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

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, 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

[*] 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

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 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).

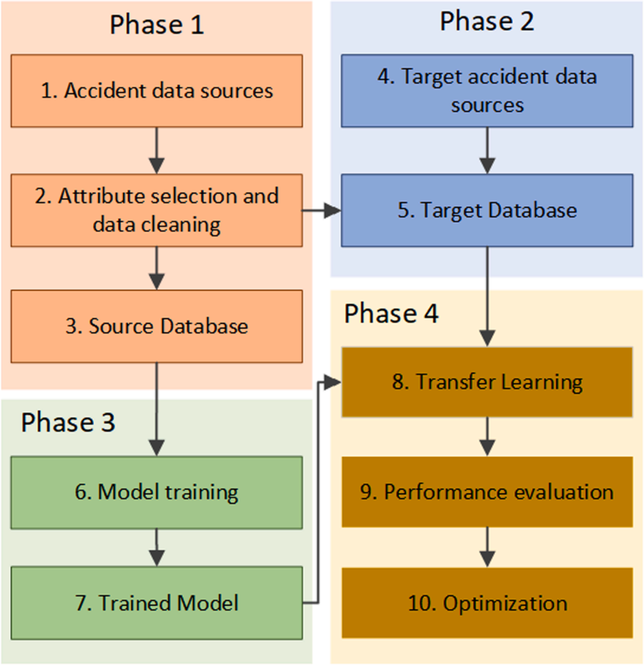

Fig. 1. Methodology workflow.

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).

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

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.

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).

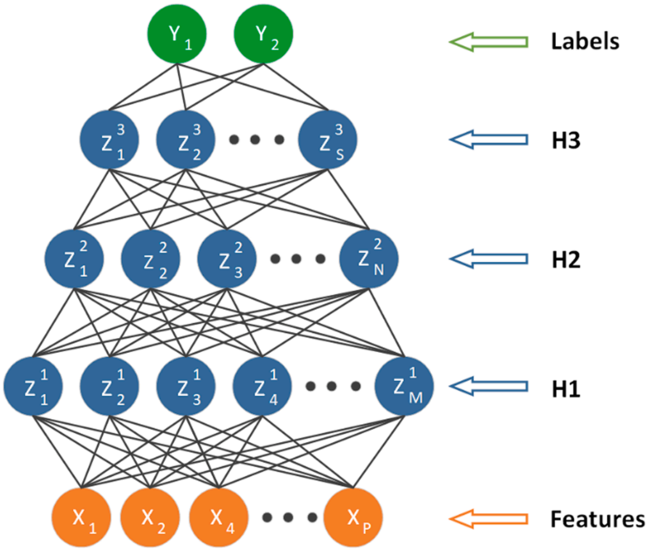

Fig. 2. Schematic representation of a DNN with three hidden layers (H1, H2, H3), P input features (X_i), and two output labels (Y_1 and Y_2).

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):

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

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

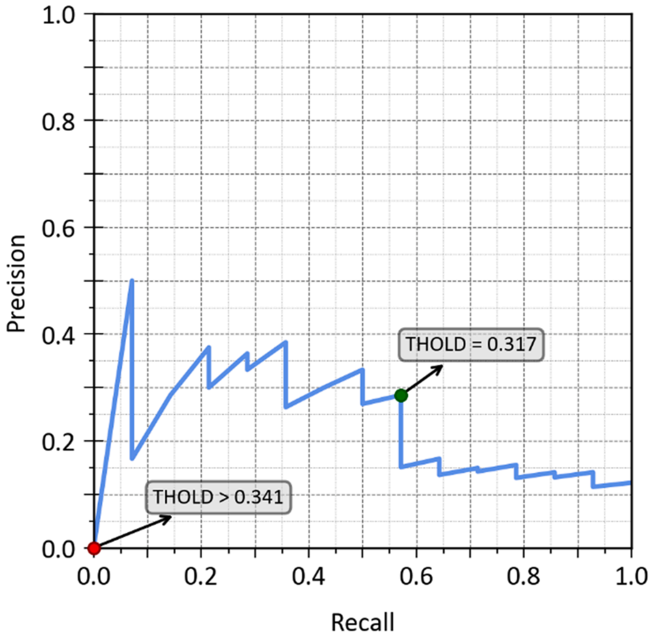

Fig. 3. Precision-Recall curve of the Deep model for the label 1 – 10 (NPK) at 2000 integration steps. THOLD represents the decision threshold.

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

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

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

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 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

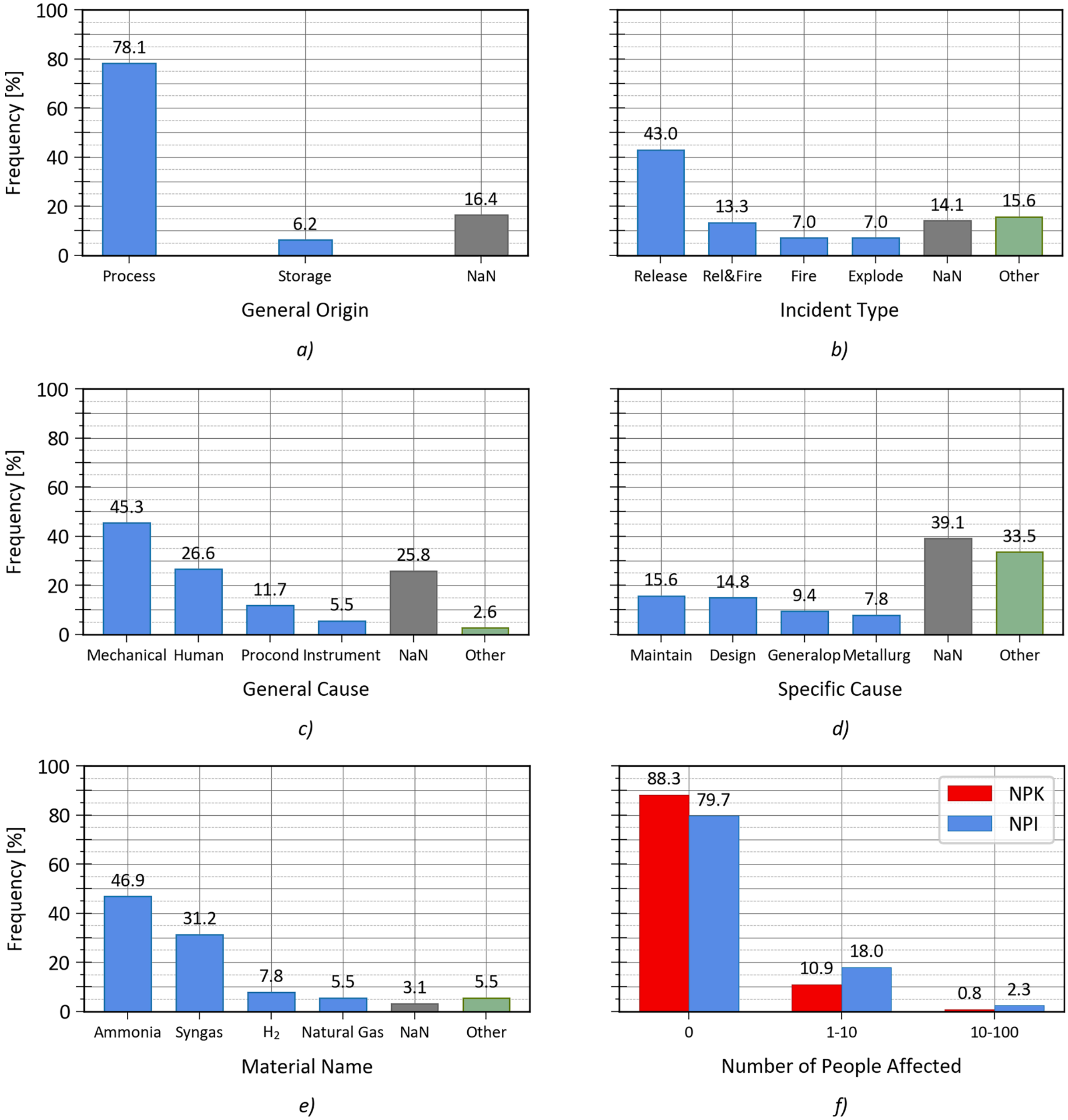

Fig. 4. Frequency distribution of the attributes GOG (a), IT (b), GC (c), SC (d), MN (e), NPK and NPI (f) (see Table 1 for the description of the attributes). Attribute codes are represented on the x-axis. "NaN" refers to missing values, "Other" refers to attribute codes that have not been represented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

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 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

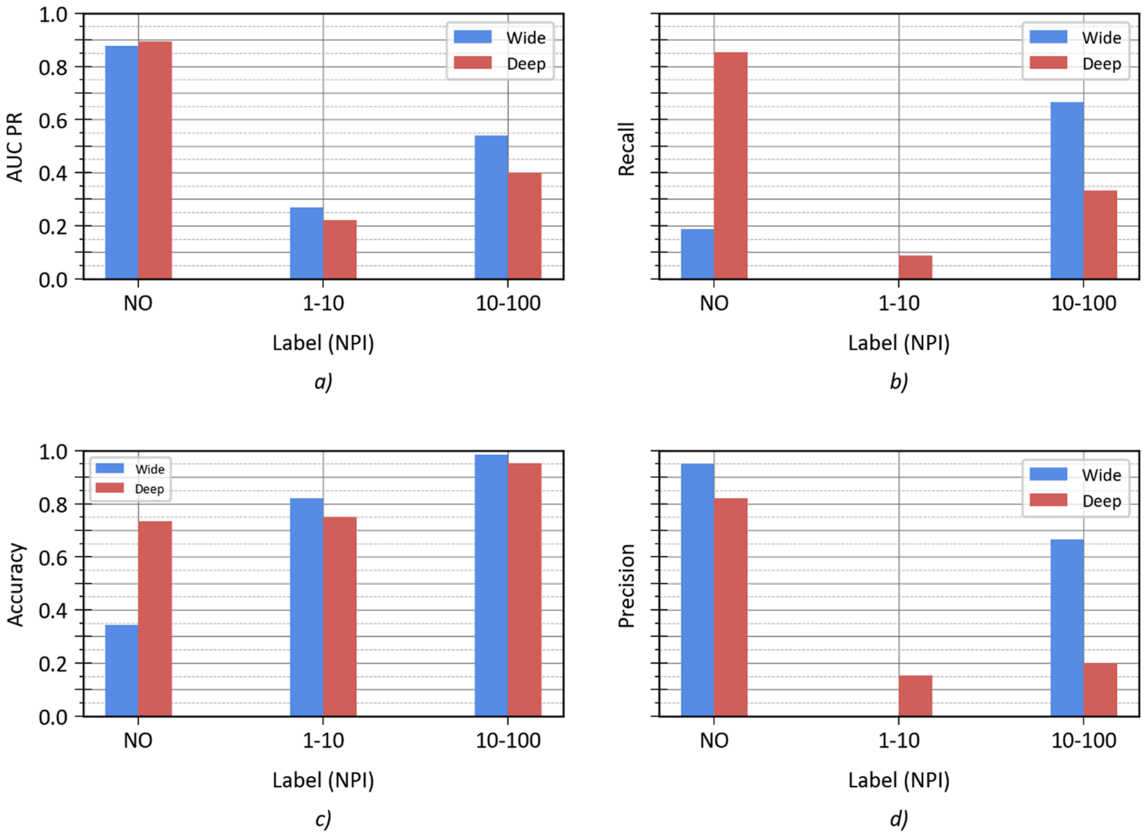

Fig. 5. Area Under the Curve Precision-Recall (a), Recall (b), Accuracy (c), and Precision (d) obtained from a small selection of simulations for the label category "Number of People Injured" (NPI). Labels are represented on the x-axis. Recall, Precision, and Accuracy are obtained at threshold = 0.5.

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.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.

The metrics in Fig. 9 indicate that 8 out of 14 events that caused 1–10

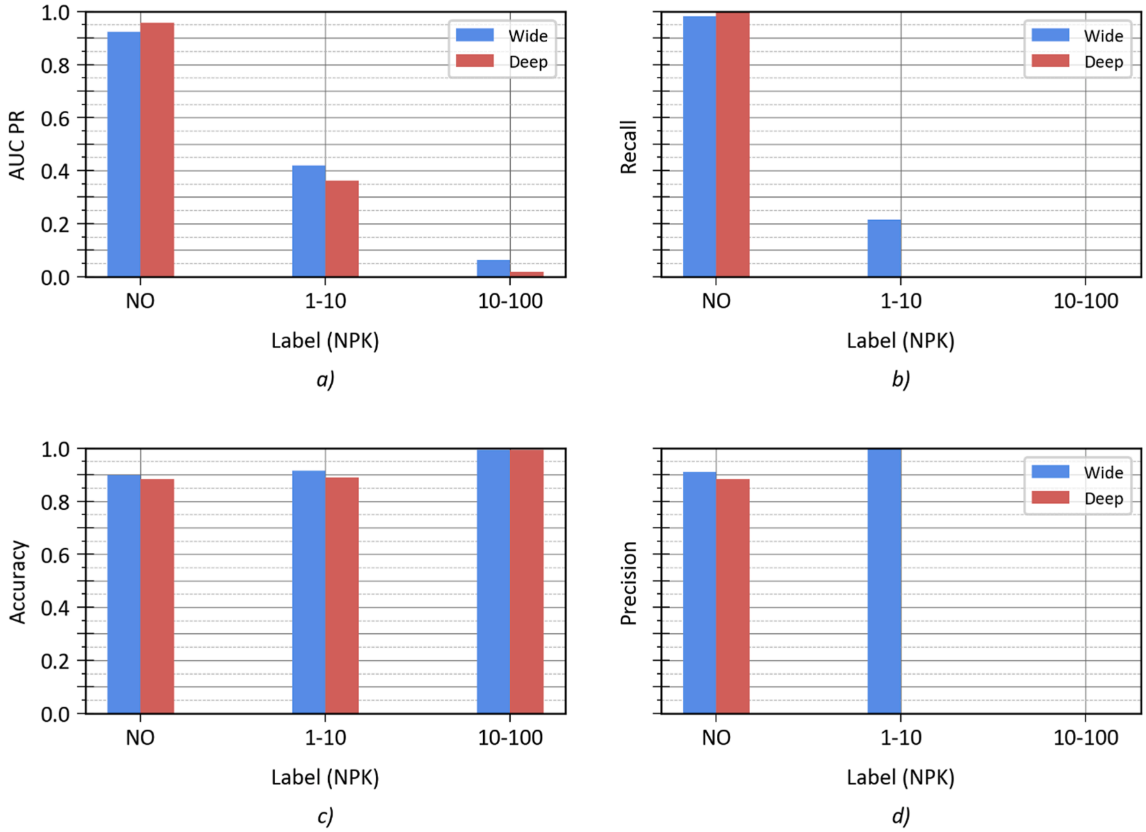

Fig. 6. Area Under the Curve Precision-Recall (a), Recall (b), Accuracy (c), and Precision (d) obtained from a small selection of simulations for the label category "Number of People Killed" (NPK). Labels are represented on the x-axis. Recall, Precision, and Accuracy are obtained at threshold = 0.5.

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

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:

- 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.

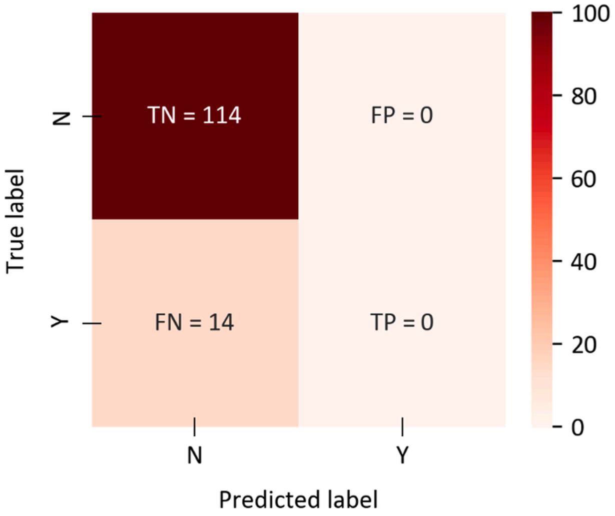

Fig. 7. Confusion Matrix produced by the Deep model for the label 1 – 10 (NPK) at 2000 integration steps. From top-left clockwise: True Negative (TN), False Positive (FP), True Positive (TP), and False Negative (FN) are obtained using a probability threshold equal to 0.5 and color-coded according to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

Hence, reducing the threshold to 0.317 has undeniably improved the

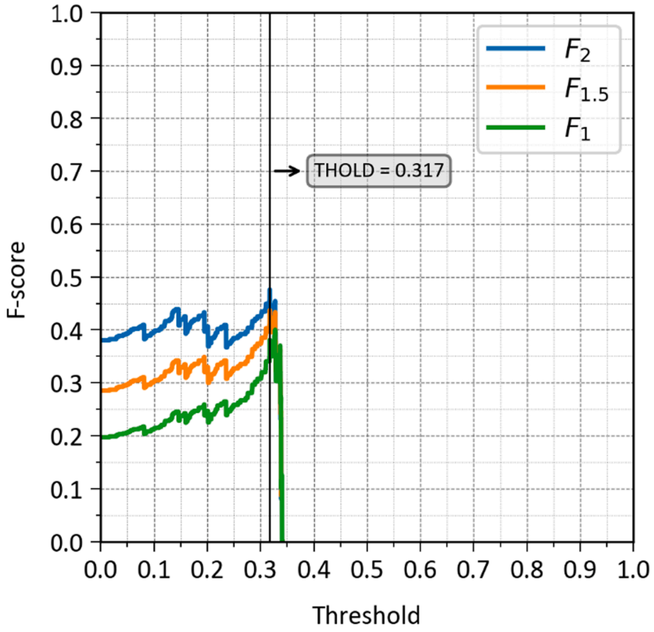

Fig. 8. F_1, $F_{1.5}$, and F_2 curves obtained by the Deep model for the label 1 – 10 (NPK) at 2000 integration steps. $F_{1.5}$, and F_2 show a global maximum for Threshold = 0.317. F_1 has a maximum at Threshold = 0.3173.

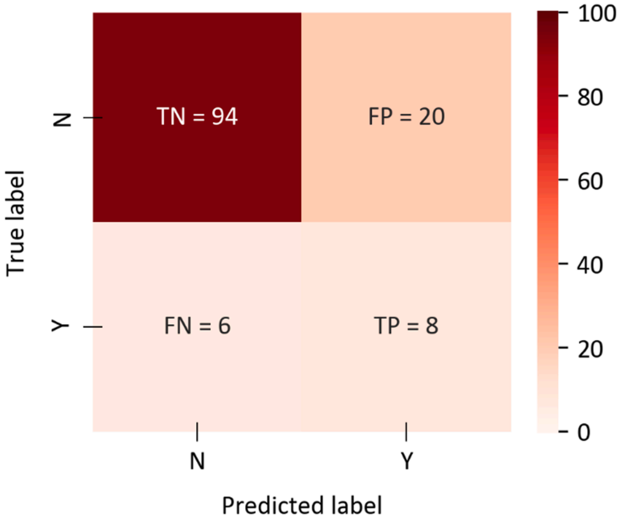

Fig. 9. Confusion Matrix produced by the Deep model for the label 1 – 10 (NPK) at 2000 integration steps. A decision threshold equal to 0.317 is used. From top-left clockwise: True Negative (TN), False Positive (FP), True Positive (TP), and False Negative (FN) are color-coded according to the color bar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

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.