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
0098-1354/

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

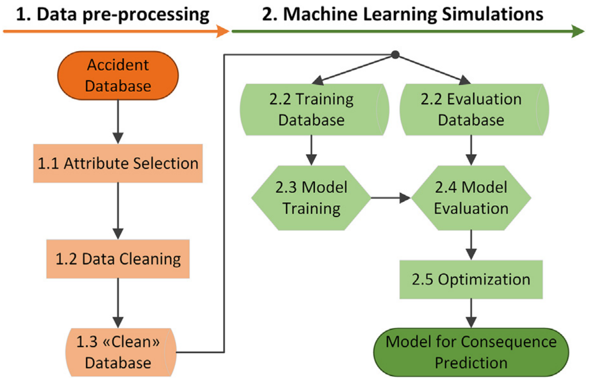

Fig. 1. Methodology workflow. Colors represent two main stages: Data pre-processing (orange), and Machine Learning Simulations (green).

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;
- β_0 = bias;
- X_j = a feature;
- β_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).

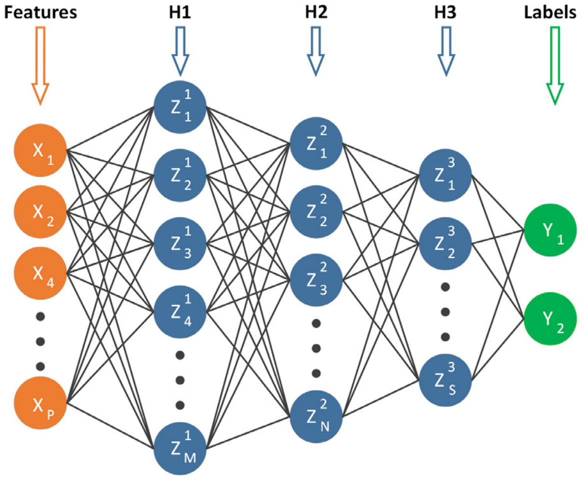

Fig. 2. Schematic representation of a Deep Neural Network. Orange, blue, and green circles represent input features (X_i), hidden units (Z_i^j), and labelsY_k. Adapted from Tamascelli et al. (2020a).

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-

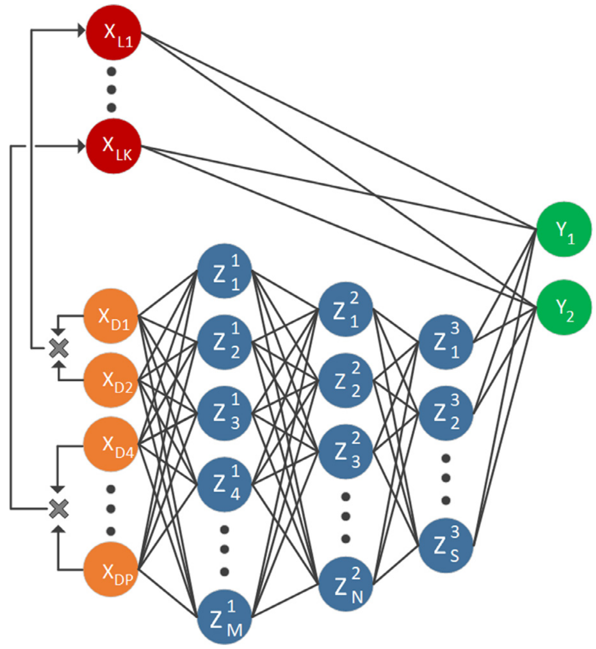

Fig. 3. Schematic representation of a Wide&Deep model, which consists of a deep part (bottom) and a wide part (top). The deep part is a DNN and takes as an input a full set of features (X_{Di}). The wide part is a Linear model and takes as an input a small set of crossed-features (X_{Li}).

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

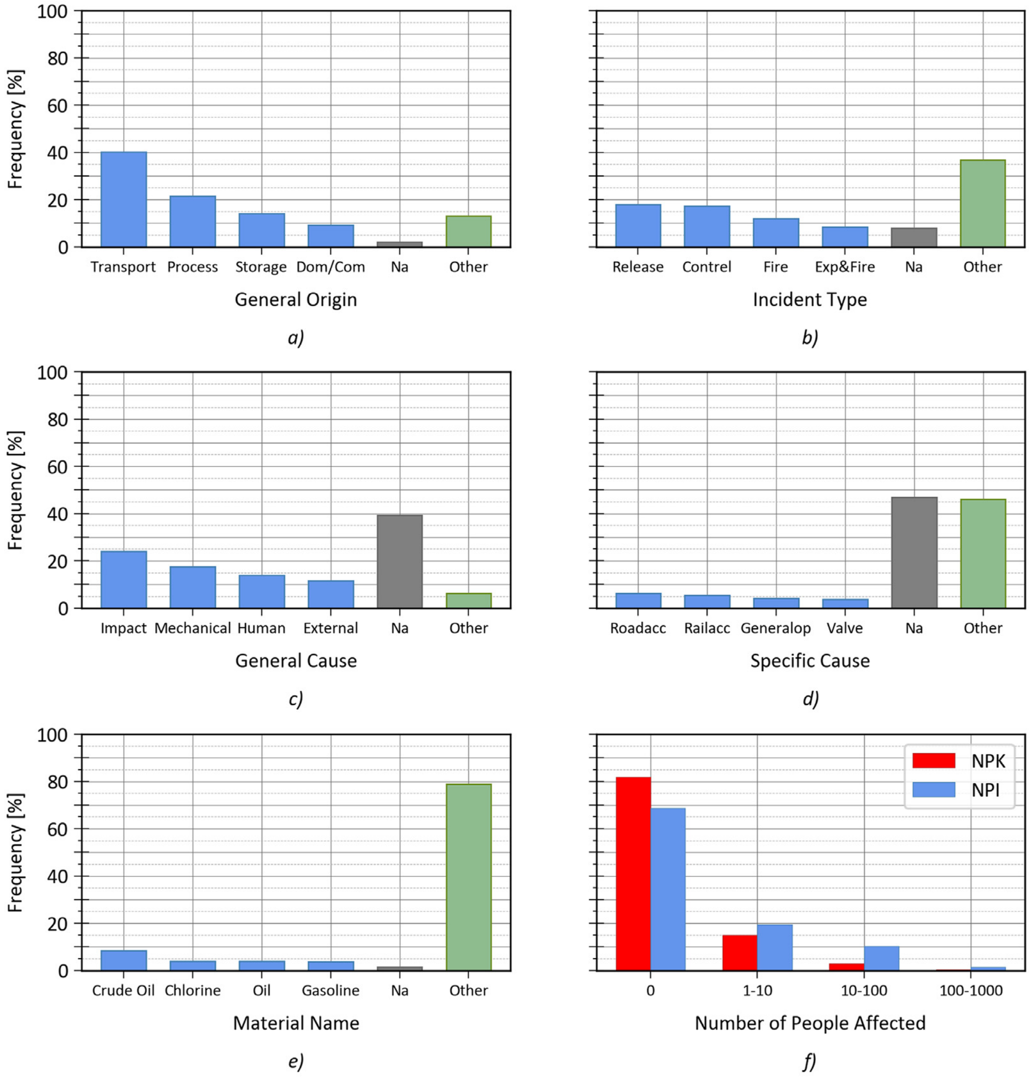

Fig. 4. Frequency distribution of the attributes GOG (a), IT (b), GC (c), SC (d), MN (e), NPK and NPI (f). “Na” refers to missing values, “Other” refers to attribute codes that have not been represented in the figure.

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.

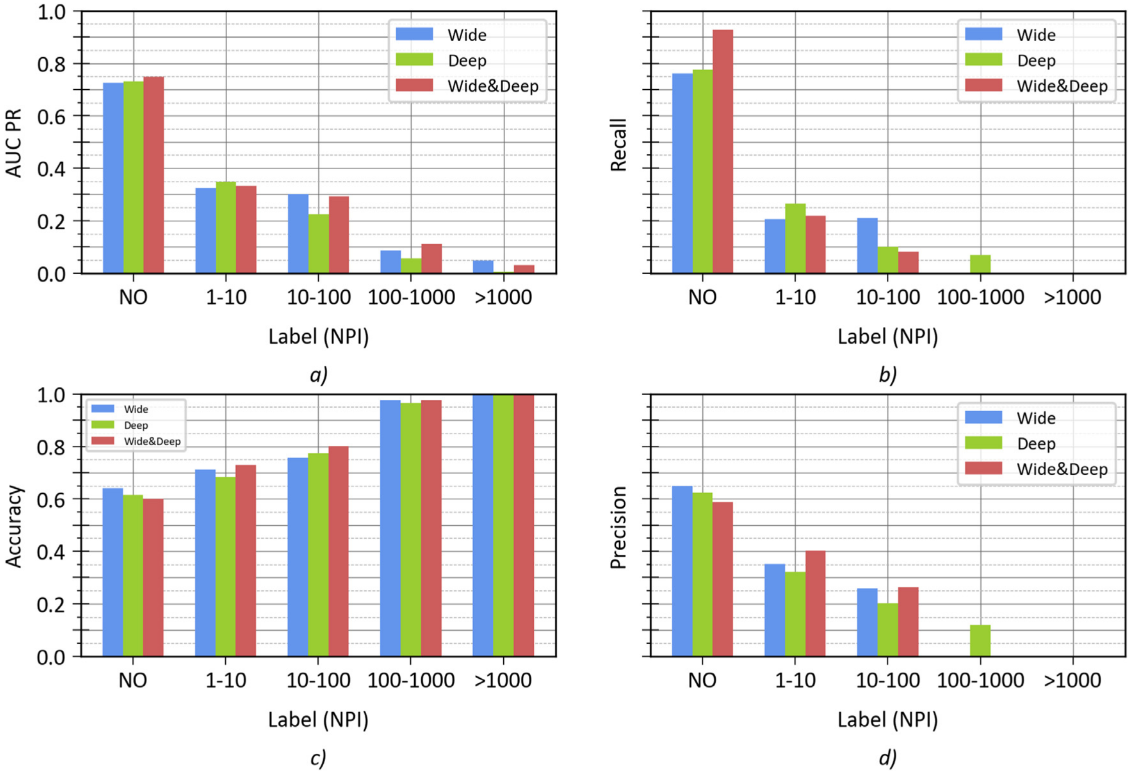

Fig. 5. Area Under the Curve Precision-Recall (AUC PR) (a), Recall (b), Accuracy (c), and Precision (d) obtained from a small selection of simulations for the category “Number of People that are Injured” (NPI). Labels are represented on the x-axis. Recall, Precision, and Accuracy are obtained at threshold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

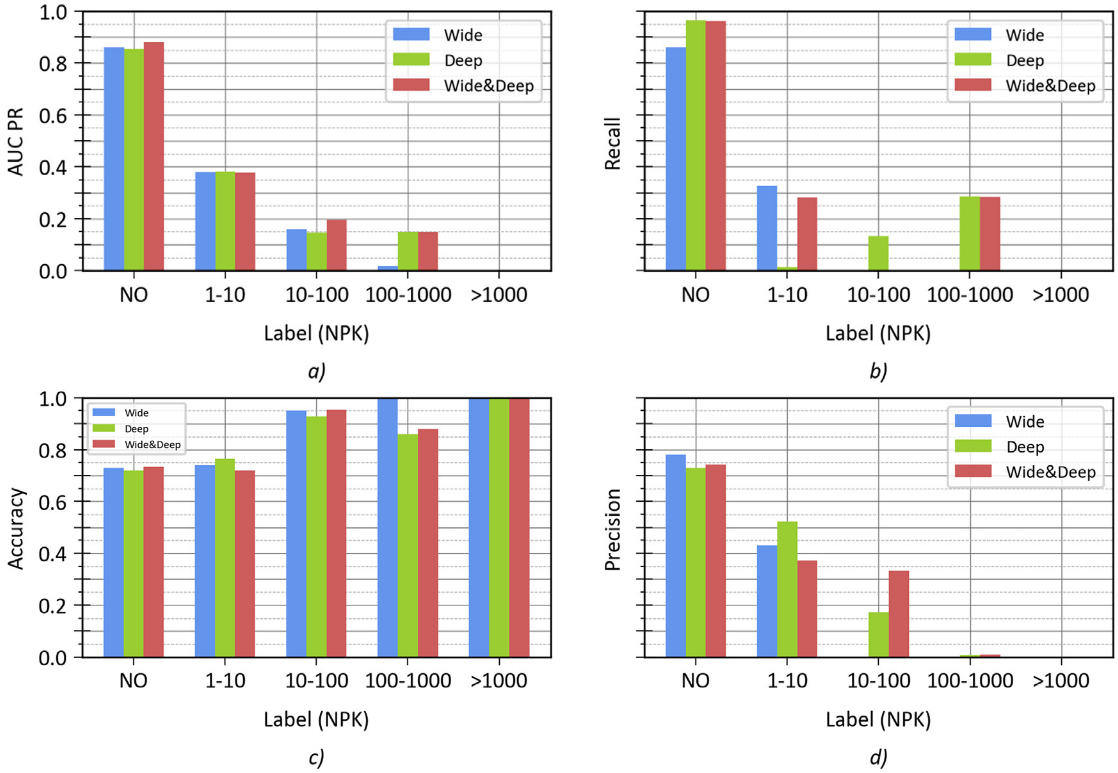

Fig. 6. Area Under the Curve Precision-Recall (AUC PR) (a), Recall (b), Accuracy (c), and Precision (d) obtained from a small selection of simulations for the category “Number of People that are Killed” (NPK). Labels are represented on the x-axis. Recall, Precision, and Accuracy are obtained at threshold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

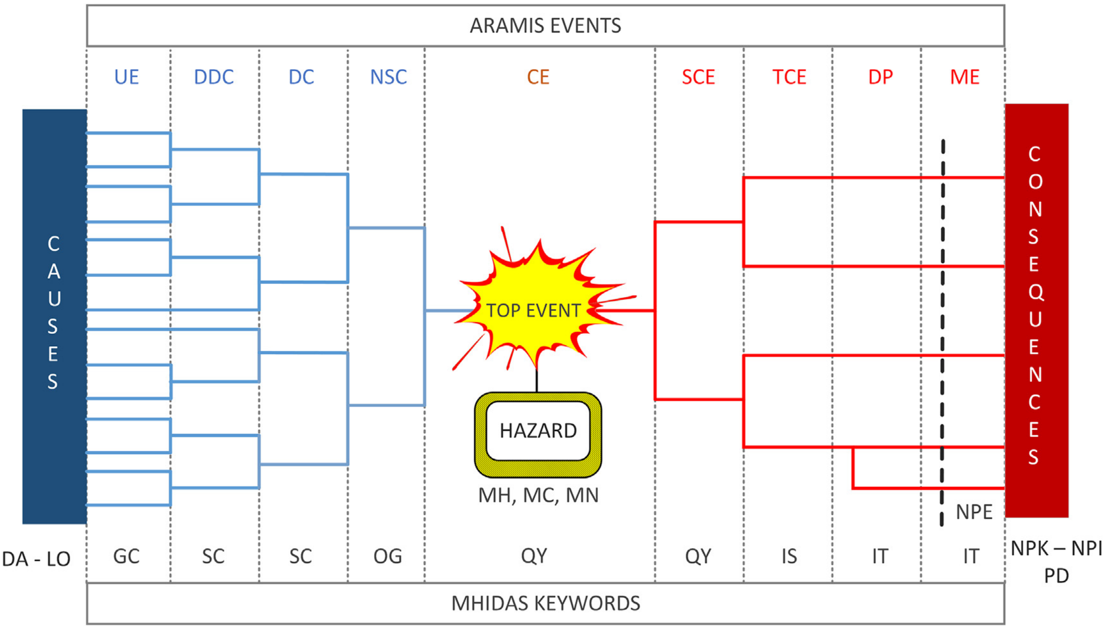

Fig. 7. Schematic representation of a Bow-Tie diagram. Names of intermediate events (top side) are defined according to MIMAH methodology (ARAMIS project team, 2004). Database attributes listed in Table 2 are associated with each intermediate event (bottom side). The bold dashed line indicates that the 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.

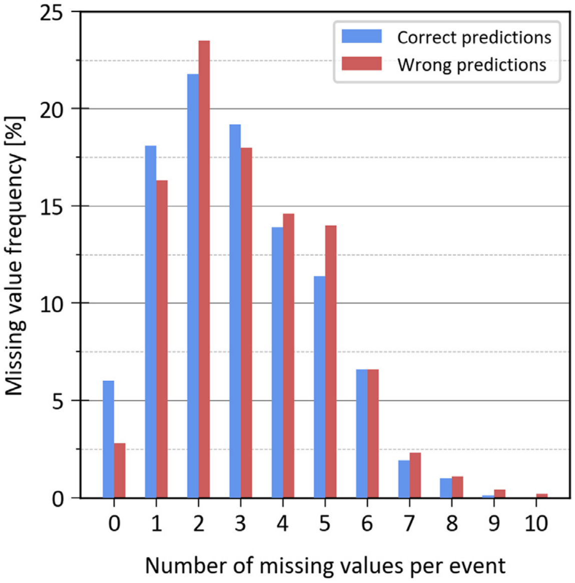

Fig. 8. Missing feature distribution among correct and wrong predictions made by the Wide&Deep model (category = "NPK", label = 1 – 10).

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

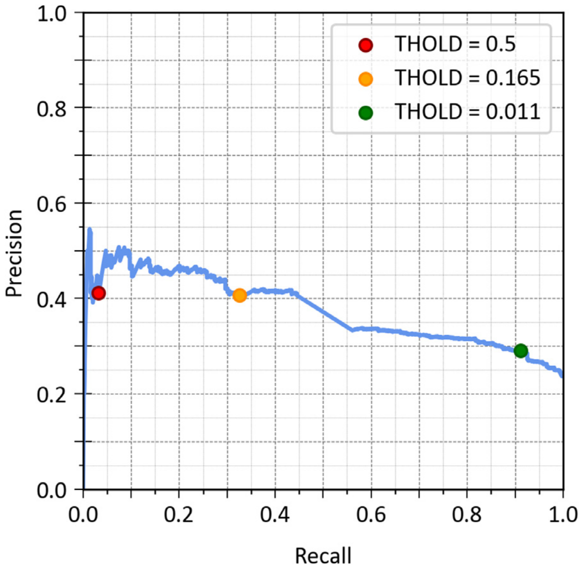

Fig. 9. Precision-Recall curve of the Deep model for the label 1 – 10 (NPK) at 2000 integration steps. THOLD represents the decision threshold.

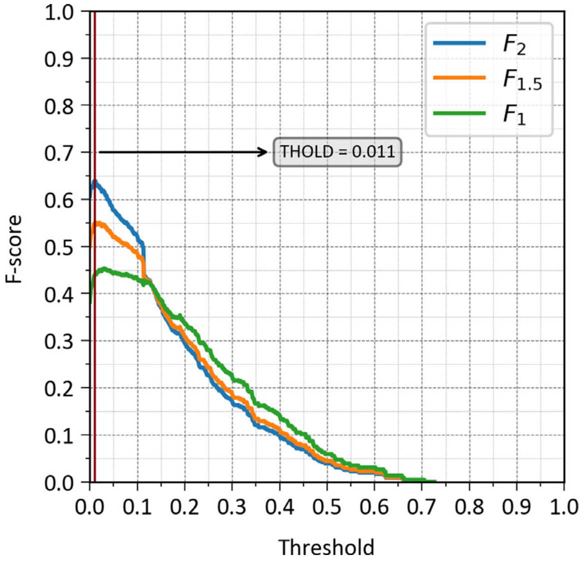

Fig. 10. F_1, $F_{1.5}$, and F_2 curves obtained by the Deep model for the label 1 – 10 (NPK) at 2000 integration steps. $F_{1.5}$, and F_2 show a global maximum for Threshold = 0.011. F_1 has a maximum at Threshold = 0.031.

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.