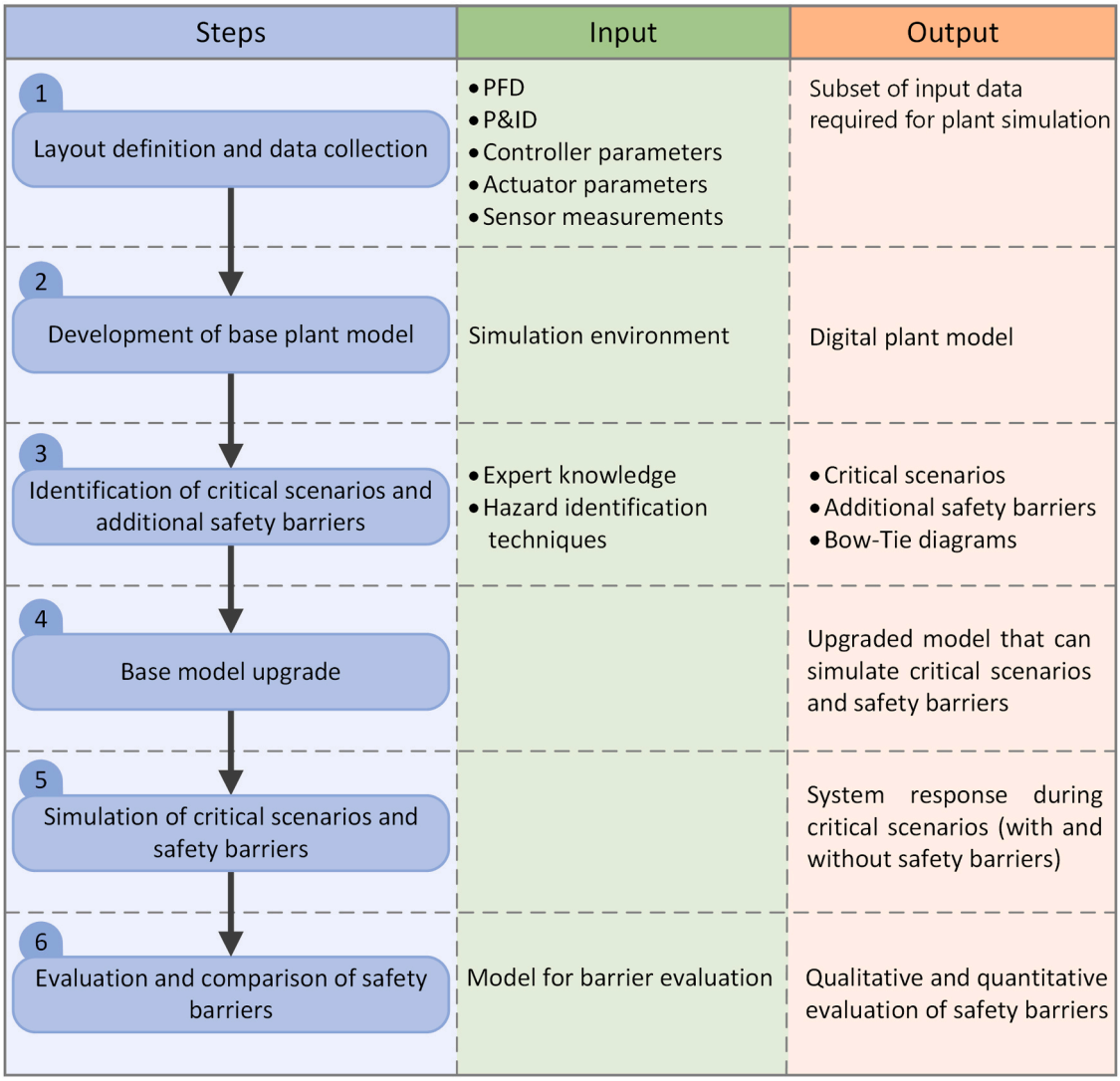


Fig. 2. Overview of the methodology.

collected and stored. The minimum set of data includes the following:

1. Process Flow Diagrams (PFD) and Piping and Instrumentation Diagrams (P&ID);
2. Parameters of the control loops;
3. Process data collected during different operating conditions.

PFD and P&ID are required to determine the process layout and the control strategy.

When considering a typical FGT section, this includes filters (i.e., fabric filters or electrostatic precipitators), reactors (e.g., spray driers, scrubbers, in-line reactors), injection devices (e.g., screw feeders), and measurement devices (e.g., thermocouples, flow meters, gas analyzers). An overview of the most used techniques for the reduction of acid gases is reported in Section 2.5.4 of the Best Available Techniques (BAT) Reference Document for Waste Incineration (European Commission, 2019).

In addition, it is critical to determine the control strategy adopted to regulate the injection of sorbent (e.g., feedback, feedforward, mixed hybrid control strategies). After the identification of the control strategies, the design parameters of controllers and actuators must be collected. That is, input-output models or, alternatively, transfer functions of controllers and actuators must be defined in terms of mathematical structure and parameters. This information may be provided by the plant personnel or may be available in technical manuals.

Finally, process data from various operating conditions must be collected and stored. These data are required to build the data-driven model of the acid gas reduction mechanism. Therefore, it is critical to ensure that data are closely related to the reaction dynamics. With reference to Fig. 3, representing a general scheme of a FGT, the minimum set of process data may include:

- The concentration of acid gases in the flue gas entering the system (stream number 1 Fig. 3), namely $C_{acid,in}$.
- The concentration of acid gases in the clean gas leaving the system (stream number 3 in Fig. 3), namely $C_{acid,out}$.
- The mass flow rate of the sorbent (stream number 2 in Fig. 3), namely $\dot{m}_{sorbent}$.

Data come in the form of time series representing the evolution of process variables with time and may be stored in a matrix-like database D, whose columns represent process variables and rows indicate time instants.

It is worth mentioning that the type and number of process variables available for collection and the total amount of observations largely depend on the specific application. Actually, different plants have different sensors and measuring points. However, the process variables mentioned above should be easy to obtain in most facilities (directly or

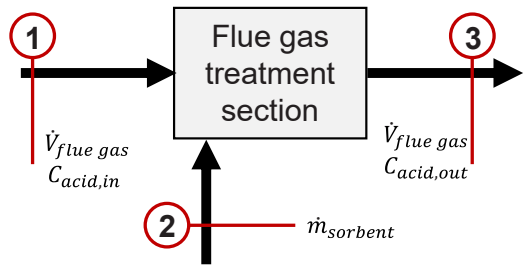


Fig. 3. General schematization of a flue gas treatment section for acid gas removal.

derived from other measured variables).

The data collection process aims to capture the plant behavior during various operative conditions, including normal operations, disturbances, and malfunctions. Thus, the dataset D should encompass a wide range of values for the process variables, maximizing information content. It is thus crucial to adequately capture the variability and diversity of the process conditions to enable an accurate and robust model development. Sufficiently representative time-series should be available, in the range of weeks up to month, depending on the features of the plant. Actually, most distributed control systems nowadays store long time series of process data (up to several years), thus providing sufficient information for the implementation of the method.

Furthermore, process data should be sufficiently granular to allow a thorough investigation of the dynamics involved in the processes. An adequate level of granularity is essential to capture accurately the intricate temporal variations and interactions within the system. A minimum granularity, in terms of sampling time, of 1 min is usually adequate to ensure that the data capture the necessary temporal resolution, enabling a detailed analysis of the processes' dynamics and facilitating accurate modeling.

3.2. Development of a base plant model

The base plant model ($\mathscr{M}$) is a digital model of system of concern. In the present study, the FGT section of a WtE was considered. The purpose of the model is to mimic the plant dynamics in terms of (i) control logic and actuators and (ii) acid gas reduction. In other words, the model takes as an input the concentration of acid gases in the flue gas at time t (i.e., $C_{acid,in}(t)$), and returns the concentration of acidic compounds in the clean gas at time $t+1$ (i.e., $C_{acid,out}(t+1)$):

$$C_{acid,out}(t+1) = \mathscr{M}\left(C_{acid,in}(t)\right) \tag{1}$$

The digital model comprises several sub-models that mimic a specific plant function. For example, there may be sub-models to replicate the controller behavior, measuring instruments, the reaction dynamics, and so forth. The number and nature of the sub-models largely depend on the specific plant under consideration. The analysis of PFDs and P&IDs is essential to define the structure of $\mathscr{M}$. In most plants, the digital model comprises at least three sub-models:

- The sub-model g that mimics the actuator;
- The sub-model f that mimics the controller action;
- The sub-model h that mimics the reaction dynamics.

In this case, Eq. (1) may be written as follows.

$$C_{acid,out}(t+1) = h\left(g(f(t)), C_{acid,in}(t)\right) \tag{2}$$

Where $f(t)$ represents the controller signal at time t, and $g(f(t))$ indicates the manipulated variable at time t (e.g., $\dot{m}_{sorbent}(t)$).

Data collected in step 1 of Fig. 2 allow the rigorous modeling of actuators and controllers. However, modeling the reaction dynamics through first principles is challenging. A viable solution to model the reaction mechanism is to rely on data-driven methods. Here, the idea is to leverage plant data collected in step 1 in Fig. 2 to build a data-driven model of the acid gas reduction mechanism. This model may take as an input (i) the concentration of acid gases entering the system and (ii) the sorbent flow rate, and return the concentration of acidic compounds in the clean gas leaving the system.

The problem described in Eq. (2) belongs to the vast area of time-series forecasting (Box et al., 2015). Therefore, h may be considered a regression model that takes a set of observations as an input and returns the value of a target variable. The selection of the model h is a critical step to ensure adequate performance (Emmert-Streib and Dehmer, 2019). However, a complete overview of available models and model selection techniques is unfeasible considering the vastity of the topic. The reader might refer to the literature on system identification (Ljung, 2010, 1999) and data mining (Kotu and Deshpande, 2019; Torres et al., 2020) to explore different modeling strategies.

Regardless of the specific model, the development of h involves at least two steps: training and evaluation. Firstly, the dataset D (i.e., process data collected in step 1 of Fig. 2) is split into two parts, namely D_t and D_e, such that $D = \begin{vmatrix} D_t \\ D_e \end{vmatrix}$. D_t is used to train the model while D_e is used in the evaluation phase. Typically, D_t contains 80% of the observations in D. Also, D_t may be conceptually divided into two parts. The first part (X_t) comprises the inputs of the model (i.e., $C_{acid,in}(t)$ and $\dot{m}_{sorbent}$), the second part (Y_t) comprises the variable that must be predicted (i.e., $C_{acid,out}(t+1)$), such that $D_t = |X_t\ Y_t|$. The same applies to D_e.

Secondly, the model is trained. Training involves the optimization of the model's internal parameters (θ) to minimize a loss function $(\mathscr{L})$.

$$\widehat{\theta} = \underset{\theta}{argmin}[\mathscr{L}(\theta, D_{train})] \tag{3}$$

Where $\widehat{\theta}$ represents the optimized model's parameters. Some widely used loss functions include the Sum of Squared Residuals (SSR), the Mean Squared Error (MSE), Mean Absolute Error (MAE), Hubert Loss, and Log-cosh loss (Wang et al., 2022). As an example, if SSR is used, the loss function is:

$$\mathscr{L}(\theta, D_{train}) = \sum_{i=1}^{N} (y_t(t+1) - h(x_t(t), \theta))^2 \tag{4}$$

Where $y_t(t+1) \epsilon Y_t$, N represents the number of observations in D_t, and $x_t(t) \in X_t$.

After training, the performance of the model must be evaluated using a new set of data. To this end, the model is used to perform predictions on the observations included in D_e. Eventually, performance indicators are calculated to quantify the prediction performance. For example, the Root Mean Squared Error (RMSE) may be calculated as follows:

$$RMSE = \sqrt{\frac{1}{M}\sum_{i=1}^{M} (y_e(t+1) - h(x_e(t), \widehat{\theta}))^2} \tag{5}$$

Where, $y_e(t+1) \epsilon Y_e$, M represents the number of observations in D_e, and $x_e(t) \in X_e$.

A user-defined acceptance criterion may be defined to discriminate between acceptable and non-acceptable performance. For example, if the RMSE is smaller than a threshold, the model h may be considered adequate to simulate the reaction dynamics.

It is worth mentioning that the procedure described above is intended to be a quick overview of the steps required to train and evaluate the model. It is not meant to be the best strategy. For example, the so-called holdout method is described above to keep the description short. The reader may adopt more advanced evaluation methods, such as holdout with validation or cross-validation (Raschka, 2018). Also, Eq. (2) assumes that the input to the data-driven model are inlet concentration of acid gases and sorbent mass flow rate. Nevertheless, the method may be promptly adapted to consider more input data, such as the flue gas temperature, the flue gas volumetric flow rate, the pressure drop across

the filters, and so forth.

3.3. Identification of critical scenarios and additional safety barriers

The FGT section is analyzed to identify (i) critical scenarios and (ii) recommendations for additional safety barriers. In this context, critical scenarios are events that have the potential to cause a significant increase in the acid gas concentration downstream of the treatment section. In other words, the analysis aims at answering the following questions:

1. Which critical events have the potential to cause a significant increase in acidic compounds in the clean gas leaving the treatment section (stream 3 in Fig. 3)?
2. Which additional safety barriers may prevent or mitigate critical events?

Traditional hazard identification techniques, such as HazOp, HazId, analysis of historical data, what-if analysis, and brainstorming (Mannan, 2005), may be used to answer these questions. Data collected in step 1 in Fig. 2 (e.g., PFD, P&ID) and the operational experience of plant personnel are the starting point of the analysis.

The selection of the actual hazard identification technique to be applied is guided by several factors, such as time constraints, objectives of the analysis, and the required level of detail (International Organization for Standardization, 2019; Mannan, 2005). Structured techniques (e.g., HazOp) provide more information and a deeper understanding of the hazards. On the other hand, unstructured methods (e.g., brainstorming) are faster and cheaper, but a higher level of expertise may be required to ensure the quality and completeness of results.

Often, the combined use of multiple hazard identification techniques leads to a more comprehensive risk identification (International Organization for Standardization, 2019). However, regardless of the specific techniques adopted, the results of the analysis should provide:

- A list of critical events, along with their causes and consequences on acid gas emission at stack;
- A list of recommended safety barriers.

Results may be condensed in a bow-tie diagram to provide a concise visual representation of critical events and safety barriers (CCPS and Energy Institute, 2018). The top event may be formulated as a “significant increase in the acidic compound concentration in the clean gas”. Among the end-point events on the right-end part of the diagram (consequences) specific possible outcomes of the top event should be listed (e.g., “half-hourly emission limit values exceeded”), while the leftmost part shows the causes of the top event. Recommended safety barriers should be included in the bow-tie to clarify their role in preventing or mitigating the critical event.

3.4. Base model upgrade

The base plant model developed in step 2 of Fig. 2 is designed to mimic the plant response during normal operating conditions. Therefore, modifications may be needed to simulate the effect of critical events and additional safety barriers identified (step 3 in Fig. 2).

Depending on the nature and extent of modifications, there are two viable solutions to update the base plant model. These include:

1. First principles modeling;
2. Data-driven modeling using data from test-runs.

If the modifications are associated with well-known systems where first principle models are available, it is possible to employ rigorous modeling techniques to incorporate the behavior of critical events and safety barriers. For example, if a critical event involves the failure of a control loop, the equations governing the controller can be modified to account for the faulty behavior.

However, when the effect of critical events and additional safety barriers is uncertain and cannot be accurately described using rigorous models, data-driven methods may be used. Data from different facilities that have experienced similar failures or implemented similar safety barriers may be used to this aim. Clearly enough, if limited data are available, carrying out specific test runs on pilot facilities or on the actual plant may be considered as an alternative in case the safe operation of the system may be granted. The reader is referred to previous studies (Bacci Di Capaci et al., 2022; Dal Pozzo et al., 2021) for details and discussion on the design of data collection campaigns for WtE flue gas cleaning systems.

Regardless of the particular updating procedure, the simulation of critical events and safety barriers necessitates the modification of existing sub-models or the development of new sub-models. This process enables the creation of an upgraded model, denoted as $\mathscr{M}''$, that can (i) simulate the effect of the critical events on the original gas treatment system and (ii) simulate the system response after the installation of all (or part of) the recommended safety barriers.

3.5. Simulation of critical scenarios and safety barriers

The upgraded model $\mathscr{M}''$ is used to simulate the critical events identified in step 3 of the methodology (see step 5 in Fig. 2). Two distinct simulation runs are performed.

The first run aims at evaluating the response of the original gas treatment system during critical scenarios. That is, all the barrier sub-models are excluded in this first run of simulations. In this phase, each critical event identified in step 3 of Fig. 2 is simulated to obtain the trends of $C_{acid,out}(t)$ and $\dot{m}_{sorbent}(t)$, describing the original plant response in the presence of critical disturbances. This first set of simulations is used as a benchmark to evaluate the improvements due to the implementation of the additional safety barriers.

The second group of simulations focuses on the system response after the installation of the safety barriers identified in step 3 of the procedure (see Fig. 2). To this end, the bow-ties produced are analyzed to identify relevant safety barriers for each critical event considered. Safety barriers are selected based on their ability to affect the operation of the specific critical event under consideration. As a result, a set of safety barriers is selected for each critical event. The upgraded plant model is then used to simulate the effect of safety barriers considering that only part of the barriers may be active during a critical event. That is, if a critical event is associated with a set of N safety barriers, the number of simulations required is $2^N - 1$. In each simulation, the model returns $C_{acid,out}(t)$ and $\dot{m}_{sorbent}(t)$, which are used to quantify the improvements due to the implementation of the safety barriers.

3.6. Evaluation and comparison of safety barriers

The output of the simulations provides a dynamic picture of the system behavior with different barrier configurations and during various critical events. These results can be used to evaluate the effectiveness of safety barriers, in both absolute and relative terms. In this context, the general definition of effectiveness introduced in Section 2 has to be declined for the specific problem of emission control as the ability of a safety barrier to ensure that the system complies with ELV. A set of indicators is built to evaluate the barrier effectiveness and allow for a quantitative comparison of alternatives. Resilience analysis is used to quantify the ability of the system to withstand external disturbances and to evaluate the improvements resulting from the installation of additional safety barriers.

Following the generic definition of resilience provided by Hollnagel et al. (2010), the resilience of the gas treatment system may be defined as its ability to fulfill its purpose in a variety of adverse conditions. In the

specific context of acid gas removal, the *purpose* of the system is to comply with ELV, and the *adverse conditions* refer to the critical scenarios identified in step 3 of the method (see Fig. 2).

The literature offers many examples of quantitative resilience metrics (Hosseini et al., 2016). Most of them rely on a time-dependent function $\varphi(t)$ that reflects the performance of the system. This performance function (also called quality function) ranges between zero and one. The performance is zero if the system is in a completely degraded state or, in other words, if it cannot fulfill its purpose. On the contrary, if the system performs as expected, the performance is one.

After a critical event, the system performance degrades, reaches a minimum, and eventually increases as mitigative actions restore normal operations, as exemplified in Fig. 4.

The mathematical formulation of the performance metric depends on the problem under assessment. The performance of the treatment system with respect to the acid compound i (i.e., $\varphi_i(t)$) may be a user-defined function that satisfies the following requirements:

- $\varphi_i(t) = 0$ if the half-hourly concentration of i exceeds the ELV;
- $\varphi_i(t) = 1$ if the absolute deviation between the half-hourly concentration of i and the controller setpoint does not exceed 10%.

The user can choose the type of function expressing the performance based on the problem requirements. The criteria for the selection of performance functions are discussed extensively elsewhere (Hosseini et al., 2016; Tran et al., 2017). In the case-study introduced in the following, an exponential function was used to penalize large deviations from the controller setpoint (see Section 4).

Given the performance, the so-called Resilience Loss (RL) can be used to quantify the loss of resilience caused by a critical event.

$$RL_i = \int_{t_e}^{t_f} [1 - \varphi_i(t)]\,dt \tag{6}$$

Where RL_i indicates the Resilience Loss with respect to the acidic compound i, t_e represents the time of occurrence of the critical event, and t_f is the recovery time, as indicated in Fig. 4. The Resilience Loss ranges between zero and $(t_f - t_e)$. Values close to zero indicate that the system has not been significantly affected by the critical event.

A performance function and a Resilience Loss can be calculated for each acid gas considered and each combination of critical scenarios and safety barriers. The performance metric $\varphi_i(t)$ reflects the system dynamics during the critical scenarios, while RL represents a quantitative indicator that reflects the system capacity to withstand internal or external disturbances. The comparison of $\varphi_i(t)$ and RL_i among alternative configurations and to the benchmark simulations allows a quantitative comparison between alternative process configurations and the identification of the best-performing safety barriers.

It is worth mentioning that other relevant features of the safety barriers, namely their availability and level of confidence, which are related to the reliability and availability of mechanical components and not to their process performance, are considered out of scope of the present analysis.

4. Case study

A full-scale case study was defined to demonstrate the application of the methodology and the potential use of the results obtained. The case study concerns an acid gas removal stage of the FGT section of Municipal Solid Waste Incinerator located in northern Italy.

In WtE operation, the concentrations of hydrogen chloride (HCl) are typically higher of at least an order of magnitude than those of SO_2 and HF (Dal Pozzo et al., 2023a). Hence, for the sake of simplicity, in the case-study only HCl removal will be considered, since in the current practice fulfilling the ELV of HCl is more critical.

A process flow diagram of the specific FGT system considered in the case-study is reported in Fig. 5. As shown in the Fig. 5, the flue gas (stream 1 in Fig. 5) enters an entrained flow-reactor where a solid sorbent (hydrated lime - stream 9) is injected into the flue gas. The entrained-flow stream of gas and solids (stream 2) enters the bag filter F-01, where solids (stream 4) are removed from the clean gas (stream 3). HCl in the gas stream is neutralized according to the following reaction:

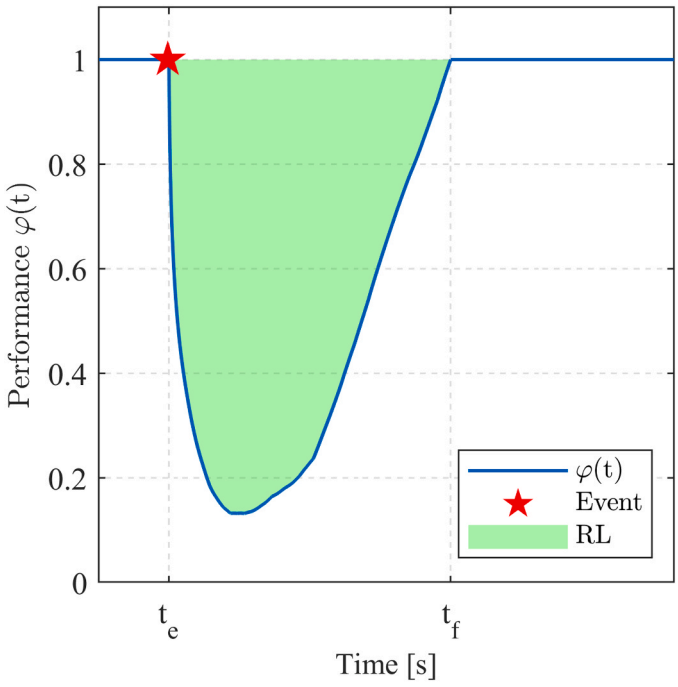


Fig. 4. Typical behavior of the FGT system following a critical event.

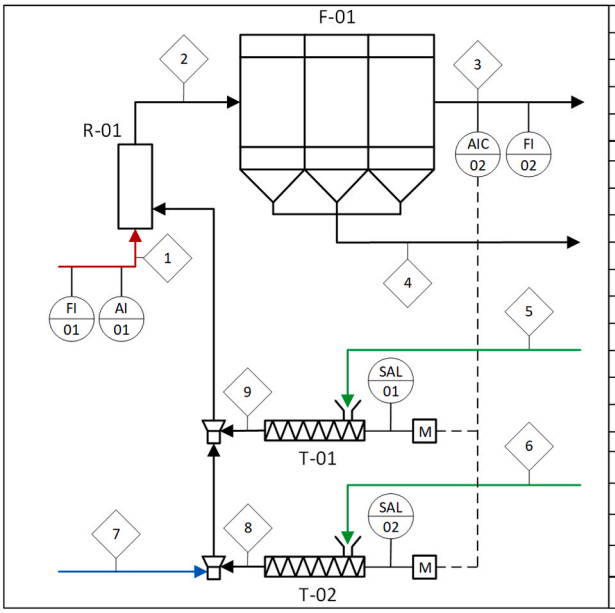
$$2HCl + Ca(OH)_2 \leftrightarrow CaCl_2 + 2H_2O \tag{7}$$

The gas-solid reaction takes place in the entrained flow reactor (R-01) and in the cake formed on the filter bags (F-01).

The sorbent mass flow rate is controlled by means of a simple feedback control loop. Specifically, a PI controller (AIC 02) is used to regulate the speed of the feeder motors (M) based on the concentration of acidic compounds in the clean gas leaving the system (stream 3). Two screw feeders are installed in parallel. During normal operations, only one of the two screw feeders operates (T-01), while the other is used as a backup during maintenance or in case of failure of the main feeder. Low-speed alarms (SAL) are installed to detect a blockage or failure of the feeder, allowing a swift start-up of the backup feeder by the control room operator. The configuration shown in Fig. 5 is among the solutions most frequently installed for acid gas removal in European incinerators according to recent surveys (Beylot et al., 2018; Dal Pozzo et al., 2018a) and is listed among the BAT for acid gas treatment (European Commission, 2020). Thus, the case-study introduced is highly representative of the current industrial practice.

The methodology outlined in Section 3 was applied to the analysis of the case study. First, the relevant documentation concerning the selected facility was collected, as indicated in step 1 of the methodology (see Fig. 2). Specifically, the plant personnel provided PFDs, P&IDs, Operating and Control Philosophy, and details on the controller and actuator parameters. Furthermore, a data collection campaign was designed and performed to extract relevant process data from the plant Distributed Control System (DCS). In particular, the following process variables were collected with a sampling interval of 30 s:

- Volumetric flow rate, temperature, and HCl concentration of the flue gas from the furnace (stream 1 in Fig. 5);



Equipment	Description
R-01	Reactor
F-01	Bag filter
T-01	Main screw feeder
T-02	Backup screw feeder
Stream	**Description**
1	Flue gas from the furnace
2	Gas-solid stream (flue gas and sorbent)
3	Clean gas
4	Solid waste
5	Sorbent feed ($Ca(OH)_2$)
6	Sorbent feed ($Ca(OH)_2$)
7	Compressed air
8	Injected sorbent
9	Injected sorbent
Instrumentation	**Description**
FI	Flow indicator
AI	Composition indicator
AIC	Composition controller
SAL	Slow speed alarm

Fig. 5. Process Flow Diagram of the FGT system for acid gas removal considered in the case-study. Red, green, and blue streams respectively indicate the flue gas entering the FGT section, the sorbent feed, and the compressed air used to inject the sorbent.

- Volumetric flow rate, temperature, and HCl concentration of the clean gas (stream 3 in Fig. 5);
- Mass flow rate of the sorbent (stream 9 in Fig. 5).

A total of four days of observations were collected and stored in tabular format. The collected data were selected to maximize the information stored in data, ensuring the adequacy and significance of the collected data. Matlab Simulink was used to develop the base plant model. Fig. 6 shows the model structure as it appears in the simulation environment. Specific sub-models were developed for each of the equipment items present in the process flow diagram of the plant section considered, shown in Fig. 5.

The input to the base plant model is the molar flow rate of HCl entering the gas treatment system (stream 1 in Fig. 6). The "DCS" block mimics the controller behavior (i.e., AIC-02 in Fig. 5), returning the controller command (signal 2 in Fig. 6) based on the outlet HCl concentration (signal 5 in Fig. 6). The "Screw feeder" block mimics the actuator behavior. It converts the command from the controller into the sorbent mass feed rate injected in the reactor (stream 3 in Fig. 6). Finally, the "Reaction" block represents the data-driven model of the acid neutralization mechanism. Specifically, the model used in this study is a linear Autoregressive with Extra Input model (ARX). The "Reaction" block takes as an input the sorbent feed rate and the molar flow rate of HCl in the flue gas (stream 1 in Fig. 6), and returns the molar flow rate of HCl in the clean gas (stream 4 in Fig. 6), which is eventually converted into the concentration of HCl leaving the system (signal 5 in Fig. 6). Further details on the base plant model used in this study are reported elsewhere (Dal Pozzo et al., 2021).

HazOp analysis has been used to identify critical events that may lead to a significant increase in HCl emissions and the safeguards and/or safety barriers to be installed.

Although the list of critical events identified through the HazOp represents a detailed description of the potential hazards present in the system, some of them may not be credible or may have a marginal impact on HCl emissions. Provided that quantitative information on the causal analysis of FGT systems failure is unavailable in the open literature, an expert elicitation procedure was adopted to complement the HazOp analysis and validate the most relevant process deviations. Expert surveys have been recognized in literature as a relevant tool for a preliminary semi-quantitative evaluation of hazards and related safety barriers (Argenti et al., 2017; Hokstada et al., 1998; Misuri et al., 2020). An ad-hoc survey was prepared and administered to a group of experts with heterogeneous and relevant backgrounds (WtE plant operators, technology suppliers, consultants, academics) that were invited to

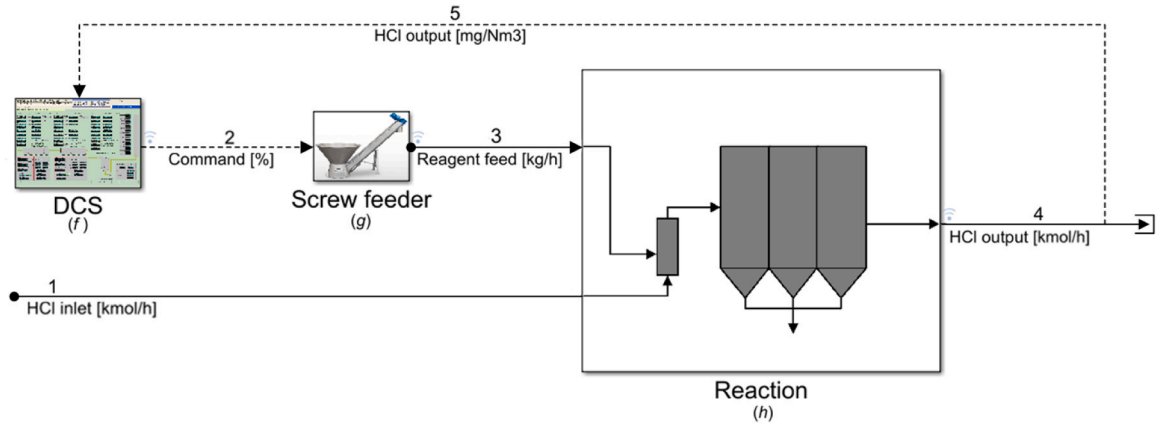


Fig. 6. Translation of the reference FGT system in Fig. 5 into the simulation environment. Items *h*, *g*, and *f* respectively indicate the submodels that mimic the reaction mechanism, the screw feeders, and the control logic. Dashed lines indicate signals and continuous lines represent process streams.

participate anonymously. Considering the specific process scheme in Fig. 5, the experts were asked about the likelihood that given process deviations could trigger a loss of control event, resulting in a temporary overrun of the emission setpoint at stack (sufficient or not to exceed the half-hour emission limit value for the plant). Next, they were asked about the likelihood that given prevention or mitigation measures could avoid such loss of control events. The experts were able to express their answers on a scale 1–5, corresponding to a verbal scale of likelihoods of occurrence from "Extremely unlikely" (i.e., 1) to "(Virtually) certain" (i.e., 5). The transcription of the questionnaire, along with general data collected on the background of survey participants, are reported in the Supplementary Material. The results of the survey supported the identification of the critical scenarios and safety barriers considered for implementation, as discussed in Section 3.3.

Following the identification of critical scenarios and safety barriers, the base plant model was upgraded (step 4 in Fig. 2), and simulations were performed to evaluate the system response with and without the recommended safety barriers (step 5 in Fig. 2). Specifically, two sets of simulations were executed utilizing the upgraded plant model. The first set of simulations models the behavior of the original FGT system during critical scenarios in the absence of any additional safety barriers. The second set of simulations replicates the system response to critical events after the installation of safety barriers. At the end of each simulation, the upgraded plant model returns $C_{HCl,out}(t)$ the concentration of HCl in the clean gas leaving the plant during different critical events and under different system configurations (i.e., with or without safety barriers).

After the simulations, the results were analyzed to evaluate the consequences of the critical events and the benefits derived from the installation of safety barriers. Specifically, the following performance metric was used to assess the performance of the system selected for the case-study:

$$\varphi_{HCl} = \begin{cases} 1 & if\ \overline{C}_{HCl,out}(t) \leq 7.15\ mg_{HCl}/Nm^3 \\ A \bullet exp\left(-B \bullet \overline{C}_{HCl,out}(t)\right) & if\ \overline{C}_{HCl,out}(t) > 7.15\ mg_{HCl}/Nm^3 \end{cases} \quad (8)$$

Where $\overline{C}_{HCl,out}(t)$ indicates the half-hourly HCl concentration at stack at time t, and 7.15 mg_{HCl}/Nm^3 represents the controller setpoint increased by 10% to allow a limited oscillation of the controlled variable. The parameters A and B have been estimated through least squares minimization with the following boundary conditions: $\varphi(7.15) = 1$ and $\varphi(10) = 0$, where 10 mg_{HCl}/Nm^3 represents the ELV. The fitting procedure leads to $A = 3.360 \bullet 10^7$ and $B = 2.424$, which implies $\varphi(10) = 1 \bullet 10^{-3}$. The formulation of the performance metric was inspired by the understanding that the system ability to sustain external disturbances diminishes quickly as the concentration of HCl approaches the ELV. Therefore, the performance metric is designed to degrade exponentially after $\overline{C}_{HCl,out}(t)$ exceeds the allowed level of oscillations and to approach 0 when $\overline{C}_{HCl,out}(t)$ reaches the ELV.

Based on the above defined performance function, the resilience was calculated for each simulated scenario using Eq. (6), enabling quantitative assessment and comparison of safety barriers.

5. Results

In the following, the application of the methodology outlined in Section 3 to the case study introduced in Section 4 is illustrated.

5.1. Critical scenarios and safety barriers

The results of HazOp analysis, used to identify critical events that may lead to a significant increase in HCl emissions and the safeguards and/or safety barriers to be installed, have been condensed into a bow-tie diagram, which is shown in Fig. 7. It is worth mentioning that the bow-tie has been simplified for visualization purposes. The complete bow tie is reported in Figure A1.

The results of the expert survey are shown in Fig. 8. Specifically, Fig. 8.a reports the results related to the credibility of the critical events identified by the HazOp.

It should be remarked that the interviewees generally considered resilient the system in Fig. 5, as only three process deviations (inlet HCl spike +200%, critical waste composition, and clogging of reactant transport line) were deemed likely to cause a temporary overrun of emission setpoint, and only a single deviation (clogging of reactant transport line) was considered likely to cause an overrun severe enough to exceed the half-hour ELV. Among process deviations related to inlet flue gas composition, spikes of HCl were considered significantly more likely. This finding is in agreement with the high HCl to SO_2 ratio typically found in waste-to-energy flue gases (Dal Pozzo et al., 2016) and supports the assumption to consider only HCl in the assessment (see Section 4). The clogging of the reactant transport line was considered the most critical process deviation, followed by failure/blockage of the screw feeder. However, it is worth noticing that an obstruction of the screw feeder was identified by the experts as the most frequent failure experienced in these systems (see section S2 of the Supplementary Material).

Combining the information coming from the HazOp analysis and the expert survey, two critical scenarios were selected for the analysis:

- Critical scenario 1: spike in inlet HCl concentration;
- Critical scenario 2: failure of the screw feeder for reactant delivery.

The survey allowed gathering information also on the effectiveness of possible safety barriers in the critical loss of control of acid gas emission scenarios discussed above. As shown in Fig. 8, the experts were

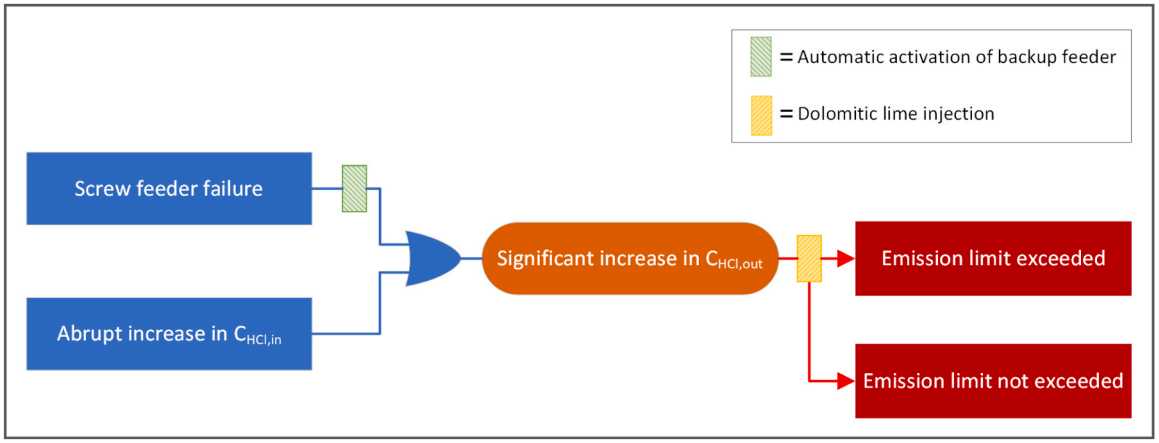


Fig. 7. Simplified bow-tie diagram of the reference FGT system considered in the case-study.

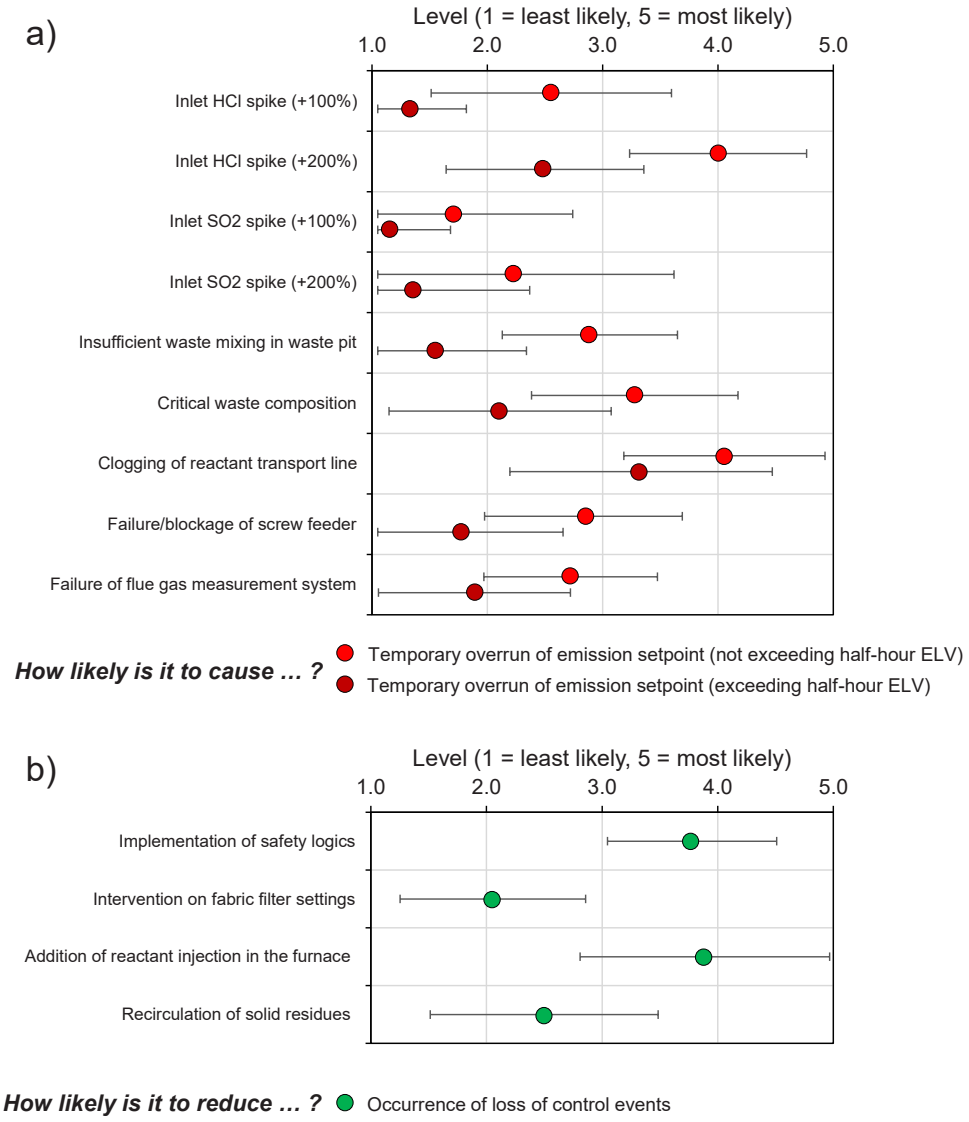


Fig. 8. Results obtained from the expert survey concerning: a) the likelihood of the critical process deviations identified by HazOp to generate loss of control events; b) the likelihood of the listed safety barriers to mitigate loss of control events. Numerical scale (1−5) to be interpreted as in section S2 of the Supplementary Material.

Table 1
List of the safety barrier types considered in the survey.

Type of safety barrier	Description
Intervention on fabric filter settings	Increase of the maximum allowable pressure drop at the fabric filter. **Effect:** *fabric filter cleaning is stopped, allowing longer residence time of the reactant on filter bags and a temporary increase of reactivity in the system.*
Recirculation of solid residues	Re-injection upstream of the fabric filter of part of the process residues collected by the filter. **Effect:** *solid process residues, partially unreacted, which are normally sent to disposal, are recirculated, increasing the overall sorbent-to-acid gas ratio in the system.*
Implementation of safety logics	Implementation of improved safety logics and backup safety systems (e.g., safety logics activating start-up of backup elements). **Effect:** *failure of any element in the control loop triggers the intervention of a backup system that maintain the required feed rate of sorbent:* e.g., *automatic activation of backup sorbent feeders in case of fault of the primary feed control loop.*
Addition of reactant injection in the furnace	Pre-treatment of flue gas in an additional reaction stage upstream of the existing FGT system. **Effect:** *reactant injection in an additional injection point upstream of the FGT system is activated, curtailing spikes of acid gases coming from the combustion chamber before they enter the FGT system.*

asked to evaluate a set of safety barriers, assessing their likelihood to reduce the occurrence of loss of control events and to reduce the consumption of reactant required to mitigate such events. The safety barriers considered in the survey are listed in Table 1.

The first two safety barriers in Table 1 share the common rationale of increasing the residence time of the solid reactant in the system, hence inducing higher solid conversion and increasing HCl removal at equal reactant consumption (Chibante et al., 2010). Although these interventions can help avoiding an excessive consumption of reactants in the control of HCl emissions, the experts consider these systems scarcely effective in reducing the frequency of loss of control events.

Higher scores in terms of likelihood to reduce the occurrence of loss of control events were given to measures that increase redundancy in the FGT system: addition of a pre-treatment HCl removal stage in the furnace (mean score 3.9), and implementation of safety logics (mean score 3.8). The former class of measures is aimed at controlling the effects of high acid gas loads from waste combustion (e.g., critical scenario 1 identified in Section 5.1), while the latter is mainly focused on mitigating the effects of failures of system components (e.g., critical scenario 2 identified in Section 5.1). Therefore, a safety barrier for each of the two classes of interventions was selected as an example for the simulation.

In the case of critical scenario 1, HCl peaks from waste combustion can effectively be mitigated by furnace sorbent injection (Biganzoli et al., 2015). The injection of dolomitic lime in the furnace, a widely applied retrofit solution to improve FGT performance (Dal Pozzo et al., 2023b), was considered for application.

In the case of critical scenario 2, the installation of a safety logic for the automatic activation of the backup feeder by a low-speed alarm was considered to mitigate the possible failure of the main screw feeder of the solid sorbent. It was assumed that such configuration can activate the backup screw feeder in 15 s, compared to at least 5 min in case of a manual intervention by plant operators, which is considered as the base case.

5.2. Base model upgrade

As discussed in Section 3.4, some modifications were introduced in the base plant model described in Section 4 in order to simulate the critical events and the additional safety barriers.

In critical scenario 1, a single pulse disturbance was added to signal 1 in Fig. 6 to simulate the critical scenario. The pulse was considered to start 35 min after the beginning of the simulation, and to have a duration of 15 min and an amplitude of 3300 mg_{HCl}/Nm^3, which represents a deviation of 5.5 times the average HCl concentration levels in the flue gas of the reference plant.

As discussed above, a safety barrier consisting in dolomitic lime injection in the furnace was introduced in the model to control the HCl concentration in the flue gas entering the FGT system in the presence of HCl spikes. According to Dal Pozzo et al. (2020), the following correlation can be used to link the dolomitic sorbent feed rate and the corresponding HCl conversion:

$$\chi = \frac{SR^{1.38} - SR}{SR^{1.38} - 1} \tag{9}$$

where χ is the conversion of HCl and SR is the Stochiometric Ratio, representing the ratio between the actual feed rate of dolomitic sorbent and its theoretical demand to achieve full HCl removal according to stoichiometry (Vehlow, 2015). The exponent in Eq. (9) is an empirical parameter derived from tests at WtE facilities (Dal Pozzo et al., 2020). This correlation can be used to obtain the final HCl concentration in the flue gas leaving the furnace after the activation of the dolomitic lime injection system. However, it does not reveal the dynamic of the phenomenon. Therefore, a simplified data-driven approach was followed to obtain the time trend of the HCl concentration in the flue gas after the activation of the safety barrier. Specifically, non-linear least squares were used to fit 4th-order polynomial functions to experimental data. These data consist of 10 experimental runs of dolomitic lime injection performed at different SR values (see Dal Pozzo et al., 2020). The following procedure was used to obtain the optimal fitting:

1. Experimental data were divided into three distinct groups based on their average SR value. Selected SR values are SR = 1, SR = 1.8, and SR = 2.5.
2. Experimental data were scaled in the range (0, 1) through min-max normalization.

$$\widehat{C}_{HCl}(t) = \frac{C_{HCl}(t) - min(C_{HCl}(t))}{\max(C_{HCl}(t)) - \min(C_{HCl}(t))} \tag{10}$$

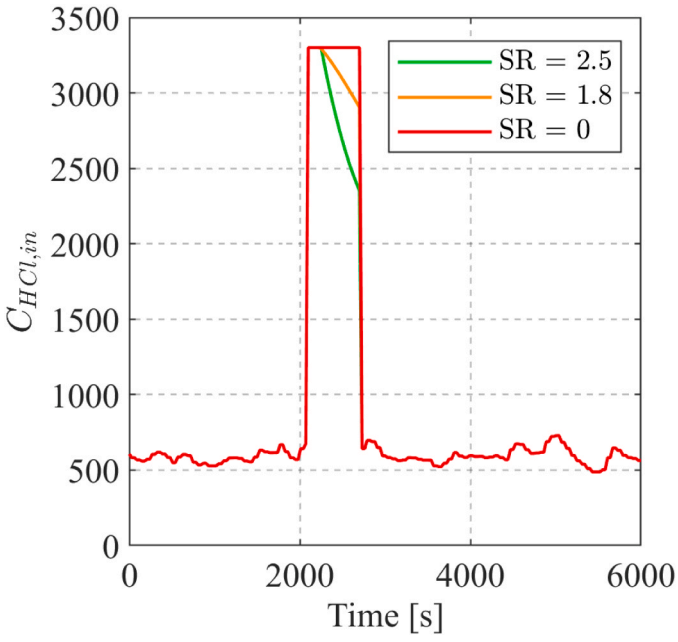


Fig. 9. Effect of the safety barrier considered (dolomitic lime injection) on the HCl concentration in critical event 1. The concentration of HCL considering two different configurations of safety barrier (SR 1.8 and SR 2.5) is compared to the baseline concentration in the absence of safety barriers.

Where $C_{HCl}(t)$ represents the HCl concentration in the flue gas leaving the furnace after the activation of the sorbent injection system, and $\widehat{C}_{HCl}(t)$ indicates the scaled concentration.

3. Scaled experimental data that belong to the same SR group were used to fit 4th-order polynomial functions through non-linear least squares.

$$\widehat{C}_{HCl}(t) = a \bullet t^4 + b \bullet t^3 + c \bullet t^2 + d \bullet t + e \tag{11}$$

where a, b, c, d, and e represent the function parameters.

The fitting procedure led to the parameters shown in Table A1, while the resulting fittings of experimental data is shown in Figure A2 in Appendix 2. Now, the following equations are available:

$$\begin{cases} \dfrac{C_{HCl}(t) - min(C_{HCl}(t))}{max(C_{HCl}(t)) - min(C_{HCl}(t))} = a \bullet t^4 + b \bullet t^3 + c \bullet t^2 + d \bullet t + e \\ \dfrac{SR^{1.38} - SR}{SR^{1.38} - 1} = 1 - \dfrac{C_f}{C_0} \end{cases} \tag{12}$$

where C_f is the final HCl concentration and C_0 is the initial concentration. It is worth noting that the second equation is Eq. (9). Considering that in a simulation SR is user-defined, the parameters a, b, c, d, and e are known. Also, assuming that $C_0 = \max(C_{HCl}(t))$ represents the HCl concentration when the furnace injection system starts and $C_f = \min(C_{HCl}(t))$ indicates the HCl concentration when the injection system stops, Eq. (12) can be used to calculate $C_{HCl}(t)$ and, therefore, to model the barrier dynamics.

Fig. 9 shows the effect of the activation of dolomitic lime injection on critical event 1 (a 15-minute-long spike of HCl). The red curve in the figure (i.e., SR = 0) represents the HCl concentration in the flue gas entering the FGT system during critical scenario 1 when no safety barrier is activated. The orange and green lines show the system behavior after the installation of the safety barrier, which is activated two minutes after the beginning of the peak and stays active until the end of the disturbance. Two different barrier configurations were investigated: SR = 1.8 (orange line) and SR = 2.5 (green line).

With respect to critical event 2, the failure of the screw feeder was simulated as a period of variable duration in which the sorbent mass flow rate (stream 3 in Fig. 6) is set to 0 kg/h. This is achieved by modifying the sorbent mass flow rate as follows:

$$\dot{m}_{sorbent}(t) = \begin{cases} \dot{m}_{sorbent}(t) & if\ t < t_f \ \vee \ t > t_b \\ 0 & if\ t_f \leq t \leq t_b \end{cases} \tag{13}$$

where $t_f = 45$ min indicates the time of failure and t_r represents the time of activation of the backup screw feeder. As mentioned in Section 5.1, in the base case it was assumed that the activation of the backup feeder is manual. A time window of 5 min ($t_b = t_f + 300$ s) seems plausible for operators to acknowledge the alarm, interpret the situation, and take action.

The overall effect of the specific safety barrier identified for this event (automatic activation of the backup feeder) is to reduce the time required to activate the backup screw feeder. This behavior can be simulated by reducing t_b in eq. (13). A response time of 15 s was deemed sufficient for the Safety Instrumented System to activate the backup feeder by plant personnel and instrumentation experts ($t_b = t_f + 15$ s) when the safety barrier is present.

5.3. Simulation of critical scenarios

Two sets of simulations were performed, as mentioned in Section 3.5. The first group of simulations evaluates the response of the original FGT system during critical scenarios (i.e., with no additional safety barrier). The second group of simulations aims to assess the system response after installing the safety barriers.

The results of the simulations of the first critical scenario and safety barrier are shown in Fig. 10a. The red line represents the HCl concentration in the clean gas leaving the original FGT system during the first critical event (i.e., HCl peak). The orange and green lines indicate the response of the system in case of activation of the dolomitic lime injection.

The system performance in the second critical scenario with and without considering the safety barrier is shown in Fig. 10b. Also in this case, the red line represents the response of the original system, while the green line indicates the system response with automatic activation of the backup screw feeder.

5.4. Assessment and comparison of safety barriers

In order to allow the qualitative and quantitative comparison of alternatives, the results of the simulations were used to compute the

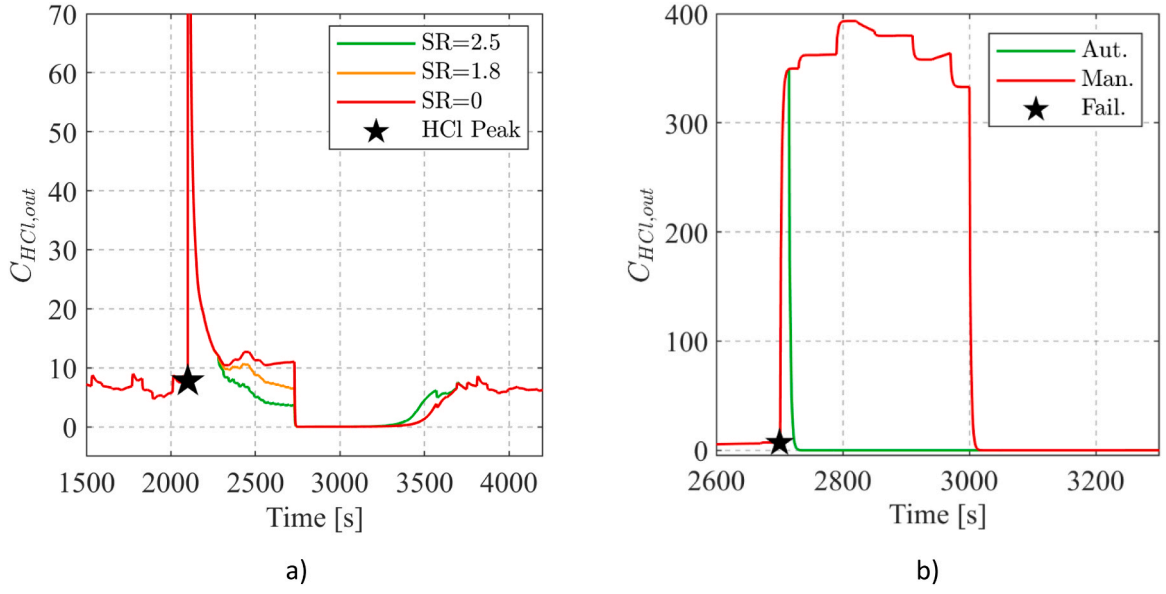


Fig. 10. Simulation of critical scenarios with and without safety barriers: a) critical scenario 1 with or without dolomitic lime injection, b) critical scenario 2 with or without automatic activation of the backup screw feeder.

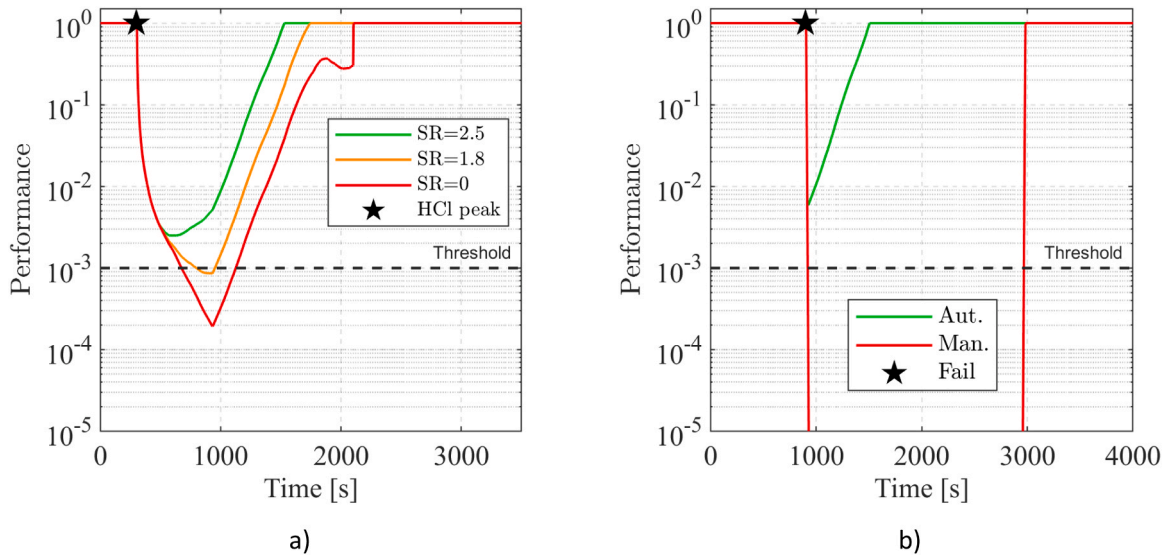


Fig. 11. System performance as defined in Eq. (8) for a) critical scenario 1 with or without dolomitic lime injection and b) critical scenario 2 with or without automatic activation of the backup screw feeder. Dashed threshold line corresponds to ELV HCl concentration. In panel (b), the values of performance for the manual case fall below the lower limit for y-axis was set at 10^{-5} to allow readability.

Table 2
Resilience loss for the two critical scenarios with and without safety barriers.

Critical scenario	Safety barrier	RL [s]
1	No	1651
1	Dolomitic lime injection (SR=1.8)	1302
1	Dolomitic lime injection (SR=2.5)	1103
2	No	2072
2	Automatic backup feeder	487

resilience indicators defined in Section 3.6: the performance metric ($\varphi(t)$) and the Resilience Loss (RL).

The time-trend of the performance indicator for the first critical event and safety barrier is shown in Fig. 11a. The red line represents the performance of the original system, and the orange and green lines indicate the performance after the installation of the dolomitic lime injection.

Similarly, the results of the second critical event and safety barrier are shown in Fig. 11b. It is worth mentioning that in critical event 2 the performance of the original system (i.e., the red curve in Fig. 11b), defined by Eq.(8), drops to 10^{-60}, with HCl concentrations significantly higher the ELVs. In order to increase the readability of the plot in Fig. 11b, the lower limit of the y-axis was set at 10^{-5}. The values calculated for the Resilience Loss in each scenario are summarized in Table 2.

6. Discussion

In several industrial applications, the current practice concerning the optimization of full-scale industrial processes is highly empirical and based on test-runs. However, as discussed above, this approach could hardly be applied to investigate the system performance in the vicinity of emission limits due to the risk of exceeding emission limits during the tests and to the negative consequences related to such events. As shown in Sections 5.3 and 5.4, the use of a digital model combined to hazard identification techniques allowed the identification and dynamic simulation of critical events and, more importantly, the performance assessment of safety barriers. In particular, the results obtained show the possibility of simulating the dynamic behavior of environmentally critical systems with and without safety barriers, providing a quantitative feedback on the increase in the operability, environmental safety and resilience of the system deriving from the installation of such barriers. Thus, the proposed approach can offer to plant managers, control room operators, and safety practitioners a crucial support in the decision-making process for the installation of safety barriers.

When considering the specific results obtained in the case-study, it is clear that in the case of the first critical event identified, as shown in Fig. 11a, the original system cannot withstand the deviations considered. Actually, the performance (red line) decreases rapidly after the critical event and reaches a minimum of $2\bullet10^{-4}$, which indicates that the system could not comply with the ELVs. The performance curves obtained at SR = 1.8 (orange) and SR = 2.5 (green) show that a safety barrier consisting in a dolomitic sorbent injection system in the furnace has the potential to mitigate the first critical scenario. In fact, the minimum performance increases if larger SRs are used. Also, the safety barrier ensures that the minimum performance occurs earlier, which indicates a faster recovery. However, the results also show that a stochiometric ratio equal to 1.8 (orange line) is insufficient to avoid exceeding emission limits. Indeed, the system performance briefly crosses the threshold of $1\bullet10^{-3-3}$ and reaches a minimum of $8.57\bullet10^{-4}$. On the contrary, the system performance obtained with SR equal to 2.5 (green line) reaches a minimum of $5\bullet10^{-3}$, which implies that the emission limit has never been exceeded. This finding confirms that the proposed approach can not only evaluate the dynamic response of safety barriers, but also guide the optimal tuning of their configuration. It should also be remarked that carrying out test-runs at the existing facility to explore system behavior in the conditions addressed would have been hardly feasible, since compliance to ELVs during tests is not granted.

Regarding the second critical scenario, the original system undergoes a complete degradation of performance during the whole critical event (red line in Fig. 11b). On the contrary, the automatic startup of the backup feeder ensures a minimum performance of $6\bullet10^{-3}$, which guarantees compliance with the ELVs.

The analysis of the performance metrics (Fig. 11) shows that the proposed safety barriers can effectively mitigate the critical scenarios considered in the case-study carried out. The Resilience Loss may be used to quantify the improvements brought by the additional safety barriers considered for implementation. Table 2 reveals that the dolomitic lime injection increases the system resilience by 21% (SR = 1.8), and by 33% (SR = 2.5) respectively when considering the first critical scenario. Similarly, the second safety barrier improves system resilience by 76%. It must be stressed these results do not suggest that the second safety barrier should be preferred over the first one. Actually, each safety

barrier is installed to deal with a specific event, which means that the second safety barrier does not affect the first critical scenario and vice versa. Performance comparison of design alternatives should be limited to those referring to the same critical event.

It is also important to mention some limitations of the proposed approach that need to be considered. Firstly, it is evident that the proposed approach, being based on data-driven models, specifically addresses the retrofitting of existing plants, rather than the design of new plants. Nonetheless, even when considering the case-study, the potential relevance of the method emerges, in spite of this limitation. Actually, in the framework of rapidly evolving regulations on emission control worldwide (Huang et al., 2021; Van Caneghem et al., 2019), existing WtE facilities need to increase the performance of their FGT systems in terms of both removal efficiency and reliability.

Secondly, the proposed approach addresses specifically the quantification of the effectiveness of the safety barrier. In addition to effectiveness, the assessment of safety barriers should take into account other criteria, namely response time, availability and level of confidence (de Dianous and Fievez, 2006). The response time, intended as the duration between the deployment of the safety barrier and the complete achievement of its safety function (de Dianous and Fievez, 2006), can be estimated from the results of the simulations (see again Fig. 11). The time between the detection of the ELV exceedance and the activation of the barrier is a required input of the simulations and, as discussed in Section 5.2, it can be obtained from tests (as in the case of dolomitic lime injection) or from operating experience (as in the case of the backup screw feeder). The time between the activation of the barrier and the full achievement of its safety function is an output of the simulations, as they are dynamic by nature and trace the evolution of barrier effectiveness over time. Conversely, aspects related to the level of confidence and the availability of the barrier are not assessed in the proposed approach, since they are associated with inherent properties and maintenance strategies of the barrier components and not with the effect of the safety barrier on the functionality of the FGT system, which is the key mechanism addressed by the simulations.

Thirdly, in the proposed approach, the evaluation of safety barriers is approached solely from an environmental perspective, while economic aspects have been disregarded in the case study. This choice aimed to demonstrate the feasibility and usefulness of the approach without introducing additional complexity. However, economic aspects must be considered when evaluating alternatives. For example, the user may combine performance and resilience assessment with cost-benefit analysis or more comprehensive techniques such as Life Cycle Assessment (LCA) (International Organization for Standardization, 2006). Alternatively, further studies may focus on improving the performance metric proposed in Eq. (8) to consider the costs associated with a particular process configuration.

Lastly, in the case-study a single barrier was considered for each critical event. On the one hand, a more realistic approach would be to consider and compare different safety barriers, from the safety, environmental and economic perspective. On the other hand, considering a single barrier provides a straightforward application of the methodology to different critical scenarios. Thus, since the intent of the case study is to provide a full-scale notional application of the methodology, the latter approach was privileged. Nevertheless, the approach developed and the specific models may be used as well to address the comparison and selection of safety barriers in a more comprehensive decision-making framework.

All in all, the application of the methodology demonstrated the possibilities arising from the integration between hazard identification techniques (e.g., HazOp and Bow-Tie analysis) and advanced simulation tools (i.e., dynamic modeling and resilience analysis) in the context of environmental risk management. The proposed framework is flexible and different choices in terms of both risk identification and process modeling can be adopted, also depending on the characteristics of the reference system and the related data availability.

Moreover, the analysis of the case study suggests that the dynamic modeling of critical events and evaluation of safety barriers through resilience analysis offers an interesting opportunity to improve environmental risk management. The approach goes beyond the static view of safety barriers (i.e., effective-not effective, and characterized by a context-independent Probability of Failure) towards a dynamic vision of the risk, where the effectiveness of safety barriers is closely linked to the dynamics of the underlying phenomena. The methodology fits perfectly in a Dynamic Risk Management framework since it is inherently updatable and can be reiterated to account for changes in the environment (e.g., changes in process conditions or plant layout) (Grøtan and Paltrinieri, 2016) and to incorporate new information as they become available (e.g., considering new critical events as more knowledge is accessible) (Paltrinieri et al., 2014).

7. Conclusions

The approach described in this study offers a comprehensive and structured framework for the dynamic evaluation of safety barriers in environmentally critical facilities based on digital modelling. The method is based on a pre-defined flowcharts of activities, covering most of the risk management phases, from the identification of critical scenarios to the evaluation of the system response. In addition, the method is sufficiently generic to allow some flexibility (e.g., with respect to modeling techniques and tools) in order to be adapted to diverse needs. The approach has several advantages and novelty elements, such as the focus on environmental risk management from a safety engineering perspective (which is often disregarded in the literature) and the integration between traditional risk management techniques and modern data-driven models, which allows the definition and simulation of critical scenarios and safety barriers that would be impossible to evaluate through first-principles or field tests. Furthermore, the methodology requires a relatively small set of data, which is often promptly available in most gas treatment facilities. In addition, the use of resilience analysis for the dynamic evaluation of safety barriers and the intrinsic updatability of the approach are further elements of novelty that contribute to dynamic risk management. The method has been tested on a full-scale real life industrial case study to demonstrate its feasibility and effectiveness. The results – which appear informative, yet easy to interpret – allow qualitative and quantitative evaluation and comparison of safety barriers. In the context of growing attention to environmental issues and widespread digitalization of production processes, this study suggests that data-driven models may effectively support traditional risk management approaches to improve environmental safety and accomplish tasks that are impractical or impossible to perform through first principles.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix 1. Results of the HazOp analysis

The full results of the HazOp Analysis are condensed in the bow-tie diagram below.

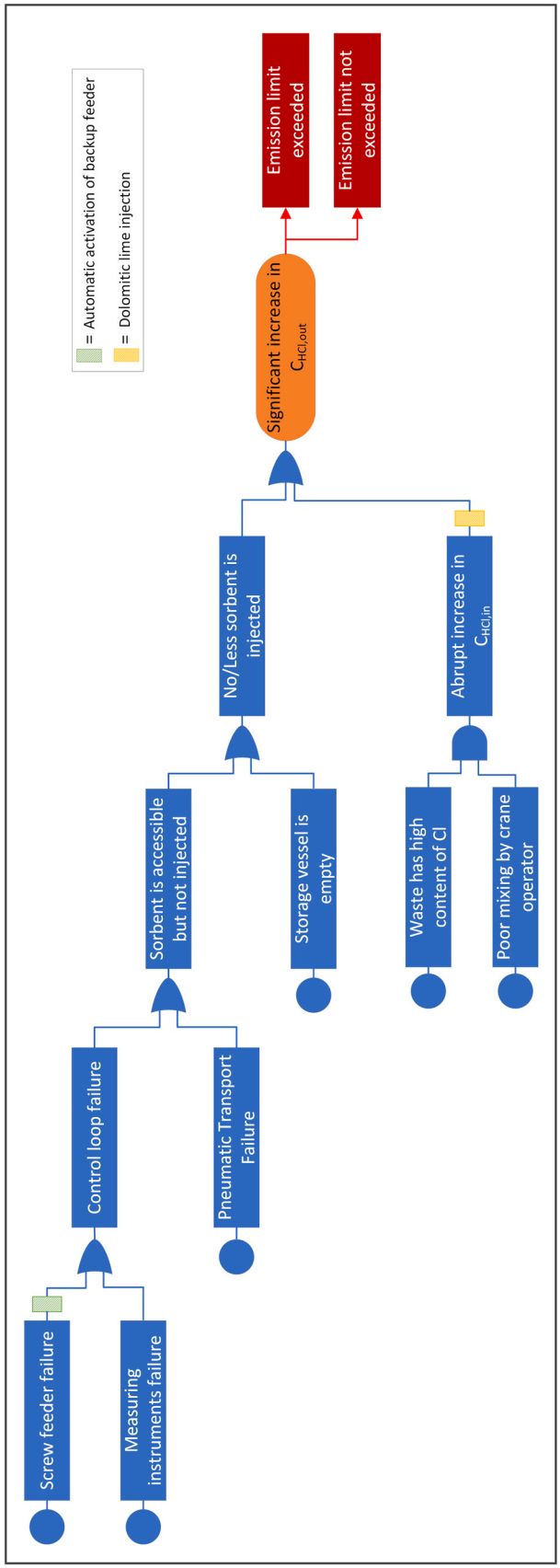


Fig A1. Complete bow-tie diagram for the reference system.

Appendix 2. Supporting information on furnace sorbent injection

Table A1
Fitting parameters for different SR groups.

Group	*a*	*b*	*c*	*d*	*e*
SR∼1	-3.09×10^{-7}	5.99×10^{-5}	-2.46×10^{-3}	-9.71×10^{-4}	9.83×10^{-1}
SR∼1.8	-1.68×10^{-6}	1.33×10^{-4}	-2.50×10^{-3}	-3.16×10^{-2}	1.01×10^{0}
SR∼2.5	-4.41×10^{-7}	-5.08×10^{-5}	5.60×10^{-3}	-1.37×10^{-1}	1.07×10^{0}

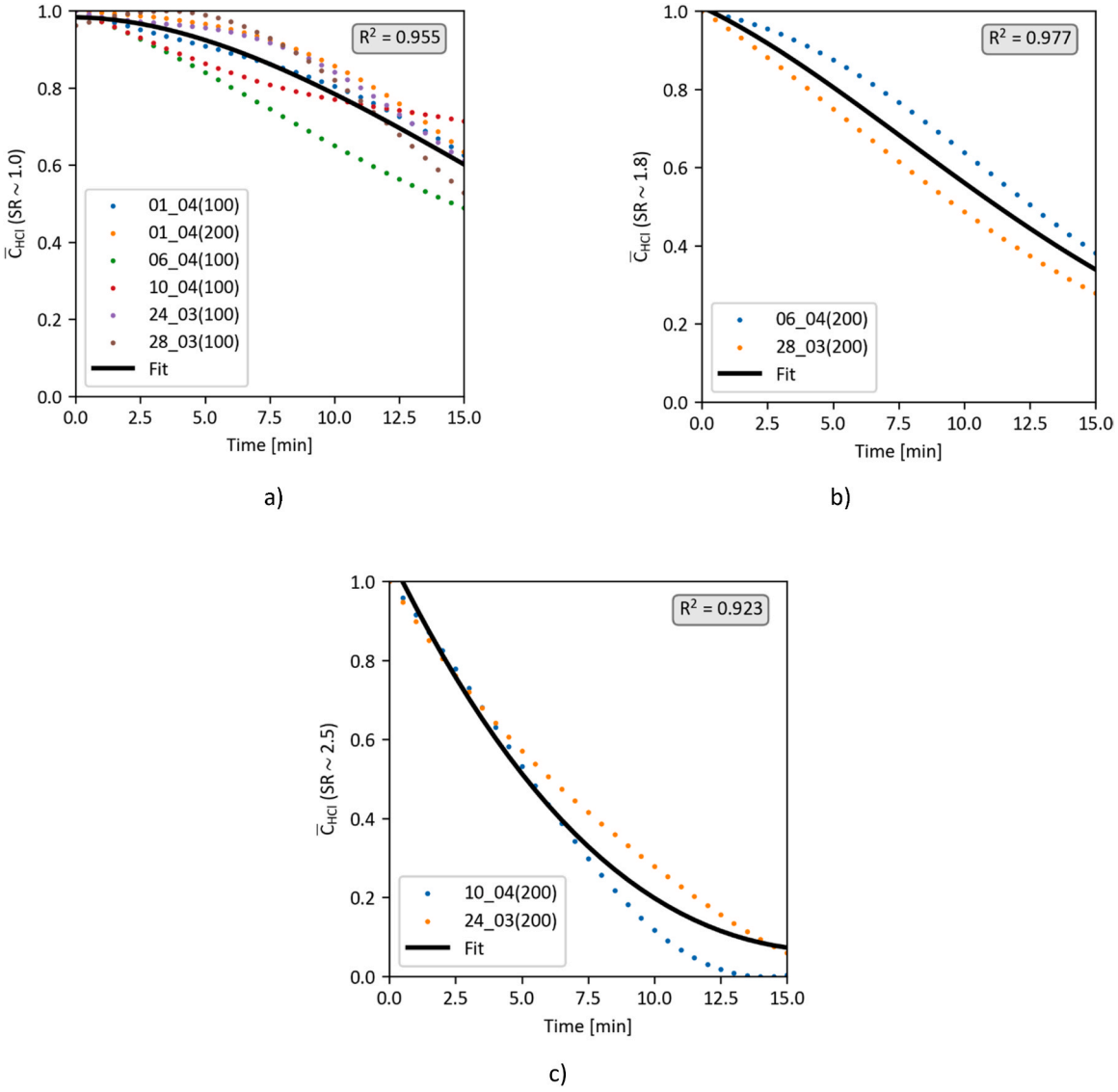


Fig A2. Experimental data and fitting curves for SR∼1 (a), SR∼1.8 (b), SR∼2.5 (c). Experimental runs are named xx_yy(zzz), where xx represents the day of collection, yy indicates the month, and zzz indicates the dolomitic sorbent feed rate in kg/h. The coefficient of determination of the fitting functions (R^2) is displayed in the upper right corner.

Appendix C. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.psep.2023.11.021.

References

Antonioni, G., Dal Pozzo, A., Guglielmi, D., Tugnoli, A., Cozzani, V., 2016. Enhanced modelling of heterogeneous gas–solid reactions in acid gas removal dry processes. Chem. Eng. Sci. 148, 140–154. https://doi.org/10.1016/j.ces.2016.03.009.

Argenti, F., Landucci, G., Cozzani, V., Reniers, G., 2017. A study on the performance assessment of anti-terrorism physical protection systems in chemical plants. Saf. Sci. 94, 181–196. https://doi.org/10.1016/j.ssci.2016.11.022.

Bacci Di Capaci, R., Pannocchia, G., Pozzo, A.D., Antonioni, G., Cozzani, V., 2022. Data-driven models for advanced control of acid gas treatment in waste-to-energy plants. IFAC-Pap. 55, 869–874. https://doi.org/10.1016/j.ifacol.2022.07.554.

Bai, Y., Wu, J., Yuan, S., Reniers, G., Yang, M., Cai, J., 2022. Dynamic resilience assessment and emergency strategy optimization of natural gas compartments in utility tunnels. Process Saf. Environ. Prot. 165, 114–125. https://doi.org/10.1016/j.psep.2022.07.008.

Bergström, J., van Winsen, R., Henriqson, E., 2015. On the rationale of resilience in the domain of safety: a literature review. Reliab. Eng. Syst. Saf. 141, 131–141. https://doi.org/10.1016/j.ress.2015.03.008.

Beylot, A., Hochar, A., Michel, P., Descat, M., Ménard, Y., Villeneuve, J., 2018. Municipal solid waste incineration in france: an overview of air pollution control techniques, emissions, and energy efficiency. J. Ind. Ecol. 22, 1016–1026. https://doi.org/10.1111/jiec.12701.

Biganzoli, L., Racanella, G., Marras, R., Rigamonti, L., 2015. High temperature abatement of acid gases from waste incineration. Part II: Comparative life cycle assessment study. Waste Manag 35, 127–134. https://doi.org/10.1016/j.wasman.2014.10.021.

Box, G.E.P., Jenkins, G.M., Reinsel, G.C., Ljung, G.M., 2015. Time series analysis: forecasting and control. Wiley Series in Probability and Statistics. Wiley.

Bubbico, R., Lee, S., Moscati, D., Paltrinieri, N., 2020. Dynamic assessment of safety barriers preventing escalation in offshore Oil&Gas. Saf. Sci. 121, 319–330. https://doi.org/10.1016/j.ssci.2019.09.011.

Bucelli, M., Landucci, G., Haugen, S., Paltrinieri, N., Cozzani, V., 2018. Assessment of safety barriers for the prevention of cascading events in oil and gas offshore installations operating in harsh environment. Ocean Eng. 158, 171–185. https://doi.org/10.1016/j.oceaneng.2018.02.046.

CCPS, Energy Institute, 2018. Bow ties in risk management. Bow Ties in Risk Management. Wiley. https://doi.org/10.1002/9781119490357.

Chibante, V.G., Fonseca, A.M., Salcedo, R.R., 2010. Modeling dry-scrubbing of gaseous HCl with hydrated lime in cyclones with and without recirculation. J. Hazard. Mater. 178, 469–482. https://doi.org/10.1016/j.jhazmat.2010.01.106.

Daintith, J., Wright, E., 2008. A Dictionary of Computing. Oxford University Press. https://doi.org/10.1093/acref/9780199234004.001.0001.

Dal Pozzo, A., Capecci, S., Cozzani, V., 2023b. Techno-economic impact of lower emission standards for waste-to-energy acid gas emissions. Waste Manag.

Dal Pozzo, A., Lucquiaud, M., De Greef, Johan, 2023c. Research and innovation needs for the Waste-to-Energy sector towards a net-zero circular economy. Energies.

Dal Pozzo, A., Antonioni, G., Guglielmi, D., Stramigioli, C., Cozzani, V., 2016. Comparison of alternative flue gas dry treatment technologies in waste-to-energy processes. Waste Manag 51, 81–90. https://doi.org/10.1016/j.wasman.2016.02.029.

Dal Pozzo, A., Guglielmi, D., Antonioni, G., Tugnoli, A., 2018a. Environmental and economic performance assessment of alternative acid gas removal technologies for waste-to-energy plants. Sustain. Prod. Consum 16, 202–215. https://doi.org/10.1016/j.spc.2018.08.004.

Dal Pozzo, A., Moricone, R., Antonioni, G., Tugnoli, A., Cozzani, V., 2018b. Hydrogen chloride removal from flue gas by low-temperature reaction with calcium hydroxide. Energy Fuels 32, 747–756. https://doi.org/10.1021/acs.energyfuels.7b03292.

Dal Pozzo, A., Lazazzara, L., Antonioni, G., Cozzani, V., 2020. Techno-economic performance of HCl and SO2 removal in waste-to-energy plants by furnace direct sorbent injection. J. Hazard. Mater. 394, 122518 https://doi.org/10.1016/j.jhazmat.2020.122518.

Dal Pozzo, A., Muratori, G., Antonioni, G., Cozzani, V., 2021. Economic and environmental benefits by improved process control strategies in HCl removal from waste-to-energy flue gas. Waste Manag. 125, 303–315. https://doi.org/10.1016/j.wasman.2021.02.059.

Dal Pozzo, A., Abagnato, S., Cozzani, V., 2023a. Assessment of cross-media effects deriving from the application of lower emission standards for acid pollutants in waste-to-energy plants. Sci. Total Environ. 856, 159159 https://doi.org/10.1016/j.scitotenv.2022.159159.

Delvosalle, C., Fievez, C., Pipart, A., Debray, B., 2006. ARAMIS project: a comprehensive methodology for the identification of reference accident scenarios in process industries. J. Hazard. Mater. 130, 200–219. https://doi.org/10.1016/j.jhazmat.2005.07.005.

de Dianous, V., Fievez, C., 2006. ARAMIS project: a more explicit demonstration of risk control through the use of bow–tie diagrams and the evaluation of safety barrier performance. J. Hazard. Mater. 130, 220–233. https://doi.org/10.1016/j.jhazmat.2005.07.010.

Emmert-Streib, F., Dehmer, M., 2019. Evaluation of regression models: model assessment, model selection and generalization error. Mach. Learn. Knowl. Extr. 1, 521–551. https://doi.org/10.3390/make1010032.

European Commission, 2019. Best Available Techniques (BAT) Reference Document for Waste Incineration, EUR 29971 EN. https://doi.org/10.2760/761437.

European Commission, 2020. Best Available Techniques (BAT) reference document for waste incineration: Industrial Emissions Directive 2010/75/EU (Integrated Pollution Prevention and Control). Publications Office. https://doi.org/doi/10.2760/761437.

Grøtan, T.O., Paltrinieri, N., 2016. Chapter 20 - Dynamic Risk Management in the Perspective of a Resilient System. In: Paltrinieri, Nicola, Khan, F. (Eds.), Dynamic Risk Analysis in the Chemical and Petroleum Industry. Butterworth-Heinemann, pp. 245–257. https://doi.org/10.1016/B978-0-12-803765-2.00020-2.

Han, Y., Zhen, X., Huang, Y., Vinnem, J.E., 2019. Integrated methodology for determination of preventive maintenance interval of safety barriers on offshore installations. Process Saf. Environ. Prot. 132, 313–324. https://doi.org/10.1016/j.psep.2019.09.035.

Hokstada, P., Øien, K., Reinertsen, R., 1998. Recommendations on the use of expert judgment in safety and reliability engineering studies. Two offshore case studies. Reliab. Eng. Syst. Saf. 61, 65–76. https://doi.org/10.1016/S0951-8320(97)00084-7.

Hollnagel, E., Paries, J., Woods David, D., Wreathall, J., 2010. Resilience Engineering in Practice: A Guidebook, Ashgate Studies in Resilience Engineering. Ashgate Publishing.

Hollnagel, E., Pariès, J., Woods, D.D., Wreathall, J., 2011. Resilience Engineering in Practice: A Guidebook, 1st ed. Ashgate, Farnham, England.

Hosseini, S., Barker, K., Ramirez-Marquez, J.E., 2016. A review of definitions and measures of system resilience. Reliab. Eng. Syst. Saf. 145, 47–61. https://doi.org/10.1016/j.ress.2015.08.006.

Huang, W., Li, H., Fan, H., Qian, Y., 2021. Causation mechanism analysis of excess emission of flue gas pollutants from municipal solid waste incineration power plants by employing the Fault Tree combined with Bayesian Network: A case study in Dongguan. J. Clean. Prod. 327, 129533 https://doi.org/10.1016/j.jclepro.2021.129533.

International Organization for Standardization, 2006. Environmental management — Life cycle assessment — Principles and framework, ISO 14040:2006(E). Geneva, CH.

International Organization for Standardization, 2018. Risk management - Guidelines, ISO 31000:2018. Geneva, CH.

International Organization for Standardization, 2019. Risk management - Risk assessment techniques, IEC 31010:2019. Geneva, CH.

Khan, F., Rathnayaka, S., Ahmed, S., 2015. Methods and models in process safety and risk management: Past, present and future. Process Saf. Environ. Prot. 98, 116–147. https://doi.org/10.1016/j.psep.2015.07.005.

Khan, F., Hashemi, S.J., Paltrinieri, N., Amyotte, P., Cozzani, V., Reniers, G., 2016. Dynamic risk management: a contemporary approach to process safety management. Curr. Opin. Chem. Eng. 14, 9–17. https://doi.org/10.1016/j.coche.2016.07.006.

Knight, J.C., 2002. Safety critical systems: challenges and directions, in: Proceedings of the 24th International Conference on Software Engineering. ICSE 2002. pp. 547–550.

Kockmann, N., 2019. Digital methods and tools for chemical equipment and plants. React. Chem. Eng. 4, 1522–1529. https://doi.org/10.1039/C9RE00017H.

Kotu, V., Deshpande, B., 2019. Chapter 12 - time series forecasting. In: Kotu, V., Deshpande, B. (Eds.), Data Science (Second Edition). Morgan Kaufmann, pp. 395–445. https://doi.org/10.1016/B978-0-12-814761-0.00012-5.

Landucci, G., Argenti, F., Tugnoli, A., Cozzani, V., 2015. Quantitative assessment of safety barrier performance in the prevention of domino scenarios triggered by fire. Reliab. Eng. Syst. Saf. 143, 30–43. https://doi.org/10.1016/j.ress.2015.03.023.

Leveson, N., Dulac, N., Zipkin, D., Cutcher-Gershenfeld, J., Carroll, J., Barrett, B., 2006. Chapter 8 - Engineering Resilience into Safety-Critical Systems. In: Hollnagel, E., Woods, D.D., Leveson, N. (Eds.), Resilience Engineering Concepts and Precepts. Ashgate Publishing Limited. https://doi.org/10.1201/9781315605685.

Liu, Y., 2020. Safety barriers: Research advances and new thoughts on theory, engineering and management. J. Loss Prev. Process Ind. 67, 104260 https://doi.org/10.1016/j.jlp.2020.104260.

Ljung, L., 1999. System identification: theory for the user. Prentice Hall information and system sciences series. Prentice Hall PTR,.

Ljung, L., 2010. Perspectives on system identification. Annu. Rev. Control 34, 1–12. https://doi.org/10.1016/j.arcontrol.2009.12.001.

Magnanelli, E., Tranås, O.L., Carlsson, P., Mosby, J., Becidan, M., 2020. Dynamic modeling of municipal solid waste incineration. Energy 209, 118426. https://doi.org/10.1016/j.energy.2020.118426.

Mannan, S., 2005. Hazard Identification. In: Mannan, S. (Ed.), Lees' Loss Prevention in the Process Industries. Elsevier, Burlington, pp. 8/1–8/79. https://doi.org/10.1016/B978-075067555-0.50096-7.

Maurya, A., Kumar, D., 2020. Reliability of safety-critical systems: a state-of-the-art review. Qual. Reliab. Eng. Int. 36, 2547–2568. https://doi.org/10.1002/qre.2715.

Misuri, A., Landucci, G., Cozzani, V., 2020. Assessment of safety barrier performance in Natech scenarios. Reliab. Eng. Syst. Saf. 193, 106597 https://doi.org/10.1016/j.ress.2019.106597.

Misuri, A., Landucci, G., Cozzani, V., 2021. Assessment of risk modification due to safety barrier performance degradation in Natech events. Reliab. Eng. Syst. Saf. 212, 107634 https://doi.org/10.1016/j.ress.2021.107634.

Paltrinieri, N., Khan, F.I., 2020. Dynamic. risk Anal. 35–60. https://doi.org/10.1016/bs.mcps.2020.04.001.

Paltrinieri, N., Khan, F., Amyotte, P., Cozzani, V., 2014. Dynamic approach to risk management: Application to the Hoeganaes metal dust accidents. Process Saf. Environ. Prot. 92, 669–679. https://doi.org/10.1016/j.psep.2013.11.008.

Patriarca, R., Bergström, J., Di Gravio, G., Costantino, F., 2018. Resilience engineering: Current status of the research and future challenges. Saf. Sci. 102, 79–100. https://doi.org/10.1016/j.ssci.2017.10.005.

Pozzo, A.D., Giannella, M., Antonioni, G., Cozzani, V., 2018. Optimization of the economic and environmental profile of HCl removal in a municipal solid waste incinerator through historical data analysis. Chem. Eng. Trans. 67, 463–468. https://doi.org/10.3303/CET1867078.

Raschka, S., 2018. Model evaluation, model selection, and algorithm selection. Mach. Learn. https://doi.org/10.48550/ARXIV.1811.12808.

Sarvestani, K., Ahmadi, O., Mortazavi, S.B., Mahabadi, H.A., 2021. Development of a predictive accident model for dynamic risk assessment of propane storage tanks. Process Saf. Environ. Prot. 148, 1217–1232. https://doi.org/10.1016/j.psep.2021.02.018.

Sklet, S., 2006. Safety barriers: Definition, classification, and performance. J. Loss Prev. Process Ind. 19, 494–506. https://doi.org/10.1016/j.jlp.2005.12.004.

Sun, H., Wang, H., Yang, M., Reniers, G., 2021. Resilience-based approach to safety barrier performance assessment in process facilities. J. Loss Prev. Process Ind. 73, 104599 https://doi.org/10.1016/j.jlp.2021.104599.

Thieme, C.A., Utne, I.B., 2017. Safety performance monitoring of autonomous marine systems. Reliab. Eng. Syst. Saf. 159, 264–275. https://doi.org/10.1016/j.ress.2016.11.024.

Torres, J.F., Hadjout, D., Sebaa, A., Martínez-Álvarez, F., Troncoso, A., 2020. Deep Learning for Time Series Forecasting: A Survey. Big Data 9, 3–21. https://doi.org/10.1089/big.2020.0159.

Tran, H.T., Balchanos, M., Domerçant, J.C., Mavris, D.N., 2017. A framework for the quantitative assessment of performance-based system resilience. Reliab. Eng. Syst. Saf. 158, 73–84. https://doi.org/10.1016/j.ress.2016.10.014.

V. De Dianous D. Hourtolou E. Bernuchon ARAMIS D1C – APPENDIX 2004 9.

Van Caneghem, J., Van Acker, K., De Greef, J., Wauters, G., Vandecasteele, C., 2019. Waste-to-energy is compatible and complementary with recycling in the circular economy. Clean. Technol. Environ. Policy 21, 925–939. https://doi.org/10.1007/s10098-019-01686-0.

Vehlow, J., 2015. Air pollution control systems in WtE units: An overview. Waste Manag 37, 58–74. https://doi.org/10.1016/j.wasman.2014.05.025.

Villa, V., Paltrinieri, N., Khan, F., Cozzani, V., 2016. Towards dynamic risk analysis: A review of the risk assessment approach and its limitations in the chemical process industry. Saf. Sci. 89, 77–93. https://doi.org/10.1016/j.ssci.2016.06.002.

Wang, Q., Ma, Y., Zhao, K., Tian, Y., 2022. A Comprehensive Survey of Loss Functions in Machine Learning. Ann. Data Sci. 9, 187–212. https://doi.org/10.1007/s40745-020-00253-5.

Yarveisy, R., Gao, C., Khan, F., 2020. A simple yet robust resilience assessment metrics. Reliab. Eng. Syst. Saf. 197, 106810 https://doi.org/10.1016/j.ress.2020.106810.

Zeng, T., Chen, G., Yang, Y., Chen, P., Reniers, G., 2020. Developing an advanced dynamic risk analysis method for fire-related domino effects. Process Saf. Environ. Prot. 134, 149–160. https://doi.org/10.1016/j.psep.2019.11.029.

Article XII.

Tamascelli, N., Javier, A., Nakhal Akel, A.J., Patriarca, R., Paltrinieri, N., Cruz, A.M. (2022). **Are we going towards "no-brainer" risk management? A case study on climate hazards**. Proceedings of the 16th Probabilistic Safety Assessment & Management Conference (PSAM16), Honolulu, Hawaii, 2022. ISBN: 9781713863755.

Are we going towards "no-brainer" risk management? A case study on climate hazards

Nicola Tamascelli[a,b], Antonio Javier Nakhal Akel[c], Riccardo Patriarca[c], Nicola Paltrinieri[a,d] and Ana Maria Cruz[d]

[a] Department of Mechanical and Industrial Engineering, Norwegian University of Science and Technology NTNU, Trondheim, Norway, nicola.tamascelli@ntnu.no

[b] Department of Civil, Chemical, Environmental, and Materials Engineering, Alma Mater Studiorum – University of Bologna, Bologna, Italy

[c] Department of Mechanical and Aerospace Engineering, Sapienza University of Rome, Rome, Italy

[d] Disaster Prevention Research Institute, Kyoto University, Kyoto, Japan

Abstract: The overall risk management domain is stepping into its 4.0 phase by implementing and increasingly relaying on cyber-technological systems. Enhanced computational power provides the capability of processing collected databases for prediction and preparation purposes. In fact, early warnings can lead to suggestion for proactive strategies, or directly initiate the action of autonomous actuators ensuring the required level of system safety. But have we reached the promises of digital risk management yet, or will we ever reach them? A traditional view on safety defines it as the absence of accidents and incidents. A forward-looking perspective on safety affirms that it involves ensuring that "as many things as possible go right". However, in both the views there is an element of uncertainty associated to the prediction of future risks and, more subtle, to the capability of possessing all the necessary information for such prediction. This uncertainty does not simply disappear once we apply advanced Machine Learning (ML) techniques to the infinite series of possible accident scenarios, but it can be found behind modelling choices and parameters setting. In a nutshell, "there ain't no such thing as a free lunch", i.e., any model claiming superior flexibility usually introduces extra assumptions. This contribution will illustrate a case on climate-driven disaster data extracted from the Emergency Events Database (EM-DAT) where ML techniques are used to understand natural disaster mortality and unravel underlying causes and influential factors that can inform decision-making and be relevant for risk reduction efforts. This manuscript may allow to affirm with certain confidence that present risk management systems are not even close to a "no-brainer" condition in which the responsibility for human and system safety is entirely moved to the machine. However, this shows that such advanced techniques are progressively providing a reliable support for critical decision making and guiding society towards more risk-informed and safety-responsible planning.

Keywords: Risk management, Climate hazards, Natural disasters, Machine Learning, Clustering

1. INTRODUCTION

At the beginning of the 90s, Prof. Diekmann [1] stated the following: "New analysis tools are emerging, which have the potential to allow complex risk analyses to be performed simply. These new tools, which are underpinned by decision analysis and, lately, expert-systems technology, may lead to powerful, yet simple, approaches to the representation of risky problems." Such optimistic prediction on the future of risk analysis was also accompanied by the suggestion of a possible interdisciplinary direction. "Future approaches to risk analysis will certainly rely more on the advances being made in Artificial Intelligence (AI) and cognitive sciences. New computer tools and knowledge-representation schemes will unquestionably lead to new techniques, insights and opportunities for risk analysis."

In the same decade (1997), the Russian chess grandmaster Garry Kimovich Kasparov (former World Chess Champion, ranked world No. 1 from 1984 until his retirement in 2005) lost a chess game with

the chess playing computer Deep Blue by IBM, which was an example of Good Old-Fashioned Artificial Intelligence (GOFAI) [2]. On that game, Kasparov later stated the following: "Deep Blue was intelligent the way your programmable alarm clock is intelligent [3]. Not that losing to a 10-million-dollar alarm clock made me feel any better."

In general, risk management has tried to make use of AI, but it has unevenly progressed since the mentioned events. It neither respected Diekmann's prediction (methodological gaps are still present [4]), nor turned into "programmable-alarm-clock intelligence" thanks to the progressive refinement of Machine Learning (ML) models and the increase in available computing power [5].

This contribution aims to outline what AI, and in particular ML techniques, can bring to risk analysis and management by illustrative examples related to climate-driven events (e.g., storms, floods, drought, heatwaves) . ML techniques are used to understand natural disaster mortality and unravel underlying causes and influential factors that can inform decision-making and be relevant for risk reduction efforts.

1.1. Machine Learning and Big Data

AI is intelligence demonstrated by machines and it is divided into sub-fields based on technical considerations, such as particular goals (e.g., "robotics" or "machine learning"), the use of particular tools ("logic" or artificial neural networks), or deep philosophical differences.

This contribution focuses on the sub-field of Machine Learning (ML). ML refers to techniques aiming to program computers to learn from experience [6]. ML is known for providing meaning to raw data and solving practical problems in a reliable and efficient way. These problems require machine assistance since the amount of data and the complexity of the statistical patterns imply that humans would not be able to solve them via traditional techniques [7].

ML rely on a collection of examples of some phenomena, to be used for training and finding patterns that can help make decisions and predictions for new, unseen information [8]. ML has several practical applications in present industrial processes [9], and it may be the key to unlocking the value of safety data to perform novel risk management systems. Therefore, a computer may run a ML algorithm to assess risks for safety-critical industries (e.g., Oil and Gas). It would allow processing a large amount of information in the form of indicators from normal operations and past undesired events (from mishaps to major accidents), which would be used for training the algorithm. Due to the subjectivity of risk definition [10], risk level cannot be assigned to each event with certainty and a supervised approach may be needed. Practical examples of ML adoption in risk management refer to predict system losses and possible risks in undesired cases [4]. Among the most used ML algorithm, one can find the clustering, used to reveal (in an unsupervised way) meaningful groups within a dataset based on underlying patterns or structures [11].

Increasing attention has been dedicated to monitoring safety barrier performance through indicators, as a way to assess and control risk. Indicators may report a series of factors: physical conditions of a plant (equipment pressure and temperature), number of failures of an equipment piece, maintenance backlog, number of emergency preparedness exercises run, amount of overtime worked, etc. [12]. Øien et al. [12], Paltrinieri et al.[13], [14], and Landucci et al. [15] have produced several reviews on risk and barrier indicators. They show that definition and collection of risk indicators have become consolidated practices in "high-risk" industrial sectors. Such trend towards definition and collection of higher numbers of indicators [16] demonstrates the mentioned challenge on big data process for risk level assessment.

In recent years, several studies have focused on ML techniques to support natural disaster risk management. One widespread approach is the analysis of disaster databases and reports to extract relevant information and support risk-informed decision-making [17]. An exhaustive overview of ML applied to natural risk management may be found in [18], [19]. However, most of these investigations focus on illustrating the potential and effectiveness of their approaches. Still, little attention has been

paid to the role of ML and whether its extensive use will lead to a condition where the responsibility for human safety is entirely moved to the machine. This study attempts to bridge this knowledge gap by illustrating an example of ML for natural disaster risk management and evaluating the need for human knowledge to interpret and contextualize the results.

1.2. Climate-driven natural disasters

Data analysis of climate hazards can aid risk management by shedding light on disaster characteristics, challenges, differences amongst regions, and similar events. Climate hazard management denotes the systematic actions focused on reducing the negative effects of disasters [20].

Mitigation measures contribute to climate hazard management by minimizing, monitoring, and reducing the probability of severe consequences, the corresponding avoidable impacts, and the unfortunate outcomes of natural hazards [21]. The risk for individuals inflicted by climate disasters differs based on societal vulnerability and exposure, and environmental conditions [22]. Climate change has forced more than 20 million people to move from their homes each year [23]. The development level of a country might affect the consequences of a natural disaster. It is often remarked how those living in poverty are hardest hit despite being the least responsible for climate change.

The increasing frequency of natural hazards led to greater attention worldwide devoted to mapping and reducing natural risks [24], unraveling and explaining potential impacts on societies. Vulnerability in this context can be a risk factor, but also an outcome: disaster exposure may lead to poverty causing damage to assets and livelihoods [25]. Besides, larger climate-driven disasters often cause extensive property damages and a high number of fatalities. Research has shown that natural disaster-related damages and mortality have increased in the past decades [23], [26].

Further research is needed to develop systematic approaches on disaster causes and impacts to improve responses, anticipation capacity, design risk prevention and mitigating interventions prior to or following major climate hazards. The International Disaster Database (EM-DAT) developed by the Centre for Research on the Epidemiology of Disasters (CRED) gathers data on natural disasters and maps them into different classification categories, impacts, and causes.

The study focuses on these climate-driven disasters in terms of societal impact, both on populations and properties, as they can be of relevance for industrial systems as well. EM-DAT is analyzed by using ML algorithm to investigate potential clusters of countries that show commonalities and subsequently can drive to common natural risk management mitigations.

2. EXAMPLE OF ML-BASED FOR RISK MANAGEMENT

2.1. EM-DAT database

The EM-DAT database was created following the 1980's investigation by CRED. The study was carried out to serve the purposes of humanitarian action at national and international levels. The initiative aimed to rationalize decision-making for disaster preparedness, as well as provide an objective base to assess vulnerability and set priorities.

The database is compiled from various sources, including United Nations agencies, non-governmental organizations, insurance companies, research institutes, and press agencies, e.g., United Nations Department of Humanitarian Affairs (UN-DHA), European Union Humanitarian Office (ECHO), International Federation of the Red Cross and Red Crescent, the Office of Foreign Disaster Assistance (OFDA-USAID), International Committee of the Red Cross and Red Croissant (ICRCRC, Switzerland), International Decade for Natural Disaster Reduction (IDNDR) [27].

Currently, EM-DAT collects more than 25000 disasters between 1900 - 2020. All the events in the EM-DAT database fulfill one or more of these entry criteria [27]:

- Deaths (10 or more people deaths)
- Affected (100 or more people affected, injuries or homeless)
- Declaration/Appeal (declaration by the country of a state of emergency and/or appeal for international assistance)

The reported incidents worldwide involve 189 countries, distributed as follows:

- About 15000 accidents are related to natural impacts (e.g., drought, extreme temperature, flood, landslide, storm, wildfire, etc.),
- About 10000 accidents refer to technological impacts (i.e., industrial, transport, and miscellaneous impacts).

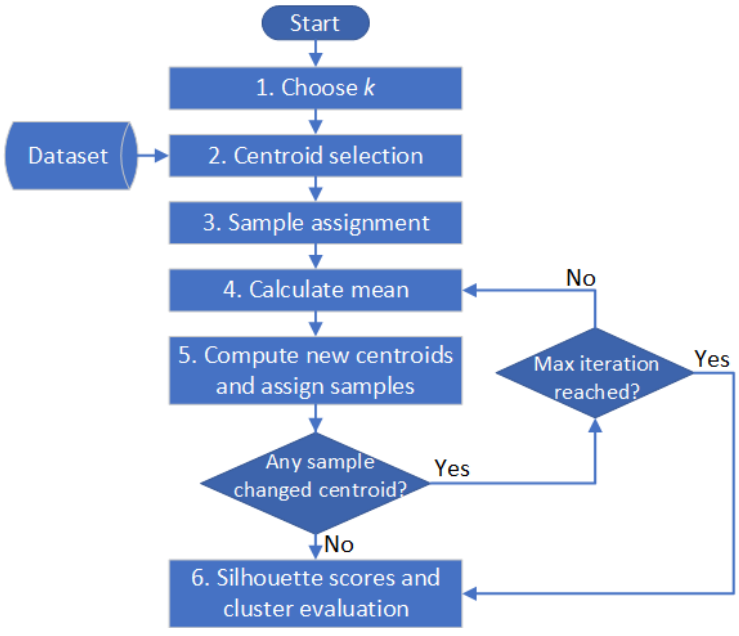
Technological events have not been considered in this study, and attention has been directed toward natural disasters. More specifically, only climate-driven disasters are examined (e.g., storms, floods, droughts, heatwaves). Other types of natural disasters (e.g., geophysical, biological, and extra-terrestrial) have been excluded from the analysis.

The database incorporates 43 parameters (e.g., location, date, damage, fatalities, disaster type, origin, reconstruction cost, insured damage, appeal, impacts) to fully detail the characteristics of the accident and allow its analysis [27].

3. METHOD

K-means is one of the most frequently used and effective clustering algorithms, as proved by results obtained in several diverse application contexts [28]. K-means has been used in this study to cluster countries found in EM-DAT, based on their similarity toward natural disaster exposure. The algorithm tries to group data by minimizing the within-cluster-sum-of-squares, which represents the distance between each data point and the cluster centroid [29]. Figure 1 depicts a flow chart in which explain the steps to perform a clustering algorithm.

Figure 1. Flowchart of k-means-based clustering



K-means is a partitioning algorithm that relies on the concept of distance and local optimization to perform clustering. One of the most common metrics to compute distances in k-means is the Euclidean distance, as it is flexible to accommodate different operational situations. Another characteristic of the algorithm is that it requires the user to specify the number *k* of clusters (step 1 in Figure 1). The algorithm will always converge, but it is vulnerable to local minima. This will depend on how centroids

are initialized. By running the algorithm with a specified number of clusters k, k random samples from the dataset are allocated as cluster centroids.

After the selection of k, the main steps of the k-means clustering algorithm are:

- Initialization: the step to choose k initial centroids (step 2 in Figure 1)
- Looping: iterative steps to stabilize centroids until reaching convergence or a maximum number of iterations (steps 3, 4, and 5 in Figure 1). This loop requires two sub-steps:
    - Assigning samples to their nearest centroid based on a selected distance measure.
    - Compute the mean of the assigned samples and create a new centroid.

K-means with Euclidean distance has been used to map countries' clusters as they appear in the EM-DAT database.

Clusters must be validated to check the logical cohesion between the clustered items and to compare the separation among them. A useful metric for validating the significance of clusters is the silhouette, whose scores represent the distance from one sample to the samples in the neighboring clusters [30]. Silhouette coefficients range between -1 and 1 where values close to 1 indicate high compactness within the cluster, which in turn implies longer distances among the sample and the neighboring clusters. Silhouette scores close to 0 indicate overlapping clusters, while negative values indicate a possible misplacement of the sample [31].

Within this case study, the algorithm runs on a set of selected features considered relevant for the scope of the analysis: World region, Disaster count, Missing data, Gross Domestic Product based on Purchasing Power Parity (GDP PPP), Population density, Disaster type, Total deaths. It is worth mentioning that GDP PPP and Population Density data are not available in EM-DAT and have been retrieved from external sources [32], [33]. In addition, categorical features have been converted into numerical features and standardized through z-score normalization.

4. RESULTS AND DISCUSSION

4.1. Clusters

The clustering algorithm allowed splitting the 189 countries involved in natural hazard accidents into 40 clusters of varied sizes. Considering the relatively large number of clusters, a complete review would be impractical. Therefore, a selection of the most interesting clusters is presented. Two criteria have been considered in the selection: cumulative number of fatalities and cluster compactness. Also, clusters that comprise only one country are treated separately.

The cluster with the highest cumulative number of fatalities and with more than two countries is:

Cluster 1. Bangladesh, France, Germany, Japan, Poland, South Korea, and Vietnam.

The cluster with the largest average intra-cluster silhouette score is:

Cluster 2. Cayman Islands, Saint Kitts and Nevis, and Turks and Caicos Islands.

On the other hand, the cluster with the smallest silhouette score is:

Cluster 3. Jamaica, Madagascar, Mauritius, Sint Maarten.

In addition, 11 clusters include only one country. Examples of these clusters are:

Cluster 4. China;
Cluster 5. India;
Cluster 6. USA.

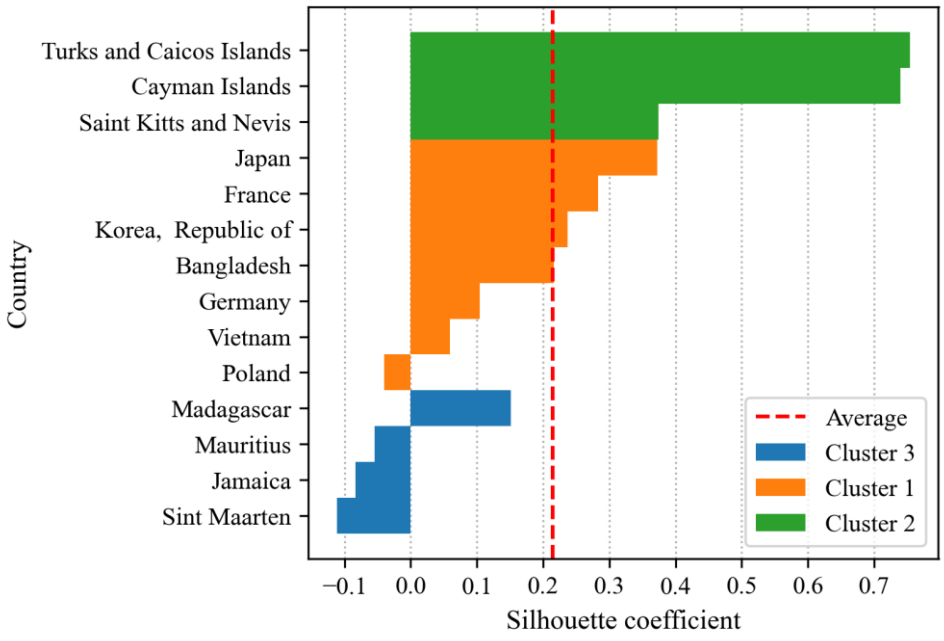
Relevant information about the countries in each cluster is summarized in Table 1. For each country, Table 1 displays the number of fatalities, the most frequent and severe natural disaster types, the location, and the income group [34]. Clusters and countries are displayed in descending order of number of fatalities.

Table 1. Relevant information about the countries in the selected clusters

Cluster	Country	Fatalities	Threats (fatalities)	Location	Income [34]
4	China	17.006.913	Flood (10.321.805) Drought (6.503.534)	East Asia	Upper-Middle
5	India	4.515.665	Cyclone (160.575) Drought (4.250.320)	South Asia	Lower-Middle
1	Bangladesh	2.590.573	Cyclone (627.048) Drought (1.900.018)	South Asia	Lower-Middle
	Japan	50.565	Cyclone (32.838) Flood (13.513)	East Asia	High
	France	28.793	Heatwave (27.517)	Western Europe	High
	Vietnam	26.025	Cyclone (19.189) Flood (3644)	Southeast Asia	Lower-Middle
	Germany	10.213	Heatwave (9361)	Western Europe	High
	South Korea	8932	Cyclone (3727)	East Asia	High
	Poland	2378	Cold wave (2085)	Central Europe	High
6	USA	41.359	Storm (30.942)	North America	High
3	Madagascar	3161	Cyclone (2834)	Sub-Saharan Africa	Low
	Jamaica	1391	Flood (730) Cyclone (604)	Caribbean	Upper -Middle
	Mauritius	81	Cyclone (28) Flash flood (11)	Sub-Saharan Africa	Upper-Middle
	Sint Maarten	4	Cyclone (4)	Caribbean	High
2	Saint Kitts and Nevis	6	Cyclone (6)	Caribbean	High
	Cayman Islands	2	Cyclone (2)	Caribbean	High
	Turks and Caicos	0		Caribbean	High

The silhouette plot of the selected clusters is displayed in Figure 2. Clusters that comprise only one country have not been included because their silhouette score equals zero.

Figure 2. Silhouette plot of the selected clusters.



The average silhouette score of the selected clusters (red vertical line in Figure 1) is equal to 0.21. Instead, the average silhouette score is 0.24 if all the 40 clusters are considered. It is also worth noting that Poland, Mauritius, Jamaica, and Sint Maarten show negative silhouette scores.

4.2. Discussion

Data in Figure 2 and Table 1 allows some observations on cluster composition, similarities and differences between members of the same clusters, and opportunities for inter-country knowledge sharing. In the remainder of this section, each cluster will be briefly commented, similarities and differences will be discussed in terms of fatalities, disaster type, development level, and economic possibilities. The discussion will focus on investigating whether the proposed method has effectively grouped countries that show similarities and subsequently can drive to common natural risk management mitigations.

The three most populous countries in the world – i.e., China, India, and the USA – belong to stand-alone clusters and have not been considered similar to any other country in the dataset. This result is not surprising considering the unique characteristics of these countries in terms of location, area, exposure to natural threats, and economy. The number of fatalities in India and China is respectively one and two orders of magnitude larger than any other cluster. Also, China has a unique exposure to riverine floods and drought, which together caused 99% of the total number of fatalities (Table 1). India on the other hand is naturally exposed to severe droughts, which have caused more than 94% of the total deaths (Table 1). From an economic perspective, China has witnessed extraordinary growth in the last three decades and is currently the second-largest economy by GDP (Gross Domestic Product) in the world after the USA [35]. On the other hand, India is the sixth-largest economy, and its annual growth rate in terms %GDP has been larger than the USA but smaller than China since 1990 [36]. Geographically, China, India, and the USA are respectively the third, fourth, and seventh-largest countries by area [37], and they cross various climate zones [38]. In light of their unique characteristics, the grouping of these countries in stand-alone clusters appears reasonable. Also, a large body of research has focused on the study of climate-driven disasters in these countries [39]–[43]. Existing studies and governmental mitigation and response plans might be good opportunities to (i) share knowledge and lesson learned between these three countries and (ii) provide critical assistance to smaller, less-developed countries which have similar exposure to climate-driven events (e.g., Vietnam concerning flooding and Bangladesh concerning storms and droughts).

Regarding clusters with more than one country (i.e., clusters 1, 2, and 3), it can be observed that some clusters show apparent internal similarities while others are more difficult to interpret. For instance, cluster 2 was chosen because its members have the largest average similarity score (Figure 2), which indicates high compactness and separability [44]. Indeed, countries in this cluster, namely Saint Kitts and Nevis, Cayman Islands, and Turks and Caicos, are extremely similar: they all are archipelagos in the Caribbean Sea, classified as high-income countries, with a relatively low number of fatalities. Due to their location, the islands have been affected by several cyclones and storms. Nevertheless, the number of climate-related deaths is extremely low. Considering the already significant success of these countries in coping with tropical storms, there might be little scope for inter-country knowledge sharing. However, islands in different clusters with similar exposure to natural threats (e.g., Fiji Islands) may be inspired by the measures adopted by the countries in cluster 2. In other words, although knowledge transfer between countries in high-compact clusters may not appear interesting, there are still interesting learning opportunities for countries that have similar exposure but that were put in a different cluster due to significant differences in, e.g., the number of fatalities.

Cluster 3 was selected for the low silhouette score of its members, which are Madagascar, Jamaica, Mauritius, and Sint Maarten. Three out of four countries show a negative silhouette score, indicating low compactness and separability [44]. However, it is still possible to spot some similarities between the members of this cluster, which are islands or archipelagos, relatively close to each other in pairs. In spite of the differences, a more detailed analysis might reveal hidden similarities and interesting learning opportunities.

Cluster 1 was chosen because it shows the largest cumulative number of fatalities between clusters with more than two countries; therefore, it is definitely the most critical and interesting within the whole database. The cluster comprises Bangladesh, Japan, France, Vietnam, Germany, South Korea, and

Poland. Figure 2 shows that the members of this cluster have positive silhouette scores except for Poland, whose score is -0.04. Therefore, Poland may be considered an outlier and will not be considered further in the analysis. Interestingly, in spite of the relatively high compactness, cluster 1 appears rather heterogeneous. It comprises countries from different locations, with diverse socio-economic backgrounds and exposure to natural hazards. It is not trivial to identify similarities in this cluster. However, this should not be perceived as a limitation. On the contrary, comparing countries with both similarities and differences in disaster situations, demographics, and economy might be extremely interesting. It is not desirable to create 'perfect' clusters of countries for natural disaster comparison. A cluster of neighboring countries with the exact same possibilities and disaster situations is not advisable because there is little room for improvements and knowledge transfer. For cross-country learning to be relevant and helpful, it is beneficial that some countries are more exposed, developed, or prepared for disasters than others. However, countries should also exhibit some similarities in the disaster patterns and threats to facilitate comparison and the creation of actionable insights.
In light of these considerations, clustering algorithms must be considered tools to reveal similarities and guide the analysis towards countries that may be more interesting to compare. However, in-depth analyses are still needed to make sense of data, interpret clusters, discover hidden similarities, and enable cross-country learning and knowledge transfer. In other words, clustering algorithms have the potential to greatly simplify the analysis by removing the need for manual screening. However, human intervention and expert knowledge are needed to convert groups of related countries into actionable insights.

Considering the number of fatalities, Bangladesh can be regarded as an outlier due to its extreme history. The total number of climate-driven natural disaster fatalities in the country has been almost 2.6 million since the year 1900. Manual analysis of the EM-DAT database reveals that despite an increasing trend in the number of critical events, fatalities have decreased in recent times. Specifically, in the time period following 1992, the number of deaths has decreased, major outliers were less frequent, and resulted in fewer deaths. Nevertheless, tropical cyclones and storm surges have been particularly severe since 1900 [45]. The decreasing fatalities in spite of an increasing number and severity of cyclones suggest a significant improvement in mitigating measures. From an economic point of view, Bangladesh is relatively less developed than the other members of the cluster. GPD value is larger than Vietnam's but significantly lower than the other counties.

Vietnam is the fourth country in cluster 1 in terms of total fatalities. Similar to Bangladesh, the country is exposed to tropical cyclones and storms, although the number and severity of critical events are lower. Similar to Japan, a relevant part of the fatalities is caused by floods. Nevertheless, Vietnam has experienced a decreasing trend of fatalities in the past 20 years, although the decrease is less pronounced than in Bangladesh. From an economic point of view, Vietnam went from being one of the poorest countries in the world to becoming a lower-middle income country [35]. However, the development in Vietnam started later compared to other developed Asian countries like South Korea and Japan, but the growth rate has been faster than in Bangladesh [46].

Japan is the second country in cluster 1 in terms of total fatalities. Due to its location and geography, the country is particularly exposed to tropical cyclones, storms, localized rains, and floods [47]. However, the relatively large number of fatalities does not indicate unpreparedness or ineffective response to natural hazards. On the contrary, the continuous exposure to natural threats pushed the country towards increasingly effective mitigation measures [48], [49]. In fact, more than 82% of the total deaths were registered before 1960. After that year, the number of fatalities decreased drastically and has remained relatively stable. However, the trend has reversed during the last 20 years, and the number of fatalities has slowly returned to grow. This change may be related to the increasing frequency and severity of natural hazards. It is also worth mentioning that a relatively new type of event, namely heatwaves, has caused the most fatalities in the last ten years. Specifically, heat waves have caused 735 deaths since 2010, while tropical cyclones and floods caused 591 and 447 fatalities over the same period. Interestingly, heatwaves caused only 135 events from 1900 to 2010. The recent increasing trend in the number of fatalities differentiates Japan from Bangladesh and Vietnam. From an economic point

of view, the Second World War had marked the beginning of extraordinary growth for Japan, which is currently one of the leading industrialized countries in the world.

Germany and France have significantly fewer fatalities than Bangladesh and Japan, and they are the only European countries in the cluster. Another difference is that most fatalities in Germany and France occurred after 2000 and were primarily caused by heatwaves. For instance, the heatwave of 2003 is responsible for 68% and 92% of the total deaths registered in France and Germany, respectively. This may indicate that rising global temperatures and climate change have affected countries that were not significantly exposed to natural threats in earlier times [50]. France and Germany have strong and stable economies and are respectively the seventh and fourth countries in the world in terms of GDP [35].

South Korea is the country with the least number of fatalities in cluster 1. Like Japan and Vietnam, South Korea is exposed to storms and floods, which are responsible for most deaths. However, extreme events are less frequent and intense in South Korea than in the other Asian countries in cluster 1. Also, the dataset analysis reveals a downward trend in the number of fatalities. Overall, South Korea is less exposed to natural hazards than Japan, Vietnam, and Bangladesh. However, more frequent and severe events are expected in the future to the effect of climate change [51]. From an economic point of view, the country grew from being a lower-income before 1980 to be a high-income economy in 1995 and currently the tenth country in the world in terms of GDP.

In light of the considerations made for countries in cluster 1, the following suggestions and learning opportunities may be identified:

1. Vietnam and Bangladesh may be considered similar with respect to exposure to tropical cyclones. In addition, both the countries are low-middle income economies. Nevertheless, Bangladesh has been more successful in mitigating the effect of extreme events. Therefore, Vietnam could be inspired and learn from the affordable mitigating measures implemented in Bangladesh.
2. Japan offers significant learning opportunities for Vietnam and Bangladesh because it has similar exposure and has invested many resources into natural disaster prevention and mitigation policies. Less developed countries could greatly benefit from the lessons learned by countries with more financial resources.
3. The number of deaths in Bangladesh and Vietnam decreased during the last two decades, while the trend has inverted in Japan. This may be due to, e.g., increased elderly population, urbanization, and coastal moving, which all imply that more people are exposed to natural hazards. Future building and infrastructure plans should consider natural risks in order to avoid turning common hazards into major catastrophes due to demographic changes and population growth.
4. Considering the effect of climate change and the increasing global temperatures, it might be beneficial for the countries that have not experienced severe heat waves (e.g., Vietnam and Korea) to learn from countries that have been severely affected (e.g., France and Germany) in order to improve awareness and preparedness to possible extreme temperature events in the future.
5. Germany and Korea appear to be the less vulnerable countries in the cluster. Therefore, they should pay close attention to the current changes in trends and improve hazard preparedness. The less vulnerable, developed countries have economies that facilitate research on innovative mitigation measures. The focus should be to create low-cost, high-impact measures since natural disasters cause more harm to poorer countries and tend to worsen poverty and unemployment.

In general, the countries in cluster 1 offered interesting insights and discussion points. This suggests that the clustering procedure has successfully identified groups of countries that share similar characteristics and can benefit from each other's experiences and lessons. However, it must be recalled that the analysis of clusters requires manual intervention and expert knowledge to, e.g., interpret and evaluate the results of the clustering procedure, identify hidden similarities and differences between countries, analyze trends and recognize learning opportunities. Therefore, the results from this example of ML clustering for risk management purposes show how the techniques used require a deep understanding of their benefits, limitations, and application boundaries. For this reason, this

contribution aims to convey the message that ML-based techniques must be considered as tools supporting and not substituting decision-making.

Awareness and knowledge of these tools properties by the user is essential to effectively exploit their results. The role of the human as user of these tools is even more central than before. ML should not be intended as a way to replace the human, but only as an improved approach assisting the human. This is conform with the concept of trustworthy AI by the European Commission [52] promoting explainable AI (XAI) human centrality by means of interpretability, info-besity (overload of information) avoidance, and transparency.

5. CONCLUSION

Considering the widespread adoption of AI and ML algorithms, many wonder whether we are proceeding toward a "no-brainer" era, where machines will be in charge of critical decisions, and human knowledge will have only a marginal role. This issue is especially important in the context of risk assessment and management, where errors may result in fatalities and significant economic losses. This study suggests that we are not yet close to such a condition since humans still play a key role in the decision-making process. In addition, we claim that ML algorithms may provide critical support and better-informed decision-making if certain conditions are met. These conditions include knowing (i) what the algorithm does, (ii) how it does it, and (iii) what the limitations are. We discuss this topic through an example of clustering of climate-driven natural disasters. EM-DAT dataset is used as the data source, and k-means is used to group countries that share similar characteristics with respect to exposure to natural disasters. The cluster analysis revealed underlying causes and influential factors that can inform decision-making and enable cross-country learning. However, the objective of this investigation is not to present and discuss an example of "perfect" clustering. On the contrary, the overall intent is to show that effective deployment of ML models must consider the role of humans in the design of the algorithms and interpretation of the results. This study shows that human knowledge still plays a pivotal role in developing and implementing ML algorithms. For example, expert knowledge is required for features selection, model hyperparameters tuning, evaluation strategy selection, and, more importantly, cluster analysis and interpretation. These steps involve human intervention and, therefore, heavily rely on human knowledge. In light of these considerations, ML algorithms are to be considered (advanced) tools, and like most tools, they are only as good as their users. Therefore, AI and ML must be considered powerful and reliable tools to extract hidden patterns from data and provide suggestions to decision-makers; however, humans are still essential to interpret those suggestions and, eventually, convert recommendations into actions.

Acknowledgements

Nicola Paltrinieri is an International Research Fellow of the Japan Society for the Promotion of Science.

References

[1] J. E Diekmann, “Risk analysis: lessons from artificial intelligence,” *Int. J. Proj. Manag.*, vol. 10, no. 2, pp. 75–80, 1992, doi: http://dx.doi.org/10.1016/0263-7863(92)90059-I.

[2] F. Hsu, M. S. Campbell, and A. J. Hoane Jr, “Deep Blue system overview,” in *Proceedings of the 9th international conference on Supercomputing*, 1995, pp. 240–244.

[3] G. Kasparov, *Deep thinking: where machine intelligence ends and human creativity begins*. Hachette UK, 2017.

[4] N. Paltrinieri, L. Comfort, and G. Reniers, “Learning about risk: Machine learning for risk assessment,” *Saf. Sci.*, vol. 118, no. July 2018, pp. 475–486, 2019, doi: 10.1016/j.ssci.2019.06.001.

[5] I. Goodfellow, Y. Bengio, and A. Courville, *Deep Learning*. MIT Press, 2016.

[6] A. L. Samuel, “Some studies in machine learning using the game of checkers,” *IBM J. Res. Dev.*, vol. 44, no. 1–2, pp. 207–219, 1959, doi: 10.1147/rd.441.0206.

[7] A. Burkov, *Machine Learning Engineering*, vol. 1. True Positive Incorporated, 2020.

[8] R. Sharda, D. Delen, and T. Efraim, *Analytics, Data Science, & Artificial Intelligence: Systems*

for Decision Support, Eleventh e. Hoboken, NJ: Pearson, 2019.
[9] F. De Felice, M. Travaglioni, G. Piscitelli, R. Cioffi, and A. Petrillo, "Machine learning techniques applied to industrial engineering: A multi criteria approach," *18th Int. Conf. Model. Appl. Simulation, MAS 2019*, pp. 44–54, 2019, doi: 10.46354/i3m.2019.mas.007.
[10] V. Villa, N. Paltrinieri, F. Khan, and V. Cozzani, "Towards dynamic risk analysis: A review of the risk assessment approach and its limitations in the chemical process industry," *Saf. Sci.*, vol. 89, pp. 77–93, 2016, doi: 10.1016/j.ssci.2016.06.002.
[11] A. J. Nakhal A, R. Patriarca, G. Di Gravio, G. Antonioni, and N. Paltrinieri, "Investigating occupational and operational industrial safety data through Business Intelligence and Machine Learning," *J. Loss Prev. Process Ind.*, vol. 73, p. 104608, 2021, doi: https://doi.org/10.1016/j.jlp.2021.104608.
[12] K. Øien, I. B. Utne, and I. A. Herrera, "Building Safety indicators: Part 1 - Theoretical foundation," *Saf. Sci.*, vol. 49, no. 2, pp. 148–161, 2011, doi: 10.1016/j.ssci.2010.05.012.
[13] N. Paltrinieri, K. Øien, and V. Cozzani, "Assessment and comparison of two early warning indicator methods in the perspective of prevention of atypical accident scenarios," *Reliab. Eng. Syst. Saf.*, vol. 108, pp. 21–31, Dec. 2012, doi: 10.1016/j.ress.2012.06.017.
[14] N. Paltrinieri and F. Khan, *Dynamic Risk Analysis in the Chemical and Petroleum Industry: Evolution and Interaction with Parallel Disciplines in the Perspective of Industrial Application*, 1st ed. Butterworth-Heinemann, 2016.
[15] G. Landucci and N. Paltrinieri, "A methodology for frequency tailorization dedicated to the Oil & Gas sector," *Process Saf. Environ. Prot.*, vol. 104, pp. 123–141, 2016, doi: 10.1016/j.psep.2016.08.012.
[16] N. Paltrinieri and G. Reniers, "Dynamic risk analysis for Seveso sites," *J. Loss Prev. Process Ind.*, vol. 49, pp. 111–119, 2017, doi: 10.1016/j.jlp.2017.03.023.
[17] X. Luo, A. M. Cruz, and D. Tzioutzios, "Extracting Natech Reports from Large Databases: Development of a Semi-Intelligent Natech Identification Framework," *Int. J. Disaster Risk Sci.*, vol. 11, no. 6, pp. 735–750, 2020, doi: 10.1007/s13753-020-00314-6.
[18] R. R. Arinta and E. Andi W.R., "Natural Disaster Application on Big Data and Machine Learning: A Review," 2019.
[19] M. Yu, C. Yang, and Y. Li, "Big Data in Natural Disaster Management: A Review," *Geosciences*, vol. 8, no. 5, 2018.
[20] O. Department of Regional Development and Environment Executive Secretariat for Economic and Social Affairs, "Chapter 2 - Natural Hazard Risk Reduction in roject Formaulation and Evaluation," 1991. .
[21] S. Sarkar and J. Maiti, "Machine learning in occupational accident analysis: A review using science mapping approach with citation network analysis," *Saf. Sci.*, vol. 131, p. 104900, 2020.
[22] P. N. Lal, R. Singh, and P. Holland, "Relationship between natural disasters and poverty a Fiji case study," 2009. [Online]. Available: https://www.iucn.org/sites/dev/files/import/downloads/poverty_a_fiji_case_study_final020509.pdf.
[23] R. Masika, *Gender, Development and Climate Change*, no. 2008. Oxford: Oxfam GB, 2013.
[24] A. M. Cruz, L. J. Steinberg, and A. L. Vetere-Arellano, "Emerging issues for natech disaster risk management in Europe," *J. Risk Res.*, vol. 9, no. 5, pp. 483–501, Jul. 2006, doi: 10.1080/13669870600717657.
[25] M. C. Suarez-Paba and A. M. Cruz, "A paradigm shift in Natech risk management: Development of a rating system framework for evaluating the performance of industry," *J. Loss Prev. Process Ind.*, vol. 74, Jan. 2022, doi: 10.1016/J.JLP.2021.104615.
[26] A. Jacobsson, J. Sales, and F. Mushtaq, "A sequential method to identify underlying causes from industrial accidents reported to the MARS database," *J. Loss Prev. Process Ind.*, vol. 22, no. 2, pp. 197–203, Mar. 2009, doi: 10.1016/j.jlp.2008.12.009.
[27] Ceneter Center for research on the Epidemiology of Disasters, "The international Disaster Database," 2021. .
[28] Y. Zhang, J. Mańdziuk, C. H. Quek, and B. W. Goh, "Curvature-based method for determining the number of clusters," *Inf. Sci. (Ny).*, vol. 415–416, pp. 414–428, 2017.
[29] T. S. Chen *et al.*, "A combined K-means and hierarchical clustering method for improving the

clustering efficiency of microarray," *Proc. 2005 Int. Symp. Intell. Signal Process. Commun. Syst. ISPACS 2005*, vol. 2005, pp. 405–408, 2005, doi: 10.1109/ispacs.2005.1595432.
[30] S. K. Kingrani, M. Levene, and D. Zhang, "Estimating the number of clusters using diversity," *Artif. Intell. Res.*, vol. 7, no. 1, p. 15, 2017, doi: 10.5430/air.v7n1p15.
[31] G. W. Milligan and M. C. Cooper, "An examination of procedures for determining the number of clusters in a data set," *Psychometrika*, vol. 50, no. 2, pp. 159–179, 1985.
[32] The World Bank, "GDP, PPP (current international $)," 2022. https://data.worldbank.org/indicator/NY.GDP.MKTP.PP.CD.
[33] OpenGeoCode, "Countries of the World COW," 2012. https://web.archive.org/web/20150319012353/http://opengeocode.org/cude/dow%0Anload.php?file=/home/fashions/public_html/opengeocode.org/download/cow.txt.
[34] The World Bank, "World Bank Country and Lending Groups," 2022. https://datahelpdesk.worldbank.org/knowledgebase/articles/906519-world-bank-country-and-lending-groups.
[35] The World Bank, "GDP (current US$)," 2022. https://data.worldbank.org/indicator/NY.GDP.MKTP.CD?most_recent_value_desc=true.
[36] The World Bank, "GDP growth (annual %) - China, India, United States," 2022. https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?end=2020&locations=CN-IN-US&start=1961&view=chart (accessed Apr. 01, 2022).
[37] Worldometer, "Largest Countries in the World (by area)." https://www.worldometers.info/geography/largest-countries-in-the-world/.
[38] H. E. Beck, N. E. Zimmermann, T. R. McVicar, N. Vergopolan, A. Berg, and E. F. Wood, "Present and future Köppen-Geiger climate classification maps at 1-km resolution," *Sci. Data*, vol. 5, no. 1, p. 180214, 2018, doi: 10.1038/sdata.2018.214.
[39] G. Shen, L. Zhou, Y. Wu, and Z. Cai, "A Global Expected Risk Analysis of Fatalities, Injuries, and Damages by Natural Disasters," *Sustainability*, vol. 10, no. 7, 2018.
[40] W. Chen, S. L. Cutter, C. T. Emrich, and P. Shi, "Measuring social vulnerability to natural hazards in the Yangtze River Delta region, China," *Int. J. Disaster Risk Sci.*, 2013.
[41] G. Greenough, M. McGeehin, S. M. Bernard, J. Trtanj, J. Riad, and D. Engelberg, "The potential impacts of climate variability and change on health impacts of extreme weather events in the United States.," *Environ. Health Perspect.*, vol. 109, pp. 191–198, 2001.
[42] H. Pandve, "India′s National Action Plan on Climate Change," *Indian J. Occup. Environ. Med.*, vol. 13, no. 1, p. 17, 2009, doi: 10.4103/0019-5278.50718.
[43] S. V. R. K. Prabhakar and R. Shaw, "Climate change adaptation implications for drought risk mitigation: a perspective for India," *Clim. Change*, vol. 88, no. 2, pp. 113–130, 2008.
[44] J. Han, M. Kamber, and J. Pei, *Data Mining: Concepts and Techniques*. Elsevier Inc., 2012.
[45] U. Haque, M. Hashizume, K. N. Kolivras, H. J. Overgaard, B. Das, and T. Yamamoto, "Reduced death rates from cyclones in Bangladesh: what more needs to be done?," *Bull. World Health Organ.*, vol. 90, no. 2, pp. 150–156, Feb. 2012, doi: 10.2471/BLT.11.088302.
[46] The World Bank, "GDP growth (annual %) - Bangladesh, Vietnam," 2022. https://data.worldbank.org/indicator/NY.GDP.MKTP.KD.ZG?end=2020&locations=BD-VN&start=1961&view=chart (accessed Apr. 01, 2022).
[47] H. Sasaki and S. Yamakawa, "Natural Hazards in Japan," in *International Perspectives on Natural Disasters: Occurrence, Mitigation, and Consequences*, 2007, pp. 163–180.
[48] T. Haruming Tyas, S. Sutisna, M. Supriyatno, I. D. K. K. Widana, and A. Fatkul Fikri, "Lesson Learned from Japan for Flood Disaster Risk Reduction in Indonesia," *Tech. Soc. Sci. J.*, 2022.
[49] Ministry of Foreign Affairs of Japan, "Disasters and Disaster Prevention in Japan," 2022. https://www.mofa.go.jp/policy/disaster/21st/2.html (accessed Apr. 02, 2022).
[50] S. C. Sheridan and M. J. Allen, "Changes in the Frequency and Intensity of Extreme Temperature Events and Human Health Concerns," *Curr. Clim. Chang. Reports*, 2015.
[51] L. Miyeon, H. J. Ho, and K. K. Yul, "Estimating Damage Costs from Natural Disasters in Korea," *Nat. Hazards Rev.*, vol. 18, no. 4, 2017.
[52] EC's High Level Expert Group on AI, "Draft Ethics Guidelines for Trustworthy AI," Brussels, Belgium, 2018.