Journal of Intelligent & Robotic Systems (2023) 108:17
https://doi.org/10.1007/s10846-023-01887-2

SHORT PAPER

A Novel Warning Identification Framework for Risk-Informed Anomaly Detection

Rialda Spahic[1,2] · Vidar Hepsø[2,3] · Mary Ann Lundteigen[1,2]

Received: 3 November 2022 / Accepted: 16 May 2023 / Published online: 1 June 2023

Abstract
Cyber-physical systems are taking on a permanent role in the industry, such as in oil and gas or mining. These systems are expected to perform increasingly autonomous tasks in complex settings removing human operators from remote and potentially hazardous environments. High autonomy necessitates a more extensive use of artificial intelligence methods, such as anomaly detection, to identify unusual occurrences in the monitored environment. The absence of data characterizing potentially hazardous events leads to disruptive noise displayed as false alarms, a common anomaly detection issue for hazard identification applications. Contrastingly, disregarding the false alarms can result in the opposite effect, causing loss of early indications of hazardous occurrences. Existing research introduces simulating and extrapolating less represented data to expand the information on hazards and semi-supervise the methods or by introducing thresholds and rule-based methods to balance noise and meaningful information, necessitating intensive computing resources. This research proposes a novel Warning Identification Framework that evaluates risk analysis objectives and applies them to discern between true and false warnings identified by anomaly detection. We demonstrate the results by analyzing three seismic hazard assessment methods for identifying seismic tremors and comparing the outcomes to anomalies found using the unsupervised anomaly detection method. The demonstrated approach shows great potential in enhancing the reliability and transparency of anomaly detection outcomes and, thus, supporting the operational decision-making process of a cyber-physical system.

Keywords Anomaly detection · Risk assessment · Risk analysis · Sensor systems · Autonomous systems · Imbalanced data

1 Introduction

The environment's safety is ever more reliant on cyber-physical systems that have applications in, among others, intelligent drones, remote sensing, and smart sensor systems. These systems are taking on permanent roles in various industries such as oil and gas, energy, and mining. They are replacing various human operations and carrying out critical responsibilities, including inspecting and monitoring remote, possibly hazardous environments. The increasing growth of sensor-collected data grows a need for artificial intelligence (AI) and data-oriented technologies along with the requirements for more autonomous systems that are safer, more perceptive, and more financially viable.

Autonomy is described as the capability of a system to operate independently from external factors [1]. Increased autonomy necessitates a more significant usage of AI [2] methods that copy intelligent human behavior [3]. With various sensors, the cyber-physical systems can efficiently gather data during ongoing operations and use AI methods to analyze the data in real time and gain situational awareness. Consequentially, increased autonomy has the potential to replace constant human supervision. As a form of AI, machine learning (ML) uses high volumes of data to learn how to execute tasks rather than being programmed to do them, allowing computing systems to become more intelligent as they encounter additional data [3]. Similarly,

✉ Rialda Spahic
rialda.spahic@ntnu.no

Vidar Hepsø
vidar.hepso@ntnu.no

Mary Ann Lundteigen
mary.a.lundteigen@ntnu.no

[1] Department of Engineering Cybernetics, Trondheim, Norway

[2] Norwegian University of Science and Technology, Høgskoleringen 1, 7491 Trondheim, Norway

[3] Department of Geoscience and Petroleum, Trondheim, Norway

anomaly detection, as a data-oriented method, detects unusual trends in data that can give insight into potentially hazardous occurrences. Detecting critical trends in good time allows for the opportunity to take necessary corrective actions in advance to ensure safe operations. Considering the variety of hazards that can affect these systems, many techniques might increase their ability to operate safely under all conditions. Therefore, AI technology must be reliable in order to responsibly integrate it into existing systems and operations.

The challenges inherent in unsupervised anomaly detection emphasize the necessity for further research into semi-supervised or alternative methodologies [4]. Although sensor data and data-driven methods are becoming essential in many safety–critical or high-risk engineering systems, data-driven methods may not be sufficient to ensure safety because they lack the underlying causal knowledge [5]. Additionally, benchmarking and comparing anomaly detection methods is eminently challenging. Due to these challenges, early warning indicator of potential hazardous events may be missed, possibly placing assets or the environment in jeopardy during operations [6].

In cyber-physical systems, particularly autonomous systems, that form decisions based on data-oriented methods, the safety and responsibility of the methods and the data that trains the methods cannot be overemphasized. Therefore, this paper proposes the development and evaluation of a Warning Identification Framework (WIF) as an extension of previous work [7, 8]. The purpose of the WIF is to facilitate the decision-making of a cyber-physical system that uses anomaly detection methods to identify warning signs of an ongoing operation. Such applications include autonomous underwater drones for inspecting pipelines and observing potential surface corrosion or cracking or intelligent sensor systems for monitoring drilling operations in mines and listening for potential seismic tremors, shaking of the ground under the stress of mining or drilling. To facilitate the decision-making of a cyber-physical system, another objective of WIF is to address the interrelated challenges of unlabeled, contextless, biased data, unsupervised methods, and consequentially unreliable anomaly detection results. WIF is anchored in risk analysis and comprises three main steps: characterization, analysis, and ranking of warning impacts detected through anomaly detection. To compare the standard hazard assessment and anomaly detection methods, we examine unlabeled seismic data with varied sensor values for tracking seismic tremors and three distinct hazard assessment methods for identifying low, medium, and high-impact hazardous occurrences.

The following is a summary of the primary contribution of this paper,

Warning Identification Framework:

1. Novel risk assessment perspective on seismic hazard identification's training and assessment role in unsupervised anomaly detection approach.
2. Identification of overlapping methods and roles in risk assessment and anomaly detection.
3. Preliminary results of three seismic hazard identification methods and their assessment role for unsupervised anomaly detection results.

This paper's structure is as follows. Section 2, Background, introduces the term anomalies, their taxonomy, and the structure of anomaly detection methods. This section also introduces the concept of risk, the phases of risk analysis, and their relationship to anomaly detection. Section 3, Challenges, discusses the context and data imbalance contributing to an ever-increasing trust mismatch inside AI and ML-based systems. Section 4, Existing Approaches to the Challenges, addresses three of the most current approaches to the challenges previously discussed: simulations for data extrapolation, rule-based anomaly detection and classification, and decision boundaries and thresholds for data and procedures. Section 5, Warning Identification Framework, introduces the framework's concept, steps, and goals. Section 6, Case Study, analyzes seismic data for tracking seismic tremors, compares three hazard assessment methods to anomaly detection results, uses WIF to comprehend the differences between the methods and discusses the results. Section 7 discusses the results and conclusions from the case study, while Sect. 8 summarizes and concludes the paper. Finally, Sect. 9 highlights future research.

2 Background

2.1 Anomalies and Anomaly Detection

Anomalies (in literature often interchangably referred to as outliers, novelties, abnormalities, discordants or deviants) are occurrences in a dataset that are odd in some sense and do not fit the dataset's general or expected trend [9]. They apply to a wide range of desired and undesired phenomena, appearing as static occurrences, time-related events, single and grouped occurrences. Despite being interchangebly used, the terms anomaly and outlier are distinguished in some studies [10–12]. For an example, Hawkins [11] provides a definition of an outlier: "An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism." In the recent study on anomaly classification, Foorthius [13] describes that the definition of anomaly is vague and dependent on the application domain due to the wide variety of ways anomalies manifest themselves. In order to understand how unsupervised anomaly detection methods in relationship with knowledge from risk analysis can be utilized to improve true anomaly discovery and potentially avoiding missed out early warning signals, it is important to understand how anomalies manifest themselves.

A broadly accepted classification of anomalies, described by Chandola et al. [14], differentiates three general categories:

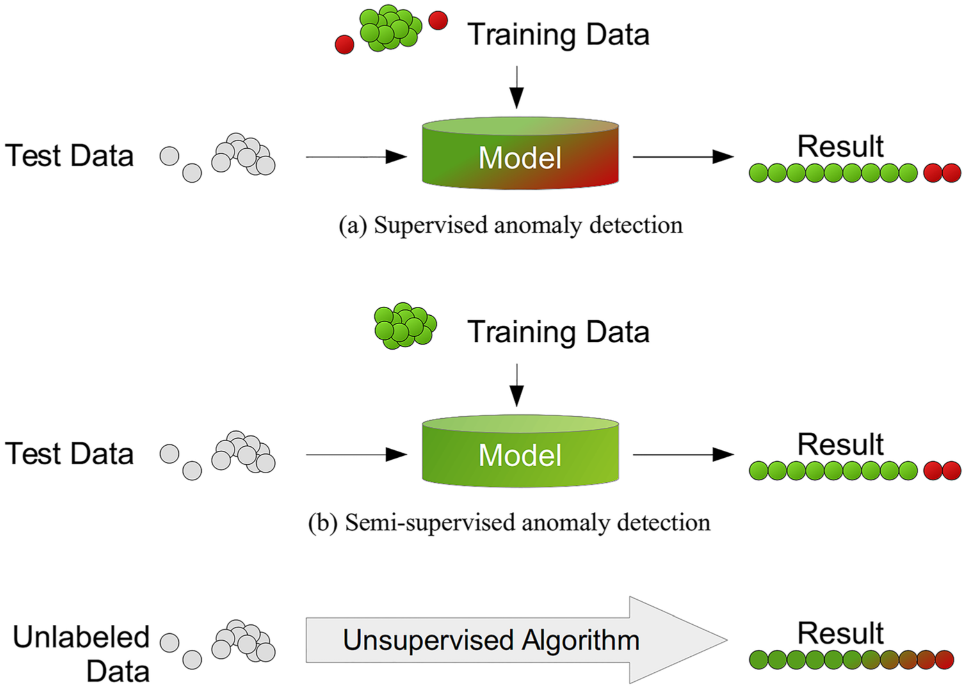
1. Global anomaly (also described as point and content anomaly [15, 16]) one or several individual data points that are deviant with respect to the rest of the data.
2. Contextual anomaly (also described as conditional anomaly [17]) data points that are deviant when an explicitly selected context is taken into account.
3. Collective anomaly (also described as group or aggregate anomaly) a collection of data points that belong together, as a group deviate from the rest of the data.

The methods that detect anomalies highly depend on the available labels in the dataset, i.e., if the data is identified with certain characteristics and classified. As illustrated on Fig. 1, Goldstein et al.[18]. describe three main setups of anomaly detection:

1. *Supervised* anomaly detection refers to a situation in which the training and test sets are fully labeled. A conventional classifier can be first taught and then implemented. This scenario is comparable to traditional pattern recognition, except classes are frequently imbalanced. Due to the assumption that anomalies are recognized and adequately labeled, however, this setup is practically irrelevant. For many applications, anomalies are either unknown beforehand or may emerge spontaneously during testing.
2. *Semi-supervised* anomaly detection employs training and test datasets, with training datasets containing only normal, anomaly-free data. Principally, anomalies are identified by deviating from a model of the normal class. A dataset comprising normal, or non-anomalous, data can be acquired either through manual curation or through frequency analysis methods, in which the most often gathered data is deemed normal due to the rarity of hazardous events [6].
3. *Unsupervised* anomaly detection is the most flexible technique because it does not require labels. In addition, no distinction is made between a training dataset and a test dataset. An unsupervised anomaly detection approach evaluates data only based on the inherent properties of the dataset. Typically, distances or densities distinguish between normal and abnormal behavior.

2.2 Risk and Risk Analysis

Risk is defined as the effect of uncertainty on objectives, where the effect can be positive, negative, or both, resulting in opportunities and threats [19]. Typically, the risk is expressed in terms of risk sources, future occurrences, their effects, and the probability they will occur. Earlier guidelines



(a) Supervised anomaly detection

(b) Semi-supervised anomaly detection

(c) Unsupervised anomaly detection

Fig. 1 Supervised, semi-supervised and unsupervised anomaly detection [18]

for the inclusion of safety aspects in standards [20] define risk as a combination of the probability of occurrence of harm and the severity of that harm, where harm is an injury or damage to people's health, property, or environment [20].

In his book, Risk Assessment Theory, Methods, and Applications, Rausand [21] describes risk analysis as one of the three main elements of risk management (see Fig. 2), the continuous process to reveal, analyze, and assess potential hazardous events in a system, and identify and introduce efficient risk control measures to eliminate or reduce possible harm [21]. The risk analysis is responsible for:

- the identification of hazards and threats related to the system of interest;
- the identification of potential hazardous events that may occur;
- the identification of causes of hazardous events;
- the identification of barriers and safeguards to prevent or reduce the hazardous events and assessment of their reliability;
- the identification of accident scenarios related to each hazardous event and their consequences.

The other two main elements of risk management are [21]:

1. Risk evaluation for assessing risk picture, comparison of the risk with established risk acceptance criteria, considerations of alternative systems.
2. Risk control and risk reduction for making decisions regarding introducing new risk-reducing measures, implementing the measures, monitoring, and communicating the risk.

Risk analysis systematically uses available information to identify hazards and estimate risk where the hazard is a potential source of harm [20]. Therefore, risk analysis can be observed as a tool to inform decision-making concerning future welfare since the risk is always related to what can happen in the future [21]. As illustrated in Fig. 2, the analysis of risk is carried out to answer the following questions:

- Hazard identification: *What can go wrong* ?
- Frequency analysis: *What is the likelihood of that happening* ?
- Consequence analysis: *What are the consequences*?

The risk analysis results have great potentials for assessing or improving the data for anomaly detection models. These potentials may be observed using the risk analysis for rule-based labeling in classification problems or transferring its knowledge to train the new models. As described by [22], transfer learning comprises different techniques that aim to gather the knowledge gained at the source problem to develop a new model using the gathered knowledge, thus minimizing the efforts of developing that new model. To the best of our knowledge, these potentials have not been leveraged effectively in prior studies (see Sect. 4) that would tackle the existing challenges in the anomaly detection methods (see Sect. 3).

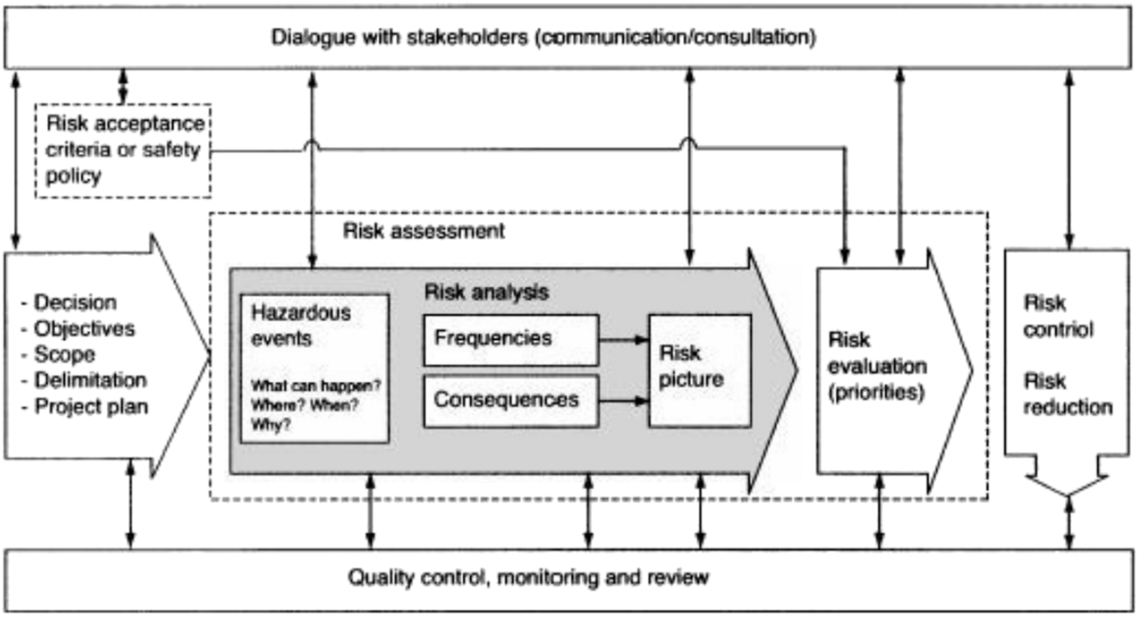

Dialogue with stakeholders (communication/consultation)
Risk acceptance criteria or safety policy
Risk assessment
- Decision
- Objectives
- Scope
- Delimitation
- Project plan
Risk analysis
Hazardous events
What can happen? Where? When? Why?
Frequencies
Consequences
Risk picture
Risk evaluation (priorities)
Risk control
Risk reduction
Quality control, monitoring and review

Fig. 2 Elements of Risk Management, adapted from [21]

2.2.1 Hazard Identification

Identifying the hazard is a critical first step toward preventing or mitigating it. Certain hazards require a triggering event to grow into a hazardous event, whilst others may develop into a hazardous event gradually [21]. A triggering event is an event or situation that must occur in order for a hazard to cause an accident [21]. Hazard identification techniques determine [23]:

- Possible cause of the harm;
- How the harm will manifest itself;
- What measures are in place to avoid or mitigate harm;
- The extent to which the harm is tolerable;
- What further actions or resources are required to avoid or mitigate harm.

The What-If Checklist, the Hazard and Operability Study, and the Failure Modes and Effects Analysis are three of the most often used techniques for hazard detection [23]. Knowing what can go wrong and identifying the properties of hazards is a crucial step in labeling the training sets for supervised classification or anomaly detection.

2.2.2 Consequence Analysis

A consequence is an adverse event that may occur due to a hazard [23]. As a result, consequence analysis examines the predicted impacts of incident outcome situations regardless of their frequency or likelihood. There is a specific amount of energy or material released in the event of containment failure. This is referred to as the source term [23]. Assume the effects are instantaneous, as with an explosion. In that situation, the analysis uses inputs such as the material type, the release pressure, and other factors to determine the impact effects. If the effects are delayed, the source term characteristics are used as inputs in a dispersion analysis followed by an analysis of the impact effects. Anomaly detection can detect anomalies representing significant information about the ongoing operation or anomalies that do not require any insight or resource allocation. Consequence analysis provides critical information on the impact of hazards or anomalies that can aid operators in allocating necessary resources.

2.2.3 Likelihood Analysis

Risk cannot be accurately assessed without first analyzing the likelihood of an event occurring, which can be challenging. Analyzing the likelihood becomes progressively more challenging for complex systems, and hazard scenarios [23]. The likelihood of often occurring events may be evaluated and validated using statistical analysis that requires large amounts of data. The common methods for likelihood analysis are fault propagation modeling methods event tree analysis and fault tree analysis. The situations, conditions, and protective mechanisms, together referred to as intermediate events that should have prevented the accident, are listed, along with their associated probability of occurrence. In anomaly detection, the likelihood and frequency analysis bring invaluable information on the underlying knowledge of detected anomaly or hazard. Although not all detected anomalies require reaction response or allocation of resources, knowing the likelihood or frequency of certain undesired events is smaller or larger under a particular operational context may eliminate the need for conjecture when classifying or labelling observed anomalies.

2.2.4 Warning Management

While warning management is not explicitly included in risk analysis, it is necessary to employ risk analysis insights as a layer of protection. A warning is used to notify the operator of a malfunctioning piece of equipment, a process deviation, or an unexpected state that demands operator intervention [23]. Alarms assist the process in remaining within normal operating parameters and ensuring its safety, differentiating between negligible, tolerable, and unacceptable risks. A risk level that is considered acceptable suggests that the risk level is usually recognized as insignificant [21]. Typically, additional risk-reduction measures are not necessary. Tolerating a risk, or tolerable risk, implies that we do not perceive it as negligible or something to be overlooked, but rather as something to be monitored and mitigated further as and when possible [21]. Except in exceptional circumstances, activities with an unacceptable level of risk are considered unsuitable, regardless of their advantages. Activities that create such risk would be prohibited, or the risk would have to be mitigated at all costs [21]. To assist in determining which alarms should be addressed first, each warning is assigned a priority, often based on the severity of the potential consequences.

Figure 3 depicts the operator response to warning. The operator must be capable of promptly detecting, diagnosing, and appropriately responding to the warning to avoid a hazardous event. A warning management system is a crucial component of cyber-physical systems involved in safety–critical activities. In increasingly complex systems, an autonomous system, such as a UAS, is anticipated to conduct detection, diagnostics, response, and reaction depending on the scenario. UAS, such as underwater drone or smart sensor systems, offer warnings to the operators if the ongoing activity requires further attention. In this instance, the UAS can autonomously determine if a given monitored occurrence is an early warning indicator and whether to

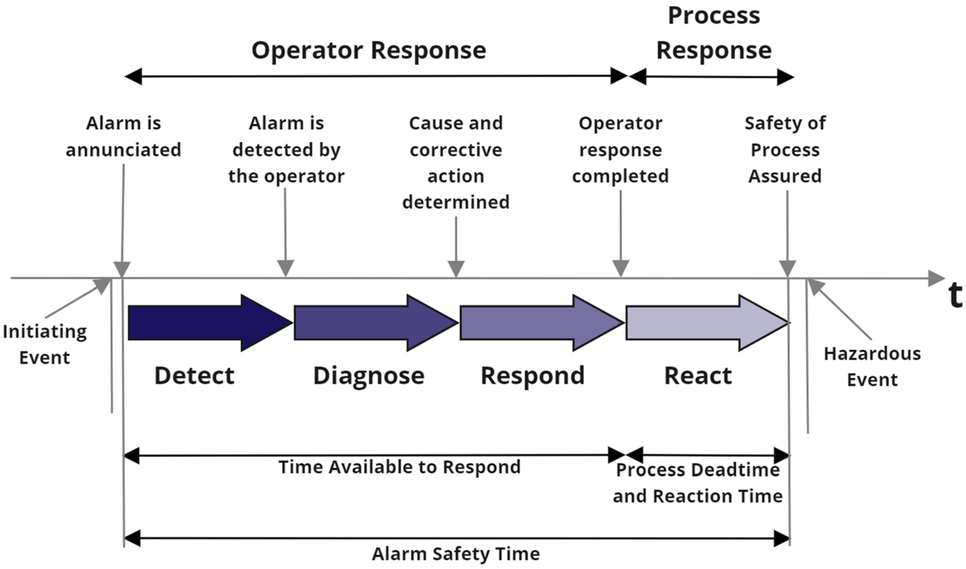

Operator Response
Process Response
Alarm is annunciated
Alarm is detected by the operator
Cause and corrective action determined
Operator response completed
Safety of Process Assured
t
Initiating Event
Detect
Diagnose
Respond
React
Hazardous Event
Time Available to Respond
Process Deadtime and Reaction Time
Alarm Safety Time

Fig. 3 Operator and Process Reaction Time, adapted from [23]

sound a warning using data-driven approaches, particularly AI.

3 Challenges

3.1 Missing Context and Data Imbalance

We observe a growing interest in research within the context of ML strategies for knowledge sharing and organizing, such as [24–26]. Righteously so, we witness a more permanent role of autonomous systems and ML in the industry. However, integrating ML into existing systems involves heterogeneous teams of, amongst others, software engineers, data analysts, and domain specialists. Every domain specialist and analyst in a heterogeneous team developing a software system that integrates ML should comprehend the context underlying ML methods and data. This context specifies the relationship between code and data, as well as the relationship between data and intent of the operation. Lacher et al. [27] point out that the context is critical to a system's capacity to respond satisfactorily as it becomes increasingly autonomous. In a *framework for discussing trust in increasingly autonomous systems* by [27], context is represented as a binding point between people, environment and the machine (i.e., the autonomous drones) (as seen in Fig. 4). *People* have varying perspectives of the machine which is based on their roles and greatly influenced by their culture (such as age and professional affiliation). The operation's context is established by the *environmental factors* because the machine will identify the situations based on the data received from sensor inputs. *The machine* is designed to perform required tasks at a high level of performance, which can be observed, measured or assessed. The results of these assessments will have an impact on people's confidence in a machine's competence. If the machine produces

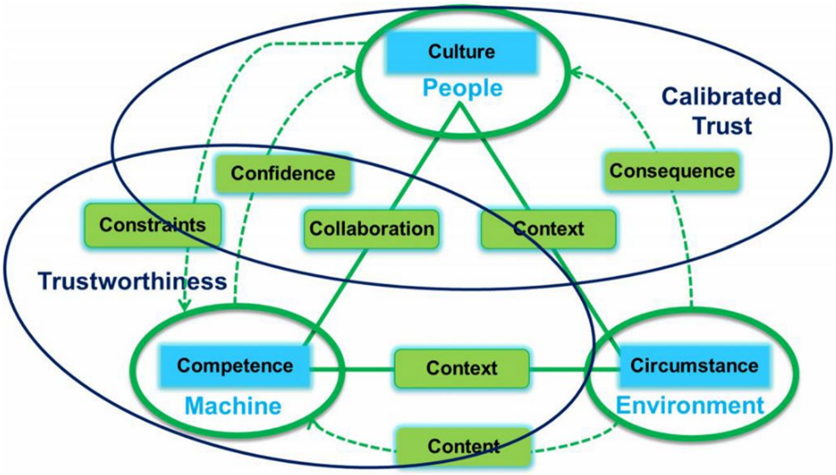


Fig. 4 A Framework for Discussing Trust in Increasingly Autonomous Systems [27]

expected results, the human confidence in the machine will increase, answering the question of the machine's reliability. Lacher et al. [27]. conclude that most of the machines will have a degree of human collaboration and the degree of trustworthiness between people and machines is a cultural, organizational and sociological challenge. According to [27], calibrated trust is founded in our perception and expectation of system performance, which has become an engineering, social, cultural, and organizational challenge. Yet, as machines become increasingly complex, trustworthiness becomes more challenging to maintain due to the difficulty to understand the functioning and set the expectations on the machine performance.

Hayes et al.[16]. offers an example of an anomaly detection algorithm missing context in the circumstance of a sensor reading detecting that a particular electrical box consumes an abnormally high quantity of energy. However, when examined in the context of the sensor's location, present weather conditions, and time of year, it is well within normal boundaries. There are various explanations for these shortcomings, which we loosely divide into two categories: technical and people/process-related. The technical reasons as the often unpredictable malfunctions of the system. However, the people/process-driven reasons for ML shortcomings are due to the more complex methodologies that the individuals or teams use to organize and transfer knowledge, including designing, developing, and maintaining the systems that employ ML. Lee et al. [28] argue that the shortcomings, particularly due to biases caused by imbalances in data, can be removed not by niche methods but rather by informing the appropriate mitigation strategy, whether technical or people/process-driven. Nevertheless, the previous studies inform that practitioners struggle to integrate newly proposed tools and methods into existing processes [28]. Authors [28] suggest that identification and categorization of different types of shortcomings, such as biases, can help to understand the roots of the unintended ML outcomes.

Due to the wide range of anomalies that can disrupt operations and the large amount of data produced by environmental sensors, real-time anomaly detection is becoming more challenging. Imbalanced or underrepresented data, such as high consequence and low probability hazardous event data, is particularly problematic because the data processing methods form biases in favor of more represented data. Classification methods, entrusted with effectively predicting outcomes from the sensor data, tend to reproduce these biases [29]. Furthermore, underrepresented data can be disregarded as noise due to the anomaly methods' inclination toward efficiency and sacrificing anomalies as tolerable collateral damage [30]. False alarms, or noise, are another known drawback of anomaly detection [31]. False alarms fall into two categories: false positives and false negatives [32]. When a normal or non-hazardous event is recorded as a hazardous event, this is called a false positive. A consequence of false positives is that a potentially hazardous event may go undiscovered due to prior false positives. A false negative is defined as the inability to notice a hazardous event. Due to the high proportion of false alarms created by anomaly detection, it is difficult to correlate specific alarms with the events that triggered them [32]. Additionally, current methods for anomaly detection focus primarily on data content, with no regard for the context behind the data [16]. These methods yield conclusions that are based on correlation without causation. *Causation* is the situation in which one event, *a cause*, causes another event to happen *an effect*. *A correlation* is the situation in which two or more events appear to be related. Therefore, basing conclusions solely on correlations is one of the critical problems in data analysis [5], as it might result in misleading predictions. However, many datasets lack labels or supervision that provides additional information and context about the data [33] making the training and testing of anomaly detection methods even more challenging.

3.2 Trust Imbalance

Many judgments made by cyber-physical systems in various scenarios are based on its analysis of the environment [34]. The biased and unjust consequences of data-driven methods are frequently the result of opaque or black-box methods that lack transparency. As a result, anomaly detection methods have recently piqued the interest of industry and academics in the hopes of gaining greater transparency and offering more context to the data and the anomaly detection methods [13]. The three of the biggest challenges of evaluating these systems are user acceptance and trust, adequate evaluation, and defining autonomy comprehensively and quantitatively [35]. Autonomous drones, for example, must operate safely and be resilient in changing environments and complex scenarios. The ability to successfully manage disturbances and emergent needs during the system's mission resilience determines the efficacy and reliability of autonomous systems [36]. A resilient and reliable system can alter its functioning in advance of or in response to changes and disturbances, allowing it to continue working even after a severe incident or in the face of persistent stress, primarily by being proactive on safety [37]. Hollnagel has outlined the three fundamental functions of a resilient system [1]:

1. *Anticipate disturbances*, prospective threats (Hollnagel uses the terms *threat* and *hazard* synonymously), and any other destabilizing conditions. This function enables the system to forecast the future and adjust risk tolerance.
2. *Monitor performance*, risks and threats while constantly improving its own risk identification model. This func-

tion enables the detection of nonpermanent transient impacts that, despite not being permanent, can still cause failures and accidents.

3. *Respond to threats*, whether they are regular, irregular, unexpected or unexampled. This function denotes a resilient system's preparedness, flexibility, and adaptability.

O'Neil [38] argues that data-driven methods should be prejudice-free, produce objective results, judge according to universal norms, and eliminate biases. However, since the methods are based on historical data, they not only incorporate biases, they reinforce them [38]. Since highly autonomous systems rely heavily on data-driven methods, these systems must include humancentered features to ensure that they society, industry, and the economy while adhering to ethical norms.

4 Existing Approaches to the Challenges

With supervised anomaly detection and labeled datasets, discriminating between anomalous and non-anomalous data is supposed to be straightforward. The dataset contains labels for anomalous and non-anomalous data points, enabling anomaly detection methods to classify the data more precisely. Distinctively, unlabeled data are analyzed using distances, density, and trends between data points. However, the difficulty of underrepresented data, or minor data (see Fig. 5), has recently eroded faith in supervised methods as well. The classifier or anomaly detection method is unable to distinguish between more represented or major and minor classes in the data, favoring the major data and thus overlooking the minor data, potentially omitting critical information.

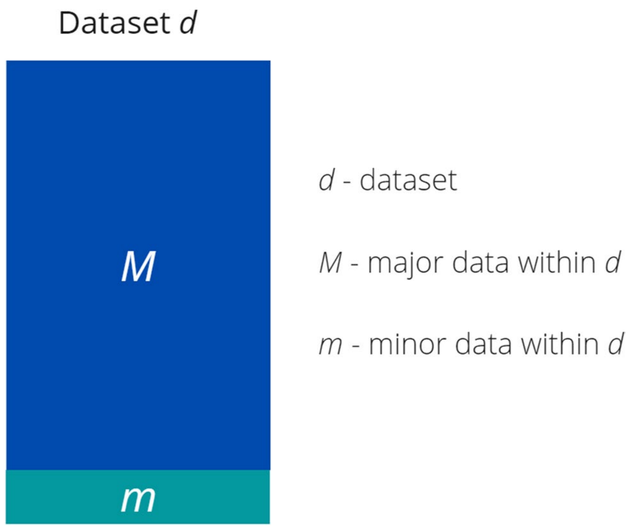


Fig. 5 Major and minor data in a dataset

Numerous approaches to the problem of underrepresented data have been developed, and the following paragraphs allude to recent studies while summarizing the approaches as follows:

- Extrapolating minor data through simulations, causal and physics data
- Setting decision boundaries and thresholds for normal data
- Semi-Supervised and Rule-Based Anomaly Detection and Classification

4.1 Extrapolating Minor Data and Simulations

Due to the rarity of hazardous occurrences with high consequences, there is a sparse indication of their properties in the sensor-collected data during environment or asset monitoring. This lack of data necessitates using simulations to replicate the natural world and generate artificial hazards and hazardous events, further extrapolating imbalanced datasets with the artificially generated data from simulations so that machine learning models can train on less imbalanced data. Eldevik et al. [5] highlight in their work on AI and safety that data-driven models alone are insufficient. Although sensor data and data-driven models are becoming an integral part of a growing number of safety–critical and high-risk engineering systems, the high-consequence and low-probability scenarios are not well reflected by data-driven models. The authors [5] propose the use of causal and physics-based knowledge for extrapolation robustness. The data-generating processes consist of stochastic and deterministic elements, providing an opportunity to utilize the deterministic processes, or those governed by known principles, and extrapolate the naturally underrepresented data with the existing causal and physics-based knowledge. The authors [5] argue that the combination of data-driven models and the causal knowledge of industry experts is essential for decision-making processes within AI systems.

The method of simulating, or extrapolating with causal and physics-based knowledge, has significant drawbacks, including runtime and curse-of- dimensionality [5]. For a high-consequence system, a model used to inform risk-based decisions must predict potentially catastrophic scenarios prior to the occurrence of the scenario. However, the runtime of these complex models is often significant, commonly taking up to several days, making it nearly impossible to initiate necessary analysis in real or near real-time. Alternately, the models can be run in advance, but this again necessitates sophisticated processes with many inputs, restricting the possibility of simulating every possible condition that an actual system can encounter before its operation. Eldevik et al. [5] emphasize that data-driven models should incorporate risk assessment into the decision-making processes as we rapidly progress toward more autonomous systems that employ AI for making safety–critical decisions.

In addition to computationally expensive simulations, Zhang et al. [39] argue that simulation experiments can be expensive to conduct in laboratories and frequently meet physical limitations for simulating the real world (i.e., simulating scenarios in the ocean vs. in a laboratory water pool). The primary limitation of simulations, whether virtual or in laboratories, is their inability to reliably mimic the complex interactions between the environment, the asset, and the ongoing hazard in the case of a hazardous occurrence.

4.2 Decision Boundaries and Thresholds

Anomalies are characterized in terms of previous behavior. This suggests that a novel behavior may first appear anomalous but ceases to be anomalous if it persists, establishing a new normal pattern [40]. Lavin et al. [40] define anomaly windows to aid in early detection. Each window is a collection of data points centered on a ground truth label for an anomaly. The earlier a detector can reliably identify anomalies, the better, which implies that these windows should be as large as possible. The trade-off with exceedingly large windows is that unreliable or random detections would be reinforced regularly [40]. This technique allows for a large window of opportunity for early detection and allows for partial credit for detections made shortly after the ground anomaly. The authors [40] emphasize that various applications may place a greater emphasis on true positives than on false negatives and false positives. For instance, in a manufacturing plant, a false negative may result in machine failure and costly production disruptions. Similarly, a false positive may necessitate an in-depth examination of the data by a technician.

Li et al.[41]. developed a novel data-driven approach to anomaly detection in cyber-physical systems by establishing a decision boundary to classify new observations using a geometric structure non-convex hull. Convex hullbased methods define a closed boundary around the normal data points. These methods make no assumptions about the underlying distribution. The convex hull-based methods do not require extensive parameter tuning, making them useful for boundary-based anomaly detection [41]. Since not all potential anomalies are known in advance, most data-driven anomaly detection techniques depend on developing a model of the system's normal behavior. This dependency may reduce the likelihood of noise or false alarms occurring during anomaly detection. The points within the convex hull are normal, whereas the points on its periphery are anomalous. However, convex hull-based algorithms produce many erroneous classifications when the input normal data is not convex [41]. The authors [41] demonstrated that incorporating a non-convex hull as a decision boundary for anomaly detection in data with non-convex forms achieved significant improvements over typical convex hull-based approaches. Shin et al.[42]. studied data bias caused by underrepresented classes in datasets. They advised using decision boundaries to increase the accuracy of anomaly detection generative adversarial network (AnoGAN) results produced from low-quality data. The primary challenge encountered by the authors [42] is *the subjective nature of establishing the decision boundary*. They evaluated the proposed method's success using the Area Under the Curve (AUC) and the F-measure through testing multiple arbitrary values for the decision boundary. AUC evaluates a classifier's ability to discriminate between classes. In contrast, F-measure evaluates the performance of a binary classification model based on predictions for the positive class. The proposed model presented in the [42] research has a slightly greater AUC and F-measure value (0.023 and 0.0231, respectively) than the initially tested AnoGAN result. While decision boundaries are frequently seen in classification and supervised algorithms that utilize labeled data, such as SVM [43], a similar approach can be applied to unlabeled data using semi-supervision.

The disadvantage of decision boundaries or thresholds is their construction. The boundaries are constructed either by an algorithm that learns from data patterns or by assuming a geometrical shape (i.e., convex hull-based methods [41]). Forming context or application-specific boundaries, as opposed to dataset-specific ones, is one approach to mitigate the disadvantages and establish more reliable decision boundaries.

4.3 Semi-Supervised and Rule-Based Anomaly Detection and Classification

Rule-based classification is a method for classifying or labeling data points using conditions such as 'if–then'. The benefit of rule-based classification resides in its interpretability and approach to generalization, rather than labeling each data entry individually. Nonetheless, this strategy requires manual inputs from domain experts and can soon become a complex task when applied to extensive data and unstructured sets.

Deng et al.[44]. explored a rule-based semi-supervised approach to anomaly detection due to a lack of labels in data and, consequentially, an emphasis on unsupervised methods that produce incomprehensible results. The authors [44] observed the challenge in selecting appropriate labels when training models for anomaly detection due to the vague definition of an anomaly being *a data point that does not share a similar pattern with the rest of the data population*. Their approach to applying rule-based classification in anomaly detection consisted of visually presenting identified anomalies and allowing users to select, label, and describe the anomalies. Although this approach yields reliable and interpretable results, it becomes a complex task when data is scaled up. While the manual labeling and conditioning of anomalous points show promising results in preventing false alarms or mistaking frequently occurring anomalous points for normal points, the process makes the system less automated and more reliant on the continual engagement of domain experts.

A more automated yet interpretable method for anomaly detection is to have the model learn from normal data and report unusual deviations, a semi-supervision process. In this instance, the model's reliability depends on the quality of the normal data it is trained on—the likelihood of frequently occurring anomalies being misinterpreted as normal increases significantly.

5 Warning Identification Framework

The Warning Identification Framework (WIF) aims to support the decision making of a cyber-physical system that uses anomaly detection methods to detect warning signals during an ongoing operation. WIF targets anomalies with a low likelihood of occurring but can have severe consequences. Typically, such anomalies are underrepresented in data, necessitating that WIF addresses data biases, a lack of labeled data, and a lack of context in data and anomaly detection methods to provide reliable results. The motivation behind WIF lies in key aspects of multiple disciplines towards operations of autonomous and intelligent sensing systems, adapted from [7]:

- Aspects of future *Risk Assessment*:
  - The recognition of knowledge, the growth of data, and the requirement for robust frameworks for the safety assessment of cyber-physical systems [45].
  - Focus on new events that become apparent in new conditions.
- Aspects of future *Reliability Engineering*, an engineering discipline for applying scientific know-how to a component, product, plant, or process in order to ensure that it performs its intended function, without failure, for the required time duration in specified environment [46]:
  - Fault prevention, removal, and tolerance.
  - Fault forecasting.
  - Reliable functioning under expected circumstances.
- Aspects of future *Resilience Engineering*, a discipline that brings together the system safety concepts, reliability of a system, analysis and handling uncertainties, risks, and survivability of a system (where a resilient system can recover quickly after a shock or an injury) [47]:
  - Anticipation of hazardous events.
  - Monitoring of hazardous events.
  - Responding to hazardous events.
- Aspects of future *Human–Machine Teaming*, a relationship between humans, the machine, and their interdependencies aiming to build trustworthy, transparent, predictable, adaptable, and reliable systems that incorporate AI [48]:
  - On-demand adjustment of autonomy.
  - Explainable functioning of a system.
  - Shared understanding of intentions.
  - Multiple approaches to a single challenge.

Anomaly detection is frequently used in applications to identify unusual data patterns that might harm the system. In comparison, risk analysis identifies hazards as potential sources of harm. Risk analysis and anomaly detection have comparable objectives. As illustrated in Fig. 6, the two disciplines share a common interest in identifying low probability events that may result in high consequences and require extensive data analysis. Therefore, the combination of risk analysis and anomaly detection provides a risk-informed approach to anomaly detection.

The interest in anomaly detection in combination with risk analysis is dependent on the capacity of anomaly detection to provide anomalous points that may be used to identify potential hazardous events, hazards, and threats. The combination of anomaly detection and risk analysis is particularly interesting for autonomous warning management. However, this paper demonstrates that the process's reverse order is equally interesting, particularly in addressing the challenges caused by imbalanced data that contributes to

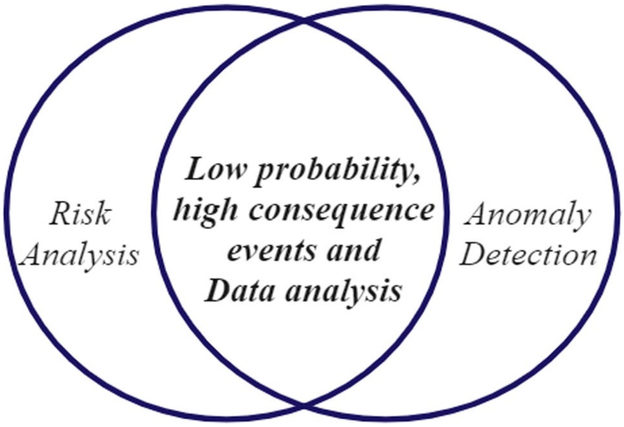


Fig. 6 Risk analysis and anomaly detection overlap

poor anomaly detection outcomes. By using insight from risk analysis, such as the list of possible hazards and their properties, anomaly detection can be guided in detecting the true anomalies that can be of interest for further inspection. The causal analysis, accompanied by an identified sequence of events leading to the potentially hazardous event, can aid in the detection of anomalies. With the analysis of the severity of potential consequences, the detected anomalies can be prioritized, consequentially decreasing the number of false alarms. In light of this, we propose selecting anomaly detection methods that consider the likelihood that true anomalies will occur infrequently. One such method is Isolation Forest, which attempts to eliminate reporting of noise by isolating rare points in the dataset on the assumption that there are fewer true anomalies. Isolation Forest is described in more detail in Sect. 6.2. We divide the process of using risk analysis as a supervisory component to anomaly detection into three steps, with an assumption that historical data exists for risk analysis as an input to WIF (as illustrated in Fig. 7):

5.1 Step 1: Warning Characterization

Given the context and circumstances of the planned operation, such as the operation goals, the assets, expected environmental compounds, location, time, and season, the first step is to answer the question, *“What can go wrong during the given operation and given the context and circumstances?”*. By answering this question, the warning characterization step, through risk analysis, aids in setting the objectives of anomaly detection, as illustrated in Fig. 7. Context and circumstances are crucial for minimizing false alarms during anomaly detection. Since not all occurrences are anomalous under all circumstances, distinguishing hazards and their contextual occurrences makes it easier to overlook expected or insignificant disturbances detected by anomaly detection methods. In addition, it is essential to identify the events or circumstances that contribute to a hazardous event, known as triggering events. While some hazards develop gradually, others occur due to another event, a trigger, typically a technical failure or human error [21]. Furthermore, while a single anomalous phenomenon may not suggest a hazardous occurrence, a collection of several phenomena may. All known or expected variables that may constitute a hazardous occurrence, or an early sign of one, should be included in the step of warning characterization.

5.2 Step 2: Warning Analysis

After determining what can go wrong and compiling a list of hazardous and potentially hazardous occurrences, the second step is to answer the question *“How does the hazard manifest?”* to gain a more profound knowledge of hazards. It is essential to collect as many attributes as possible that can explain the hazard, such as the sequence of events that may lead to their occurrence, frequency, and the likelihood of appearance. The sequence of events can highlight changes in environmental components that may lead to a hazardous event. Inner corrosion in a gas pipeline is an example of a hazard that builds gradually until a gas leak, a hazardous event, occurs [21]. Accordingly, the sequence of events may consist of multiple sensor measurements with specific values and properties that are informative of hazardous occurrence, as determined by domain experts. Rausand [21] describes the first occurrence in a sequence of events that will lead to undesirable outcomes as an initiating event or the event that disrupts the normal operations of the system and may necessitate a response in order to prevent subsequent undesirable outcomes.

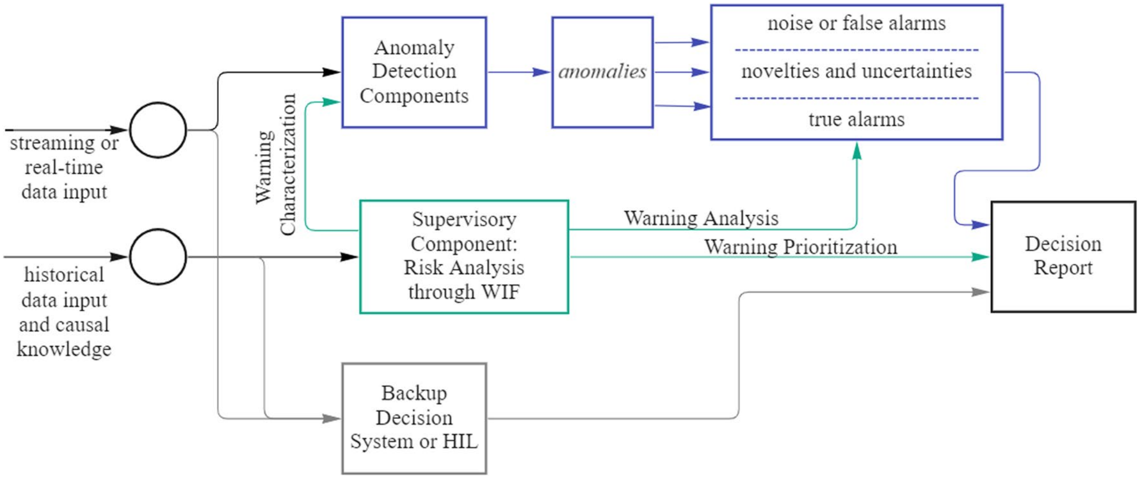

Anomaly Detection Components
anomalies
noise or false alarms
novelties and uncertainties
true alarms
streaming or real-time data input
Warning Characterization
Supervisory Component: Risk Analysis through WIF
Warning Analysis
Warning Prioritization
Decision Report
historical data input and causal knowledge
Backup Decision System or HIL

Fig. 7 Architectural pattern for systems using WIF

The second part of understanding how hazards manifest is by answering, “What is the likelihood of this hazard occurring, and how frequently does it occur?”. While frequency tells us about the number of times an event has happened within a specific timeframe, likelihood answers how probable it is that it will occur [49]. Knowing the frequency and likelihood of a hazard is valuable for dismissing false alarms and recognizing circumstances under which the hazardous events are more likely to occur. As illustrated in Fig. 7, the warning characterization step, through risk analysis, aids the anomaly detection outcome in distinguishing noise from hazards or true and false warnings detected in a group of anomalies. Nonetheless, unsupervised anomaly detection leads to identifying novelties that may have been overlooked during risk analysis.

5.3 Step 3: Warning Prioritization

Knowing whether to respond is the objective of the third step. Figure 7 illustrates the critical component of decision-making of a cyber-physical system responsible for autonomously reporting hazards during ongoing operations. Warning priority is derived from consequence analysis and is responsible for determining the impact of an identified hazardous occurrence. The impact of a hazard prioritizes a response by the autonomous system to notify the operators or supervisory system if and when the situation necessitates it, allowing for early warnings with minimal false alarms.

Figure 7 formalizes the three phases of WIF into an architectural pattern for systems employing WIF and utilizing anomaly detection (or comparable ML approaches) for safety-related decision-making. This type of architecture permits decisions to be risk-informed instead of based on ML-discovered patterns that depend on often unreliable data. Risk analysis through WIF represents a supervisory component for anomaly detection, provided by domain specialists examining historical data and causal knowledge. Incorporating a supervisory component increases the opportunities to address the challenges associated with anomaly detection, such as a high number of false alarms and the inability to differentiate noise from hazards, and other general challenges associated with machine learning methods too, such as bias, lack of context, and lack of explainability. WIF enables anomaly detection to distinguish false alarms, true alarms (potentially hazardous occurrences), uncertainties, and novelties. *Uncertainties and novelties* represent anything unknown. While some publications use the terms *anomalies* and *outliers* interchangeably, other sources [50–52] use the term *outliers* to denote uncertainties or novelties captured by anomaly detection. As part of the architectural pattern for WIF-based systems, as illustrated in Fig. 7, it is recommended to include a backup decision plan that requires human intervention, Human In the Loop (HIL), if the system fails to operate autonomously.

The suggested methods for each WIF step depend on the data, case study, and objectives. The methods for our seismic data case study are described in the following paragraphs.

6 Case Study

The application of the WIF is demonstrated using data acquired by the geophysical station supporting system towards estimating the rock burst hazard using seismic and seismoacoustic techniques [53]. Seismic hazard is one of the most challenging natural hazards to detect and anticipate [54] and can result in devastating consequences during underground activities such as mining and drilling. One of the primary responsibilities of geophysical stations is to determine the current level of seismic hazard, especially the probability of high-energy, destructive seismic tremors that might cause rock bursts during underground activities. For example, rock bursts pose a significant risk to humans onsite during mining operations and can destroy longwalls and damage equipment. The complexity of seismic processes and the imbalanced distribution of favorable”hazardous state” and unfavorable”non-hazardous state” data points pose a significant challenge for predicting seismic hazards using machine learning approaches [54]. The original Seismic dataset is a 19 attribute binary classification dataset. It is an unbalanced dataset in which the positive (hazard) class is in the minority and considered an anomaly class. In contrast, the negative (no hazard) class is considered normal [55]. The list of seismic dataset attributes is presented in Appendix A.

The prediction horizon of the data is eight hours. This eight-hour shift indicates that the prediction methods (anomaly detection and classifiers) make seismic hazard predictions one shift in advance. Continuous data collection necessitates the aggregation of raw data prior to analysis. The aggregation process replaces a series of measurements recorded at eight-hour intervals with a single value. For instance, aggregating measurement data collected over 100 shifts yields a sequence of records or vector of variables x_1, x_2,…, x_{100}, where x_t is a vector of aggregated measurement values characterizing the eight-hour interval or one shift, as denoted in the dataset. After two-month data collection process and aggregation, the seismic dataset consists of 2584 instances.

6.1 Seismic Data Hazard Assessment

Three hazard assessment methods are performed for the seismic data: seismic hazard assessment, seismoacoustic hazard assessment, and seismoacoustic hazard assessment based on only the registration of maximum energy from a geophone

Table 1 Basis of hazard assessment for quantitative method, adapted by [54]

Rockburst haard	Caved faces	Roadways
a No hazard	1. No tremors or single tremors with energies E of the order of 10^2 J-10^3 J $E_{\max} \le 10^4$J	1. No tremors or single tremors with energies E of the order of 10^2 J $E_{\max} \le 10^3$J
	2. $\sum E < 10^5$J per 5 m of longwall advance	2. $\sum E < 10^5$J per 5 m of longwall advance
b Low hazard	1. Occurrence of tremors with energies E of the order 10^2 J-10^5 J $1 \bullet 10^4\text{J} < E_{\max} \le 5 \bullet 10^4$J	1. Occurrence of tremors with energies E of the order 10^2 J-10^3 J $E_{\max} \le 5 \bullet 10^3$J
	2. $1 \bullet 10^5\text{J} \le \sum E < 10^6$J per 5 m of longwall advance	2. $1 \bullet 10^3\text{J} \le \sum E < 10^4$J per 5 m of longwall advance
c Moderate hazard	1. Occurrence of tremors with energies E of the order 10^2 J-10^6 J $5 \bullet 10^5\text{J} < E_{\max} \le 5 \bullet 10^6$J	1. Occurrence of tremors with energies E of the order 10^2 J-10^4 J $5 \bullet 10^3\text{J} < E_{\max} \le 5 \bullet 10^3$J
	2. $1 \bullet 10^5\text{J} \le \sum E < 10^7$J per 5 m of longwall advance	2. $1 \bullet 10^4\text{J} \le \sum E < 10^5$J per 5 m of longwall advance
d High hazard	1. Occurrence of tremors with energies E of the order 10^2 J-10^6 J $E_{\max} > 5 \bullet 10^6$J	1. Occurrence of tremors with energies E of the order 10^2 J-10^5 J $E_{\max} > 10^5$J
	2. $\sum E < 10^7$J per 5 m of longwall advance	2. $\sum E < 10^5$J per 5 m of longwall advance

[54]. The main aim of the three hazard assessments is to predict increased seismic activity, which can cause a rockburst. There are four distinct categories of rockburst hazard: no hazard, low hazard, moderate hazard, and high hazard. The following are the primary assessment factors influencing the hazardous occurrence probability and the condition of rockburst hazard [56, 57]:

- Coal seam thickness;
- The distance between a coal seam and a probable seismogenic layer;
- Maximum seismic energy of tremors from a particular coal seam.

6.1.1 The seismic hazard assessment method

The essence of seismic hazard assessment is observing changes in seismic activity and identifying an increase or decrease in the degree of hazard relative to a previously determined degree [54]. Seismic hazard assessment utilizes qualitative assessment (for low seismic activity) or quantitative assessment (for high seismic activity) based on the strength of seismic tremors. The level of seismic activity is calculated by the quantity and magnitude of seismic tremors recorded in the vicinity of an observed longwall during a specific period (a shift) [54]. Table 1 provides the foundation for quantitative hazard assessment.

6.1.2 The seismoacoustic hazard assessment method

The seismoacoustic method for assessing seismic hazard is based on the relationships between seismoacoustic emission and seismic hazard. In the seismoacoustic method, the following criteria are essential for assessing earthquake risk:

- recording of the seismoacoustic emission;
- the number of pulses recorded by geophones or denoted by seismic energy.

Changes in recorded seismoacoustic activity and energy are the primary evaluation criteria. In addition, deviations (denoted as DEV in Table 2) of values calculated during subsequent time intervals also influence the classification of one of the four seismic hazard states (a,b, c, and d for

Table 2 Seismoacoustic method for hazard assessment, adapted by [54]

Time	25 ≤ DEV ≤ 100	25 < DEV ≤ 100	Decrease of activity/energy after an increase of activity/energy such as 100 < DEV ≤ 200			DEV > 200	Decrease of activity/energy after an increase of activity/energy such as DEV > 200		
			1 shift	2 shift	> 2 shift		1 shift	2 shift	> 2 shift
1 shift	a	b	a	a	current hazard sate -1 after 3 changes of activity/energy drop	c	c	c	current hazard sate -1 after 3 changes of activity/energy drop
1 shifts	a	c	b	b		d	d	d	
1 shifts	b	c	c	c		d	d	d	

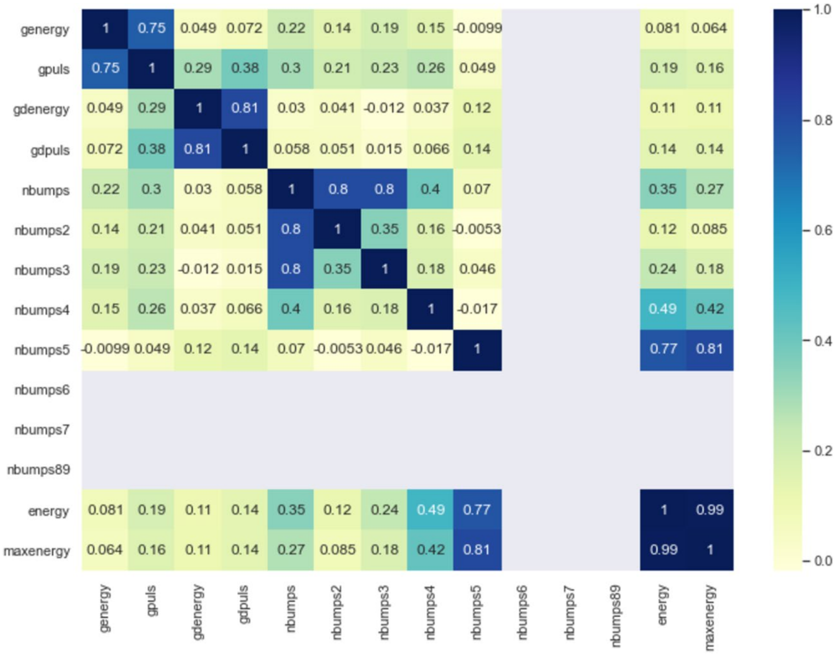


Fig. 8 Heat map correlations between dataset attributes

no, low, medium, and high-impact hazards). Identifying the hazard level is based on the percentage changes in activity/energy value deviations (see Table 2).

6.2 Anomaly Detection for Seismic Data

In order to achieve the most credible results, it is essential to select the anomaly detection method that corresponds to the data description from among the vast number available. Our approach is firstly to determine if the data is Gaussian. If the data is Gaussian, anomalies often reside away from the peak of the normal distribution [58]. The normality test of the seismic dataset, performed with Python Library for statistical calculations *Shapiro–Wilk Test for Normality* [59], in our case study indicates that the seismic data is not Gaussian with p-value approaching 0. The data distribution contains more information than the covariance matrix, which measures how much two random variables change together and is helpful for normal data but less for non-Gaussian data. Plotting noise and artificial anomalies is more difficult for this type of data. Correlation between attributes or their relevance to one another is an additional essential characteristic of data. Figure 8 demonstrates the heat map illustrating the magnitude of the correlation between attributes.

For this dataset we have selected an anomaly detection method *Isolation Forest* that isolates anomalies with minimal computational needs. [60] provides a comprehensive summary of each step of the Isolation Forest algorithm and the

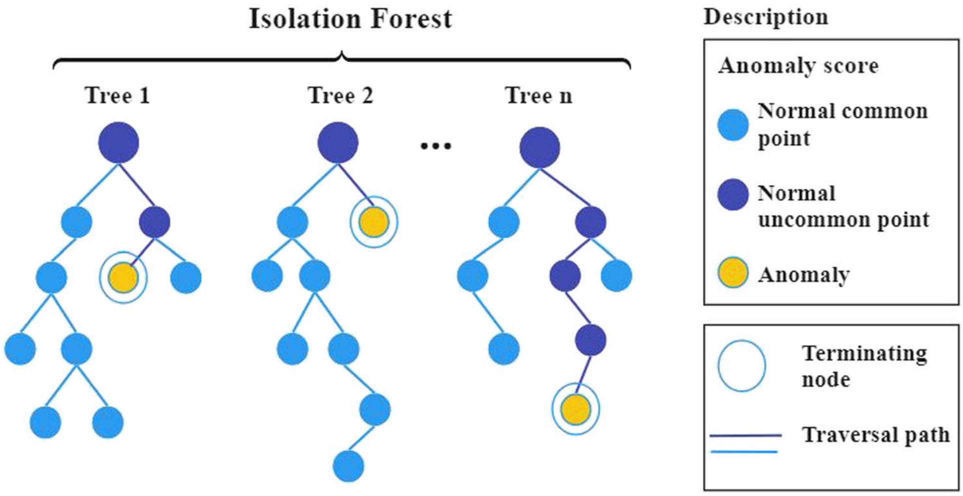


Fig. 9 Isolation Forest, illustrated. Adapted from [61]

underlying equations. Since Isolation Forest is capable of isolating outstanding data points efficiently, it can also be used to determine if these points share similarities with hazards, i.e., if hazards also appear as outstanding points and if they are apparent to both domain experts and anomaly detection methods. If they are difficult to detect, it indicates that the anomalies may not share contextual properties with hazards, so the autonomous approach may need to be modified accordingly. This knowledge can be used to extract what is not apparent for anomaly detection to detect the hazards and determine what properties to introduce to increase apparency and improve autonomous detection, since not every anomaly is a hazard and vice versa. In our case study, Isolation Forest provides a suitable testing environment for measuring an algorithm's capacity in isolating anomalies and determining whether they are comparable to the anomalies that a human domain expert would identify as seismic hazards. Isolation Forest effectively solves high-dimensional problems with multiple non-correlated attributes by constantly and recursively splitting instances until they are isolated by their normal common, normal uncommon, and anomalous occurrence [60], as illustrated in Fig. 9. Isolation Forest does not rely on distance or density metrics to identify anomalies, eliminating processing expenses and making it suitable for nonlinear datasets. It is expected that there are fewer true anomalies in the dataset; therefore, they are more susceptible to isolation, eliminating the overabundance of registered noise or false alarms. It is a widely applied and one of the most developed unsupervised anomaly detection methods. The efficiency of Isolation Forest is in the way it builds a normal data point profile and isolates the points that do not fit that profile, taking advantage of anomalous properties and uncommon values, as illustrated in Fig. 9. Algorithm Part 1, 2, and 3 show the algorithm details of Isolation Forest split into three parts: initialization of a forest, initialization of a single tree (more of which construct a forest), and calculation of traversal path length, a path between the tree node and the isolated anomaly. A group of isolation trees finds anomalies as points with path lengths, with numerous trees functioning as"domain experts" to target the anomalies [60]. Additionally, the Isolation Forest does not need to separate the majority of the training sample consisting of normal examples.

Algorithm Part 1: Creating a forest, adapted from [60]

```
Input: X - input data, $t$ - number of trees, $\psi$ - sub-sampling size
Output: a set of $t$ iTrees
  Initialize Forest
  set height limit $l = ceiling(log2\ \psi)$
  for $i = 1$ to $t$ do
      X' ← sample(X, $\psi$)
      Forest ← Forest ∪ iTree(X', 0, l)
      return Forest
  end for
```

As described in the Algorithm Part 1 and 2, the trees are produced by iteratively splitting the data until instances are isolated or a predetermined tree height is attained, resulting in a partial model. The algorithm automatically determines the tree height limit based on the sub-sampling size, which is denoted as the height limit variable. Finding the average height limit is necessary because shorter-than-average path lengths are more likely to be anomalies. Sub-sampling size ψ that controls the data size is reliably detected by Isolation Forest, keeping the performance, processing time, and memory size optimal. Algorithm Part 3 depicts the evaluating stage in which an anomaly score s is derived from the expected path length for each test instance, which is obtained by passing instances through each tree in the forest. A single path length is determined by counting the number of edges e from the root node to a terminating node as an instance traverses a tree.

When a single path is obtained for each tree in the forest, an anomaly score s is derived following the Eq. 1, where $h(x)$ denotes the path length, $E(h(x))$ is the normalized $h(x)$ from a collection of isolation trees, and $c(n)$ is the average of path lengths. Finally, the data are then sorted in descending order to identify the most significant anomalies. The results of Isolation Forest application of our case study are described in Sect. 6.3.

$$s = 2 - \frac{E(h(x))}{c(n)} \tag{1}$$

Algorithm Part 2: Creating a tree, adapted from [60]

```
Input: X - input data, $e$ - current tree height, $l$ - height limit
Output: an $an\ iTree$
  if $e \geq l$ or $X \leq 1$ then
      return $exNode$ $\{$ Size $\leftarrow$ | X | $\}$
  else
      let Q be a list of attributes in X,
      randomly select an attribute $q \in Q$,
      randomly select a split point p from $max$ and $min$ values of q in X,
      $X_l \leftarrow$ filter(X, q ¡ p),
      $X_r \leftarrow$ filter(X, q $\geq$ p),
      return $inNode$ $\{$
      $Left \leftarrow iTree(X_{l,e}$ + 1, l),
      $Right \leftarrow iTree(X_{r,e}$ + 1, l),
      $SplitAtt \leftarrow q$,
      $SplitValue \leftarrow p$
  end if
```

Algorithm Part 3: Calculating path length, adapted from [60]

```
Input: $x$ - an instance, $T$ - an iTree, $e$ - current path length; to be initialized
to zero when first called
Output: path length of $x$
  if $T$ is an external node then
      return $e + c\ (T.size)$ $\{$where c is average search path$\}$
  end if
  $a \leftarrow T.splitAtt$
  if  $X_a < T.splitValue$ then
      return $PathLength(x, T.left, e + 1)$
  else $X_a \geq T.splitValue$
      return $PathLength(x, T.right, e + 1)$
  end if
```

6.3 WIF Steps: Application of Risk Definition

6.3.1 Step 1: Warning Characterization

Answering the question *"What can go wrong during the given operation and given the context and circumstances?"* necessitates domain expert observations. For the seismic dataset, this is answered through three hazard assessment methods. Tables 1 and 2 provide the hazards as the events that *can go wrong*. These hazardous events serve as the ground truth for testing the capability of anomaly detection method to detect the same events as anomalies. The results of the hazard assessment methods are shown in Table 3. The three methods do not yield the same amount of hazardous and non-hazardous states. Upon closer inspection, the number of equal instances of the non-hazardous state resulting from seismic and seismoacoustic hazard assessment methods is 1071, and the number of equally denoted hazardous states is 393. As suggested by [54], knowledge of the present hazard state is essential for production process management and industrial safety. However, assessing and predicting seismic hazards is a highly complex procedure with a substantial element of randomness.

Table 3 Number of hazardous and non-hazardous instances labeled by hazard assessment methods

Hazard Assessment Methods			
State	Seismic	Seismoacoustic	Seismoacoustic by max energy
Non-hazardous	1682	1580	2342
Hazardous	902	1004	242

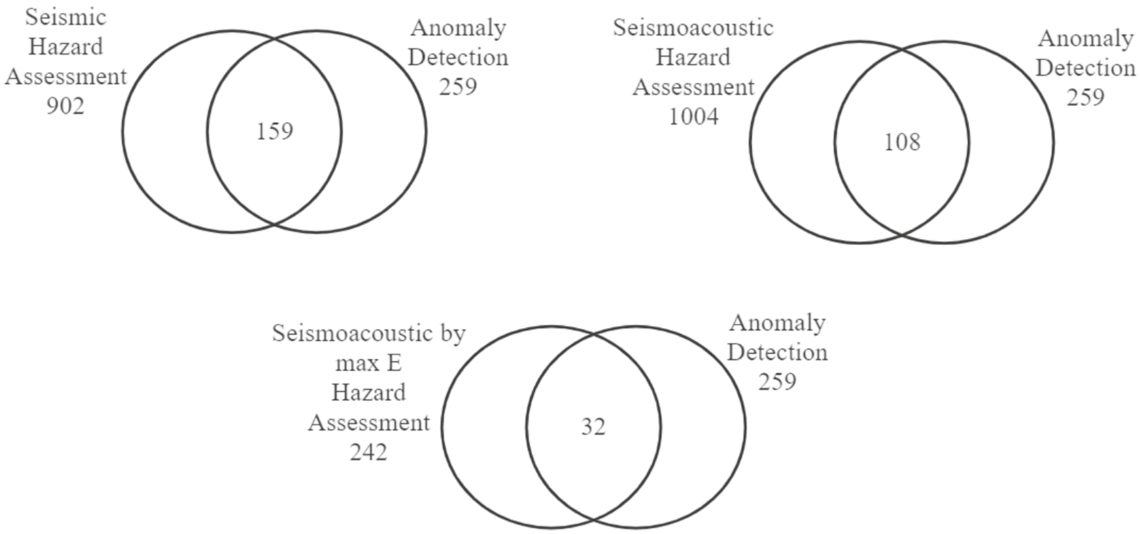

Seismic Hazard Assessment 902
159
Anomaly Detection 259
Seismoacoustic Hazard Assessment 1004
108
Anomaly Detection 259
Seismoacoustic by max E Hazard Assessment 242
32
Anomaly Detection 259

Fig. 10 Hazards identified by anomaly detection and (**a**) seismic hazard assessment, (**b**) seismoacoustic hazard assessment, (**c**) seismoacoustic by max energy hazard assessment methods

Treating each result of hazard assessment methods as different ground truths, the information derived by domain expert observations and known as the absolute truth, the anomaly detection results show significantly different numbers (see Fig. 10). Out of the 259 detected anomalies, 159 are the hazardous state identified by seismic (Fig. 10 (a)), and 108 are by the seismoacoustic hazard assessment method (Fig. 10(b)). These results lead to an early conclusion that approximately half of the anomalies detected by the anomaly detection method are considered hazardous, and the other detected anomalies are of no significance. Compared to the results of seismoacoustic by max energy results of 242 hazardous states, anomaly detection has identified only 32 (see Fig. 10 (c)). The Fig. 10 illustrates the critical difference and the main shortcoming of the anomaly detection method, the inability to independently detect true hazards and a substantial number of false alarms. The confusion matrices in Tables 4, 5, and 6 provide additional insight into these results. Despite three distinct seismic hazard assessment methods representing hazard occurrences, seismic, seismoacoustic, and seismoacoustic by maximum energy, Isolation Forest demonstrated an insufficient understanding of hazards. This evidence may prompt an early proposition that unsupervised anomaly detection may not be appropriate for seismic hazard detection despite its widespread use for unusual patterns and threat detection and that seismic hazard assessment is required as an element of supervision.

6.3.2 Step 2: Warning Analysis

Warning Analysis intends to detect patterns in which the hazards may occur and the likelihood of their occurrence. Conditional probability P, Eq. 2, is the likelihood that an event A or outcome will occur given the occurrence of a prior event or outcome B, C, D [49]. Multiplying the likelihood of the preceding event by the updated probability of the subsequent, or conditional, occurrence yields the conditional probability.

$$P(A|B,C,D) = \frac{P(A \cap B)}{P(B,C,D)} \tag{2}$$

Table 7 shows the results of conditional probability of each nonhazardous occurrence and hazard, categorized by their impact. Table 7 represents the probability for

Table 4 Confusion Matrix with Seismic Hazard Assesment as True Values, Isolation Forest Anomalies as Test Values

		Anomalies by Isolation Forest		
		Positive	Negative	Total
Hazards by Seismic	Positive	159	743	902
	Negative	100	1582	1682
	Total	259	2325	

Table 5 Confusion Matrix with Seismoacoustic Hazard Assesment as True Values, Isolation Forest Anomalies as Test Values

		Anomalies by Isolation Forest		
		Positive	Negative	Total
Hazards by Seismoac	Positive	108	896	1004
	Negative	151	1429	1580
	Total	259	2325	

Table 6 Confusion Matrix with Seismoacoustic by Max Energy Hazard Assesment as True Values, Isolation Forest Anomalies as Test Values

		Anomalies by Isolation Forest		
		Positive	Negative	Total
Hazards by Max Energy	Positive	32	210	242
	Negative	227	2115	2342
	Total	259	2325	

each state (either no hazard, low, medium, high-impact) given the occurrence of the other states. The impact of the hazard is derived following the causal knowledge described in Tables 1 and 2 on hazard detection patterns with the task of hazard prediction based on the association between the energy of recorded seismic tremors and seismoacoustic activity with the probability of seismic tremor occurrence [54].

Even though this case study has an exact number of seismic hazards, it is not always expected that seismic tremor monitoring operations will have sensor data for identifying hazards analyzed by domain experts. In the event of not having an exact number of hazards derived from sensor data by domain experts, knowing the frequency of a hazardous occurrence in a given environment may help compensate for situations in which a large number of anomalies are reported in order to determine the likelihood of the anomaly being a true hazard or a false alarm.

6.3.3 Step 3: Warning Prioritization

The three approaches to hazard assessment for seismic data provide the impact of the hazard on four levels (hazard impacts derived following the Table 1):

1. No hazard
2. Low-impact hazard
3. Medium-impact hazard
4. High-impact hazard

Table 7 Conditional probability of hazard occurring, by hazard assessment methods

Conditional Probability by Hazard Impact, expressed in percentages			
State	Seismic	Seismoacoustic	Seismoacoustic by max energy
No hazard	65.09	61.14	90.63
Low-impact hazard	34.90	36.99	8.20
Medium-impact hazard	1.80	1.85	1.16
High-impact hazard	0	0	0

Table 8 Hazard impacts by hazard assessment methods

Hazard Assessment Methods: Hazard Impact			
State	Seismic	Seismoacoustic	Seismoacoustic by max energy
No hazard	1682	1580	2342
Low-impact hazard	902	956	212
Medium-impact hazard	0	48	30
High-impact hazard	0	0	0

Table 8 shows the number of hazards, categorized by their impact, detected by the three hazard assessment methods. In comparison, Table 9 represents the number of anomalies detected by the unsupervised anomaly detection method, Isolation Forest, where each hazard assessment method is used to categorize the hazards and their impacts among the detected anomalies.

As presented in Table 8, during the hazard assessment, there were no records of high-impact hazardous occurrences. The seismic hazard assessment method has identified only low-impact hazards, and no medium or high impact hazards. According to the impact, the reactions during operations can be prioritized.

Anomaly detection has provided poor results concerning the identification of various levels of hazard impacts, presented in Table 9. For low impact hazards, anomaly detection has, on average, identified only 14,2% of the low impact hazards, and for medium-impact hazards, only 14,5% of the cases on average. These results indicate that unsupervised anomaly detection cannot reliably identify seismic hazards and distinguish them based on their severity impact. Therefore, a form of supervision, as demonstrated with different hazard assessment approaches, is necessary to introduce.

Table 9 Hazard impacts identified by anomaly detection methods, with hazard assessment methods as the ground truth

Anomalies detecting hazard impacts			
State	Seismic	Seismoacoustic	Seismoacoustic by max energy
Low-impact hazard	166	104	29
Medium-impact hazard	0	6	5
High-impact hazard	0	0	0

6.3.4 Case Study Summary and Opportunities for a Generalized Framework

The case study architecture of WIF, applied to identify seismic hazards among detected anomalies by unsupervised anomaly detection, is illustratated in Fig. 11. Isolation Forest, a method for unsupervised anomaly detection, analyzes unlabeled seismic sensor data and detects a group of anomalies. However, within the detected anomalies, there is yet no knowledge of which ones are false warnings and the ones that are true warnings or hazards. In this case study, domain expert knowledge is leveraged through hazard assessment criteria based on three methods: seismic, seismoacoustic, and seismoacoustic with maximum energy. Hazards, or true warnings, can be extracted from the given dataset and compared to anomalies detected by the unsupervised method to determine if the unsupervised method can capture the properties of hazards and report them as anomalies while ignoring false warnings. These hazard anomalies can be prioritized based on their impact, such as none, low, medium, and high.

As the use of data-driven and machine learning methods increases, the problem of unintended and harmful behavior of machine learning systems resulting from poor design of real-world AI systems becomes increasingly apparent [62]. Unsupervised anomaly detection, classification, and other data-driven machine learning methods face well-known challenges:

- biased data,
- false positives and false negatives (false alarms),
- prioritization of anomaly reporting for anomaly detection applications,
- lack of context that is tied to all of the previous challenges, and
- lack of explainability of the results produced by unsupervised methods

Introducing a supervisory component to data-driven systems is a step toward providing context to the method, reducing biases and false reporting, adding prioritization knowledge, and improving the explainability of the results as they are derived from more traditional risk, and hazard assessment approaches. The approach studied in this paper can be generalized by observing risk assessment methods and properties of hazards for a specific operation, where hazard properties may serve as a class label by which the unsupervised data-driven method can be validated. According to a technical report and recommendations on AI and safety by ISO/IEC, 2022 [63], providing explainable algorithms and results and validating them in the real world characterizes the future of AI-related systems and safety.

7 Discussion

The results of the case study show shortcomings of an unsupervised anomaly detection method through a clear difference between the identified hazards by three seismic hazard assessment methods and their results with unsupervised anomaly detection method Isolation Forest. The findings from Step 1 Warning Characterization provide crucial insights into unsupervised anomaly detection and seismic hazard assessment differences. During safety–critical operations, such as seismic hazard monitoring, it is essential to assess the difference between the discovered hazardous

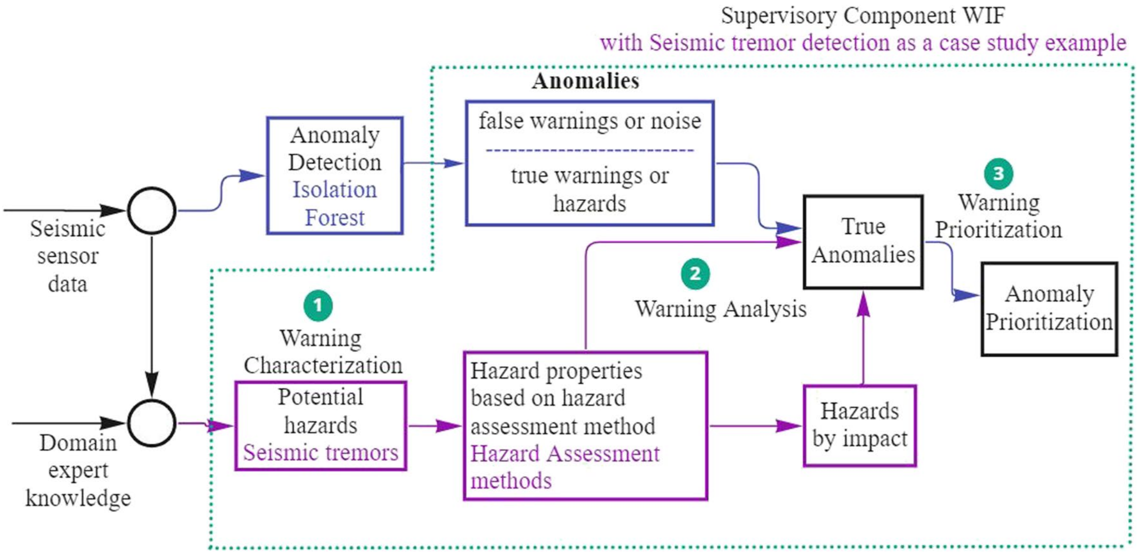

Supervisory Component WIF
with Seismic tremor detection as a case study example
Anomalies
false warnings or noise
true warnings or hazards
Anomaly Detection Isolation Forest
Seismic sensor data
Domain expert knowledge
1
Warning Characterization
Potential hazards Seismic tremors
Hazard properties based on hazard assessment method Hazard Assessment methods
2
Warning Analysis
True Anomalies
Hazards by impact
3
Warning Prioritization
Anomaly Prioritization

Fig. 11 Case Study Summary Architecture

states and adapt our expectations for the implementation of anomaly detection. Since it is not expected that an operation will have labeled training dataset, and identified hazards list by domain experts for each case, the analysis of the data becomes unsupervised. This step showed unexpected differences between the number of seismic hazards identified by domain expert inputs and anomalies identified by unsupervised anomaly detection. An unexpectedly low number of detected anomalies proved to be hazards when categorized following the hazard identification criteria presented in Tables 1 and 2. This leads to an assumption that unsupervised anomaly detection, despite being used for detection of threats and unusual patterns, cannot be trusted to detect all seismic hazards. An additional layer of context, namely through hazard assessment methods, is necessary to distinguish the anomalies that are only data discrepancies and offer no significance, from the ones that are hazardous.

The results of conditional probability obtained in Step 2, Warning Analysis, and based on hazard assessment methods, are crucial to setting the expectation of the occurrence of a hazard of varying degrees of impact. The seismic and seismoacoustic hazard assessment methods resulted in the highest probability of a non-hazardous event, followed by a low-impact hazard and a low probability of medium-impact hazard occurrence. The data in this case study provided no evidence of high-impact hazards, resulting in an expected no probability of high-impact hazardous occurrences. In comparison, the seismoacoustic by maximum energy method resulted in the highest probability of 90.53% of non-hazardous occurrence, followed by a low probability of low impact and medium-impact hazards. This step showed a limitation in the case study data where the lack of high-impact hazard evidence resulted in a 0% probability of such hazards occurring. This imbalance in hazard impacts can lead to biases during anomaly detection or hazard identification methods.

Further analysis in Step 3, Warning Prioritization, categorized the identified hazards in the varying degrees of impact: no hazard, low-impact, medium-impact, and high-impact. The anomaly detection method resulted in fewer identified hazards than the hazard assessment methods. This step showed the low reliability of the anomaly detection method as an autonomous hazard identification approach.

The case study results have validated the assertion that unsupervised anomaly detection generates a considerable amount of false alarms, that may waste operator response resources if the methods are used as a part of an autonomous drone or smart-sensor system. These results provide valuable insight into the possibilities of addressing the shortcomings of unsupervised anomaly detection methods for seismic hazard identification, where risk assessment approaches, such as hazard identification, can play a crucial role.

8 Conclusion

It is anticipated that cyber-physical and intelligent sensor systems will play a permanent role in industrial operations, including monitoring, inspecting, and intervening with assets and the environment, necessitating greater autonomy for making significant decisions in near-real and real-time. Current challenges include a lack of context, the underutilization of causal knowledge, and an excess of imbalanced data. We discussed the growing need for employing data-driven methods in a more explainable, transparent, and reliable practice.

Recent research provides different approaches to handling discussed challenges through simulations, rule-based classification, and decision boundaries. However, these approaches do not address the explain ability of the data-driven methods and introduce new complexities. The results and contributions of this paper can be summarized as follows:

1. A novel outlook on utilizing existing domain knowledge in seismic tremors through seismic hazard assessment methods as a supervisory component for unsupervised anomaly detection through the Warning Identification Framework based on risk assessment, resilience and reliability engineering, and future human–machine teaming expectations.
2. Identification of overlapping tasks for risk assessment and anomaly detection objectives that can be utilized in addressing the shortcomings of anomaly detection results.
3. A case study examining the sensor-obtained seismic data for monitoring seismic tremors and analyzing three different hazard identification methods in comparison to unsupervised anomaly detection for hazard identification.

During our analysis, we identified significant anomaly shortcomings in detection methods to detect hazardous occurrences by their levels of impact and to distinguish anomalies of no significance from the anomalies that represent hazardous occurrences. The results of this research show significant opportunities in utilizing risk assessment insights to tackle the shortcomings of unsupervised anomaly detection methods and aid a more reliable and transparent hazard detection.

9 Future Work

Future work on the Warning Identification Framework consists of testing more anomaly detection methods on different case studies and hazard identification inputs from domain experts and researching the role of uncertainty analysis for WIF applications. An additional task for WIF expansion is to utilize hazard identification or risk assessment picture for anomaly detection training, using fuzzy logic to represent varying degrees of hazard impact. The evidence of a large number of detected anomalies not representing known hazards additionally opens opportunities to observe these anomalies as a potential to uncover uncertainties not yet addressed by seismic hazard assessment methods. To expand the application, we plan to analyze the time-series image data of underwater pipeline inspections. We plan to interview domain experts in pipeline surveillance for risk assessment and hazard identification inputs and test the framework on streaming data, focusing on methods for anomalous change detection.

10 A Seismic dataset attributes

1. seismic: result of shift seismic hazard assessment in the mine working obtained by the seismic method (a lack of hazard, b low hazard, c high hazard, d danger state);
2. seismoacoustic: result of shift seismic hazard assessment in the mine working obtained by the seismoacoustic method;
3. shift: information about type of a shift (W coal-getting, N -preparation shift);
4. genergy: seismic energy recorded within previous shift by the most active geophone (GMax) out of geophones monitoring the longwall;
5. gpuls: a number of pulses recorded within previous shift by GMax;
6. gdenergy: a deviation of energy recorded within previous shift by GMax from average energy recorded during eight previous shifts;
7. gdpuls: a deviation of a number of pulses recorded within previous shift by GMax from average number of pulses recorded during eight previous shifts;
8. ghazard: result of shift seismic hazard assessment in the mine working obtained by the seismoacoustic method based on registration coming form GMax only;
9. nbumps: the number of seismic bumps recorded within previous shift;
10. nbumps2: the number of seismic bumps (in energy range $[10^2,10^3)$) registered within previous shift;
11. nbumps3: the number of seismic bumps (in energy range $[10^3,10^4)$) registered within previous shift;
12. nbumps4: the number of seismic bumps (in energy range $[10^4,10^5)$) registered within previous shift;
13. nbumps5: the number of seismic bumps (in energy range $[10^5,10^6)$) registered within the last shift;
14. nbumps6: the number of seismic bumps (in energy range $[10^6,10^7)$) registered within previous shift;
15. nbumps7: the number of seismic bumps (in energy range $[10^7,10^8)$) registered within previous shift;
16. nbumps89: the number of seismic bumps (in energy range $[10^8,10^10))$ registered within previous shift;
17. energy: total energy of seismic bumps registered within previous shift;
18. maxenergy: the maximum energy of the seismic bumps registered within previous shift;
19. class: the decision attribute ‘1’ means that high energy seismic bump occurred in the next shift (’hazardous state’),’0’ means that no high energy seismic bumps occurred in the next shift (’non-hazardous state’) generated during rule-based classification experiment by [54]

Acknowledgements This research is a part of Better Resource Utilization in 21st Century (BRU21), Norwegian University of Science and Technology (NTNU) Research and Innovation Program on Digital and Automation Solutions for the Oil and Gas Industry (www.ntnu.edubru21) and supported by Equinor.

Authors' Contributions All authors contributed to the research conception. Rialda Spahic performed material preparation, data gathering, analysis, and manuscript writing. Mary Ann Lundteigen performed writing reviews and supervision of all prior drafts of the manuscript. Vidar Hepsø contributed to the concept visualisation of the research. All authors reviewed and commented on prior manuscript versions. All authors read and approved the final manuscript.

Funding Open access funding provided by NTNU Norwegian University of Science and Technology (incl St. Olavs Hospital - Trondheim University Hospital). This research is supported by Better Resource Utilization in 21st Century (BRU21) Innovation Program at the Norwegian University of Science and Technology Research.

Data Availability The datasets generated during and/or analysed during the current study are available in the University of California (UC) Irvine Machine Learning Repository, available at https://archive.ics.uci.edu/ml/datasets/seismic-bumps.

Declarations

Ethics approval Not applicable.

Consent to Participate Not applicable.

Consent for Publication Not applicable.

Conflicts of Interest There are no financial or non-financial interests to disclose by the authors.

References

1. Vachtsevanos, G., Lee, B., Oh, S., Balchanos, M.: Resilient Design and Operation of Cyber Physical Systems with Emphasis on Unmanned Autonomous Systems. J. Intell. Robot. Syst: Theory Appl. **91**(1), 59–83 (2018). https://doi.org/10.1007/s10846-018-0881-x
2. McDermid, J.A., Yan J., Habli I.: Towards a framework for safety assurance of autonomous systems', proceedings of the workshop on artificial intelligence safety 2019, 28th International Joint Conference on Artificial Intelligence, Vol. 2419, CEUR-WS.org, Macao, China
3. Oxford University Press. Oxford Learner's Dictionaries, Oxford University (2021). URL https://www.oxfordlearnersdictionaries.com/. Accessed 20 Feb 2023
4. Fraser, K., Homiller, S., Mishra, R.K., Ostdiek, B., Schwartz, M.D.: Challenges for Unsupervised Anomaly Detection in Particle Physics. Journal of High EnergyPhysics, Springer Science and Business Media (2022) p. 3. URL https://doi.org/10.1007/jhep03%282022%29066
5. Eldevik, S., Pedersen, F.B.: Safety implications for artificial intelligence why we need to combine causal-and data-driven models, DNV GL AS Oil & Gas Safety Risk Magement (2018). URL https://ai-andsafety.dnv.com. Accessed 20 Feb 2023
6. Spahic, R., Hepsø, V., Lundteigen, M.A.: Enhancing Autonomous Systems' Awareness Conceptual Categorization of Anomalies by Temporal Change During Real-Time Operations. The Eighteenth International Conference on Autonomic and Autonomous Systems pp. 25–30 (2022). ISBN:978–1–61208–966–9
7. Spahic, R., Hepso, V., Lundteigen, M.A.: Reliable Unmanned Autonomous Systems: Conceptual Framework for Warning Identification during Remote Operations. 2021 IEEE International Symposium on Systems Engineering (ISSE) pp. 1–8 (2021). https://doi.org/10.1109/ISSE51541.2021.9582534. URL https://ieeexplore.ieee.org/document/9582534/
8. Spahic, R., Hepsø, V., Lundteigen, M.A.: *Proceedings of the 32nd European Safety and Reliability Conference (ESREL 2022)*, ed. by M.C. Leva, E. Patelli, L. Podofillini, S. Wilson. pp. 273–280. Research Publishing, Singapore, Singapore (2022). https://doi.org/10.3850/978-981-18-5183-4_R08-03-390-cd. URL https://rpsonline.com.sg/rps2prod/esrel22-epro/html/toc.html
9. Aggarwal, C.C.: Outlier Analysis. chap. 1. Springer, Cham, pp. 1–34 (2017). https://doi.org/10.1007/978-3-319-47578-3_1

10 Taha, A., Hadi, A.S.: Anomaly detection methods for categorical data: A review. ACM Comput Surv **52**(2), 1–35 (2019). https://doi.org/10.1145/3312739

11 Hawkins, D.M.: Identification of Outliers. Springer, Netherlands, Dordrecht (1980). https://doi.org/10.1007/978-94-015-3994-4

12. Beckman, R.J., Cook, R.D.: Outliers. Technometrics **25**(2), 119–149 (2012). https://doi.org/10.1080/00401706.1983.10487840
13. Foorthuis, R.: On the nature and types of anomalies: a review of deviations in data. Int J Data Sci Anal **12**(4), 297–331 (2021). https://doi.org/10.1007/s41060-021-00265-1
14. Chandola, V., Banerjee, A., Kumar, V.: Anomaly detection. ACM Comput Surv (CSUR) **14**(1), 1–22 (2009). https://doi.org/10.1145/1541880.1541882

15 Fisch, A., Eckley, I., Fearnhead, P.: Subset Multivariate Collective And Point Anomaly Detection. J Comput Graph Stat **31**, 574–585 (2019). https://doi.org/10.1080/10618600.2021.1987257

16. Hayes, M.A., Capretz, M.A.: *Proceedings 2014 IEEE International Congress on Big Data, BigData Congress 2014* (Institute of Electrical and Electronics Engineers Inc.), pp. 64–71 (2014). https://doi.org/10.1109/BigData.Congress.2014.19
17. Xiuyao, S., Mingxi, W., Jermaine, C., Ranka, S.: Conditional anomaly detection. IEEE Trans. Knowl. Data Eng. **19**(5), 631–644 (2007). https://doi.org/10.1109/TKDE.2007.1009
18. Goldstein, M., Uchida, S.: A Comparative Evaluation of Unsupervised Anomaly Detection Algorithms for Multivariate Data. PLOS ONE **11**(4), e0152173 (2016). https://doi.org/10.1371/JOURNAL.PONE.0152173
19. ISO 31000, Risk management — Guidelines, International Organization for Standardization. Tech. rep., International Organization for Standardization (2018). URL https://www.iso.org/obp/ui/#iso:std:iso:31000:ed-2:v1:en
20. ISO:51. Safety aspects Guidelines for their inclusion in standards ISO/IEC Guide 51:2014(E) (2014)
21. Rausand, M.: Risk Assessment Theory, Methods, and Applications. John Wiley & Sons Inc, Hoboken (2011). https://doi.org/10.1002/9781118281116

22 Michau, G., Fink, O.: Unsupervised transfer learning for anomaly detection: Application to complementary operating condition transfer. Knowl. Based Syst. **216**, 106816 (2021). https://doi.org/10.1016/J.KNOSYS.2021.106816

23. Scharpf, E.W., Thomas, H.W., Stauffer, T.R.: *Practical SIL Target Selection: Risk Analysis Per the IEC 61511 Safety Lifecycle*, 2nd edn. (exida.com LLC, Sellersville, Pennsylvania) (2012)
24. Garcia, R., Sreekanti, V., Yadwadkar, N., Crankshaw, D., Gonzalez, J.E., Hellerstein, J.M.: *Common Model Infrastructure.* London, UK (2018)
25. Schelter, S., Biessmann, F., Januschowski, T., Salinas, D., Seufert, S., Szarvas, G.: On Challenges in Machine Learning Model Management. Bulletin of the IEEE Computer Society Technical Committee on Data Engineering pp. 5–13 (2018). URL http://sites.computer.org/debull/A18dec/p5.pdf
26. Derakhshan, B., Rezaei Mahdiraji, A., Abedjan, Z., Rabl, T., Markl, V.: *Proceedings of the ACM SIGMOD International Conference on Management of Data* (Association for Computing Machinery), pp. 1701–1716 (2020). https://doi.org/10.1145/3318464.3389715
27. Lacher, A.R.: A Framework for Discussing Trust in Increasingly Autonomous Systems. Tech. rep., The MITRE Corporation (2017)
28. Lee, M.S.A., Singh, J.: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (2021)
29. Madras, D., Creager, E., Pitassi, T., Zemel, R.: Fairness through causal awareness: Learning causal latent-variable models for biased data. FAT* 2019 Proceedings of the 2019 Conference on Fairness, Accountability, and Transparency. pp. 349–358 (2019). https://doi.org/10.1145/3287560.3287564
30. Makhlouf, K., Zhioua, S., Palamidessi, C.: On the applicability of ML fairness notions. arXiv pp. 1–32 (2020)
31. Sekar, R., et al.: *Proceedings of the 9th ACM conference on Computer and communications security CCS '02* (Association for Computing Machinery, New York, NY, USA), p. 265–274 (2002). https://doi.org/10.1145/586110.586146
32. Patcha, A., Park, J.M.: An overview of anomaly detection techniques: Existing solutions and latest technological trends. Comput. Netw. **51**(12), 3448–3470 (2007). https://doi.org/10.1016/J.COMNET.2007.02.001
33. G¨opfert, C., Ben-David, S., Bousquet, O., Gelly, S., Tolstikhin, I., Urner, R.: When can unlabeled data improve the learning rate? arXiv **1905.11866** (2019). URL https://arxiv.org/abs/1905.11866v1
34. Henne, M., Schwaiger, A., Weiss, G.: Managing Uncertainty of AI-based Perception for Autonomous Systems. AISafety@IJCAI (2019)
35. Phillip Durst, S.J., Gray, W.: ERDC/GSL SR-14–1 "Levels of Autonomy and Autonomous System Performance Assessment for Intelligent Unmanned Systems". Tech. rep., The US Army Engineer Research and Development Center (ERDC), Vicksburg, MS (2014). URL www.erdc.usace.army.mil.
36. Marshall, C., Roberts, B., Grenn, M.: *3rd International Conference on Control, Automation and Robotics, ICCAR 2017* (Institute of Electrical and Electronics Engineers Inc.), pp. 438–443 (2017). https://doi.org/10.1109/ICCAR.2017.7942734

37. Hollnagel, E., Woods, D.D., Leveson, N.: *Resilience Engineering: Concepts and Precepts* (ASgate Publishing Ltd) (2007). URL https://books.google.no/books?hl=en&lr=&id=rygf6axAH7UC&oi=fnd&pg=PP1&dq=hollnagel+resilience+engineering+concepts+and+precepts+2007&ots=iq5GQV42bb&sig=mK37zLFtfiltAKZMV-JoNpT96Po&redir_esc=y#v=onepage&q=hollnagelresilienceengineeringconceptsandprece
38. O'Neil, C.: Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy, p. 272. Crown Publishers, New York (2016)
39. Zhang, T., Chen, J., Li, F., Zhang, K., Lv, H., He, S., Xu, E.: Intelligent fault diagnosis of machines with small & imbalanced data: A state-of-the-art review and possible extensions. ISA Trans. **119**, 152–171 (2022). https://doi.org/10.1016/j.isatra.2021.02.042
40. Lavin, A., Ahmad, S.: *IEEE 14th International Conference on Machine Learning and Applications, ICMLA 2015* (Institute of Electrical and Electronics Engineers Inc., Miami, Florida, USA), pp. 38–44 (2015). https://doi.org/10.1109/ICMLA.2015.141
41. Li, P., Niggemann, O., Hammer, B.: *Proceedings of the IEEE International Conference on Industrial Technology*, vol. 2019-Febru (Institute of Electrical and Electronics Engineers Inc.), pp. 1311–1316 (2019). https://doi.org/10.1109/ICIT.2019.8754997
42 Shin, D.H., Park, R.C., Chung, K.: Institute of Electrical and Electronics Engineers Inc. IEEE Access **8**, 108664–108674 (2020). https://doi.org/10.1109/ACCESS.2020.3000638
43. Omar, S., Ngadi, A., Jebur, H.H.: Machine Learning Techniques for Anomaly Detection: An Overview. Int. J. Comput. Appl. **79**(2), 33–41 (2013). https://doi.org/10.5120/13715-1478
44. Deng, J., Brown, E.T.: *EuroVis 2021* (The Eurographics Association, Chicago, IL, U.S.A.) (2021). https://doi.org/10.2312/evs.20211050
45. Zio, E.: The future of risk assessment. Reliab. Eng. Syst. Saf. **177**(March), 176–190 (2018). https://doi.org/10.1016/j.ress.2018.04.020
46. Kiran, D.: *Total Quality Management* (Elsevier), pp. 391–404 (2017). https://doi.org/10.1016/B978-0-12-811035-5.00027-1. URL https://linkinghub.elsevier.com/retrieve/pii/B9780128110355000271
47. Hollnagel, E.: Resilience Engineering A New Understanding of Safety J. Ergon. Soc. Korea 35(3),185-191 (2016). https://doi.org/10.5143/jesk.2016.35.3.185. URL http://jesk.or.kr
48. P. Mcdermott, C. Dominguez, N. Kasdaglis, M. Ryan, I.T. Mitre, A. Nelson, Human-Machine Teaming Systems Engineering Guide. Tech. rep., The MITRE Corporation (2018). URL https://www.mitre.org/publications/technical-papers/human-machine-teaming-systems-engineering-guide
49. Schweder, T., Hjort, N.L.: *Confidence, Likelihood, Probability,* pp. 1–22. Cambridge University Press (2016). https://doi.org/10.1017/CBO9781139046671.002. URL https://www.cambridge.org/core/product/identifier/CBO9781139046671A008/type/bookpart
50. K.G. Mehrotra, C.K. Mohan, H. Huang, *Anomaly Detection Principles and Algorithms*, 1st edn. Springer, Cham, pp. 21–32 (2017). https://doi.org/10.1007/978-3-319-67526-8
51. Aggarwal, C.C.: Outlier Analysis. pp. 399–422. Springer, Cham (2017) https://doi.org/10.1007/978-3-319-47578-3_13.
52. Markou, M., Singh, S.: Novelty detection: A review Part 1: Statistical approaches. Signal Process. **83**(12), 2481–2497 (2003). https://doi.org/10.1016/j.sigpro.2003.07.018
53. Sikora, M., Mazik, P.: Towards the better assessment of a seismic hazard—the Hestia and Hestia map systems. Mechanizat. Automat. Min. **3**(457), 5–12 (2009)
54. Kabiesz, J., Sikora, B., Sikora, M., Wrobel, L.: Application of rule-based models for seismic hazard prediction in coal mines. Acta Montanist. Slovaca **18**(4), 262–277 (2013)
55. Sathe, S., Aggarwal, C.: LODES: Local density meets spectral outlier detection. 16th SIAM International Conference on Data Mining 2016, SDM 2016 pp. 171–179 (2016). https://doi.org/10.1137/1.9781611974348.20. URL https://epubs.siam.org/terms-privacy
56. Bukowska, M.: The probability of rockburst occurrence in the Upper Silesian Coal Basin area dependent on natural mining conditions. J. Min. Sci. **42**(6), 570–577 (2006). https://doi.org/10.1007/S10913-006-0101-0
57. Li, Z.L., He, X.Q., Dou, L.M., Wang, G.F.: Rockburst occurrences and microseismicity in a longwall panel experiencing frequent rockbursts. Geosci. J. **22**(4), 623–639 (2018). https://doi.org/10.1007/S12303-017-0076-7
58. Frontera-Pons, J., Veganzones, M.A., Pascal, F., Ovarlez, J.P.: Hyperspectral Anomaly Detectors Using Robust Estimators. IEEE J. Sel. Top. Appl. Earth Obs. Remote Sens. **9**(2), 720–731 (2016). https://doi.org/10.1109/JSTARS.2015.2453014
59. The SciPy community. scipy.stats.shapiro — SciPy v1.9.1 Manual (2023). URL https://docs.scipy.org/doc/scipy/reference/generated/scipy.stats.shapiro.html. Accessed 10 Feb 2023
60. Liu, F.T., Ting, K.M., Zhou, Z.H.: *Proceedings IEEE International Conference on Data Mining, ICDM* (IEEE), pp. 413–422 (2008). https://doi.org/10.1109/ICDM.2008.17
61. Mohamed, D., El-Kilany, A., Mokhtar, H.M.: A Hybrid Model for Documents Representation. Int. J. Adv. Comput. Sci Appl **12**(3), 317–324 (2021). https://doi.org/10.14569/IJACSA.2021.0120339
62. Amodei, D., Olah, C., Brain, G., Steinhardt, J., Christiano, P., Schulman, J., Dan, O., Google Brain, M.: Concrete Problems in AI Safety. ArXiv **1606.06565** (2016). URL https://arxiv.org/abs/1606.06565v2
63. ISO/IEC, ISO/IEC TR5469:202x(E) Artificial Intelligence Functional safety and AI systems. Tech. rep., International Electrotechnical Comission (2022). URL https://www.iso.org/standard/81283.html

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Rialda Spahic is a Ph.D. candidate at the Norwegian University of Science and Technology, Department of Engineering Cybernetics. Her background is in data analysis and system development, with a research focus on machine learning, the reliability of artificial intelligence, and autonomous systems. She received bachelor's and master's degrees in computer science and engineering from the International University of Sarajevo, Bosnia and Herzegovina.

Vidar Hepsø is an Adjunct Professor in Digitalization and oil and gas transformation at the Norwegian University of Science and Technology (NTNU), Department of Petroleum Engineering and Applied Geophysics, and an R&D Project Manager working with socio-technical implications of Remote Operations, electrification, and digitalization in Oil and Gas and Wind operations at Equinor. He received his Ph.D. in Anthropology of Science and at NTNU.

Mary Ann Lundteigen is a professor in the area (and the intersection between) instrumentation systems and functional safety at the of Engineering Cybernetics at the Norwegian University of Science and Technology (NTNU). Her background combines industrial and academic experience in safety instrumented systems, reliability analysis, and the digitalization of automation systems. She has an MSc in Engineering Cybernetics from the Norwegian Institute of Technology (NTH) (1993) and received her Ph.D. in Reliability, Availability, Maintainability, and Safety at NTNU.