POLITECNICO DI TORINO
Repository ISTITUZIONALE

Decision support impact and error prediction in control room: A behavioural data analysis

Original
Decision support impact and error prediction in control room: A behavioural data analysis / Amazu, C.W., Mietkiewicz, J., Abbas, A.N., Baldissone, G., Fissore, D., Demichela, M., Madsen, A.L., Leva, M.C.. - In: RELIABILITY ENGINEERING & SYSTEM SAFETY. - ISSN 0951-8320. - STAMPA. - 272 - 2:(2026), pp. 1-22. [10.1016/j.ress.2026.112616]

Availability:
This version is available at: 11583/3016138 since: 2026-09-30T09:01:54Z

Publisher:
Elsevier

Published
DOI:10.1016/j.ress.2026.112616

Terms of use:

This article is made available under terms and conditions as specified in the corresponding bibliographic description in the repository

Publisher copyright

(Article begins on next page)

18 November 2026

Reliability Engineering and System Safety 272 (2026) 112616



Contents lists available at ScienceDirect

Reliability Engineering and System Safety

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



Decision support impact and error prediction in control room: A behavioural data analysis

Chidera W. Amazu [a,b], Joseph Mietkiewicz [b,c,*], Ammar N. Abbas [a,b], Gabriele Baldissone [a], Davide Fissore [a], Micaela Demichela [a], Anders L. Madsen [c,d], Maria Chiara Leva [b]

[a] Politecnico di Torino, Turin, Italy
[b] Technological University Dublin, Dublin, Ireland
[c] Hugin Expert A/S, Aalborg, Denmark
[d] Aalborg University, Aalborg, Denmark

ARTICLE INFO

Keywords:
Operational data
Bayesian network
Logistic regression
Decision support
Control rooms
Alarm handling
Human factors

ABSTRACT

Data from psychophysiological measures can offer new insight into control room operators' behaviour, cognition, and mental workload status. This is particularly helpful when assessing capacity to respond to critical plant conditions such as alarm response scenarios. However, wearable tools such as eye tracking and electroencephalography caps can be perceived as intrusive and unsuitable for daily operations. Therefore, this article examines the potential of using real-time data from process and operator-system interactions during abnormal scenarios, recorded and retrieved from the distributed control system's historian or process log, to provide insight into operator behaviour and predict their response outcomes without intruding on daily tasks. Data for this study were obtained from a design of experiment using a formaldehyde production plant simulator and four human-in-the-loop support configurations. A comparison between configurations in terms of both behaviour and performance is presented. Then, a step-wise logistic regression and a Bayesian network model were used to predict operator error. The results identified predictive metrics, discussed in terms of their value as precursors of overall system performance in alarm response scenarios. Knowledge of relevant and predictive behavioural metrics accessible in real time can better equip decision-makers to predict outcomes and provide timely support measures for operators.

1. Introduction

1.1. Background and motivation

Operators in modern process control room configurations play a crucial role in preventing accidents and managing unexpected events in contemporary process plants. Typically, operators in these settings monitor plant processes, manage alarms, diagnose issues, and execute appropriate control actions. However, they face several challenges that hinder their ability to perform these roles effectively. One key challenge arises from the tools designed to support their tasks. More specifically, tools such as alarm systems, procedures and displays, which are intended to enhance their situational awareness, decision making, and ability to manage future incidents, may not always provide adequate support [1,2].

In addition, in increasingly automated plants, operators face extended monitoring periods during which automation assumes their active role, leaving them out of the loop [3]. When automated systems subsequently fail, these disengaged operators must suddenly respond to critical alarms, a transition that, without adequate support, can have catastrophic consequences in safety-critical environments. These issues underscore the necessity of adopting a predictive approach to anticipate factors that could impede correct alarm responses by operators and to develop strategies to manage such situations. Predictive insights can be achieved using machine learning or deep learning techniques. However, a fundamental step is to identify the variables required for these predictions, which accurately describe the operator-system interaction.

Several performance shaping factors have been recognised as influential in evaluating the operator-system interaction in process control, most of which are based on traditional human reliability analysis

* Corresponding author.
E-mail address: joseph.mietkiewicz@tudublin.ie (J. Mietkiewicz).

https://doi.org/10.1016/j.ress.2026.112616

Available online 19 March 2026
0951-8320/© 2026 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (http://creativecommons.org/licenses/by/4.0/).

methods, such as SPAR-H [4]. Some researchers have also identified psycho-physiological measures that describe these interactions, exploring their use in predicting operators' workload levels [5], situational awareness, or comparing different support options. However, these approaches have limitations. Traditional human reliability analysis methods are typically applied either prospectively during design phases or retrospectively after incidents have occurred, and while they can incorporate context-specific factors, many HRA techniques rely on expert judgment with insufficient empirical validation and abstract contextual information rather than real-time, objective operational data. On the other hand, psycho-physiological measures, such as those obtained by eye tracking, pose problems such as intrusiveness of the device, lack of applicability in real-time, and the need for controlled setups [6]. Therefore, the goal is to identify real-time measures that describe the operator-system interaction and can help predict an operator's ability to respond correctly to alarms. This would enable the provision of timely, real-time recommendations to operators, enhancing safety and efficiency in process control environments.

Operational data, specifically those that describe the human system interaction, such as information on the acknowledgement of alarms, response times, number of alarms, etc., are useful to decision makers as they have the potential to provide insights into metrics for understanding operator behaviour, workload, situational awareness, and errors [6–8]. They provide practical and unbiased information. The research of Braarud and Chen specifically highlighted the value of real-time operator-system interaction data for classifying the operator workload [6,8]. Their studies identified important features including the number of alarms per station, human-system interaction (HSI) activities (navigating process images, selecting objects, and entering values), and speech characteristics.

In process plants, obtaining the necessary operational data presents several challenges. For example, recovery from safety or load-related alarms is typically not automated due to the wide variety of potential causes. These manual interventions are logged in audit trails and stored in historians or databases, yet they are rarely analysed, resulting in a blind spot in digitisation and Industry 4.0 initiatives [9]. To address this, human-system interaction data from historians (specialised time-series databases that store operational data from distributed control systems) should be examined; however, little research has explored these data for error prediction or as potential indicators of process safety and control performance [10].

This gap may stem from the limitations in how historian data are recorded and their original intended purpose. Even in cases where such data are available, as demonstrated in a previous analysis of a gas exploration and production (E & P) company, the methods used to record the data can pose significant challenges during analysis, making it difficult to extract meaningful behavioural metrics.

1.2. Related work and research gaps

1.2.1. Simulator-based frameworks and advanced measurement approaches

Building on the limitations of traditional HRA methods noted above, recent frameworks have emerged for systematic empirical data collection from simulators. Jung et al. [11] developed the HuREX framework using nuclear power plant full-scope simulators with licensed operators, generating 37 human error probabilities (HEPs) across 21 generic task types. Park et al. [12] introduced the SHEEP framework using simplified simulators with student participants to complement full-scope studies. Both frameworks systematically collect data to quantify relationships between performance shaping factors (PSFs), such as crew experience, task complexity, and procedure quality, and error likelihood.

However, these frameworks calculate generic, population-level HEPs for task types (e.g., "probability of omitting a valve isolation step") through post-hoc analysis of simulator recordings for use in probabilistic safety assessment models. The outcome variable is binary error occurrence on predefined procedural steps, with PSFs used as categorical grouping variables to stratify HEPs. While valuable for prospective risk modelling, this approach does not extract real-time behavioural metrics, such as alarm acknowledgement patterns, navigation sequences, or response timing, that could enable the prediction of task-specific, individual operator outcomes during ongoing scenarios.

Beyond the eye-tracking approaches mentioned previously [6], Srinivasan et al. [13] demonstrated that fusing eye-tracking and EEG data enables accurate classification of operators' cognitive workload during abnormal situation management, while Iqbal et al. [14] developed EEG-based methodology for dynamic workload assessment. These multimodal physiological approaches achieve high accuracy but remain impractical for routine industrial deployment [15]. Extending the work of Braarud et al. on HSI activities for workload classification, Schmidt et al. [16] specifically examined manual intervention patterns in process plants, confirming that while these data are systematically logged in historians, they remain largely unexploited for performance assessment, the "blind spot" in digitisation initiatives mentioned earlier [9].

1.2.2. Machine learning for human reliability prediction

The application of machine learning to HRA represents an emerging frontier for data-driven methods. Garg et al. [17] compared Logistic Regression, Bayesian Networks, and Neural Networks for predicting HEPs from performance shaping factors (PSFs) using SACADA and HuREX data. Groth et al. [18] proposed hybrid algorithms combining Bayesian Networks with simulator data and cognitive models. Morais et al. [19] developed NLP techniques to automatically classify accident reports, reducing expert analysis time from days to minutes. Liu et al. [20,21] advanced Bayesian network approaches by modelling dependencies among PSFs, addressing the limitation that traditional HRA treats PSFs as independent multipliers. Their frameworks use Exploratory Factor Analysis to cluster dependent PSFs and quantify interactions through Monte Carlo sampling. However, these approaches rely on subjective PSF assessments (e.g., expert ratings of stress level" or procedure quality") rather than objective behavioural observations.

Kim and Park [22], demonstrated the limitations of traditional maximum likelihood estimation for PSF modelling from simulator data, proposing a Bayesian logistic regression approach that incorporates prior knowledge from established HRA methods such as SPAR-H and CREAM. Their results show that integrating expert prior knowledge with empirical data produces more stable PSF multiplier estimates, particularly when sample sizes are limited, further supporting the value of the Bayesian approaches for human reliability modelling in data-constrained settings.

1.2.3. Decision support systems in control rooms

While the introduction highlighted challenges with inadequate support tools [1,2], recent research has examined specific decision support interventions. Dorgo et al. [23] demonstrated that ML-based alarm rationalisation achieves 50% reduction in total alarms, explicitly linking alarm overload to major incidents. Qing et al. [24] found that computer-based procedures reduce workload but significantly decrease situation awareness, particularly perception and projection levels, compared to paper-based procedures. However, these studies examine support tools individually without systematic comparison across configurations and complexity levels.

1.3. Research gaps and questions

Building on this foundation, three critical gaps motivate the present research:

1. **Accessibility for Industrial Implementation**: While physiological approaches provide objective measurement [13,14], and simulator frameworks enable systematic data collection [11,12], the field lacks

a validated methodology for extracting predictive behavioural metrics from DCS historians, data that exists in every modern facility without additional instrumentation.
2. **Real-Time Predictive Capability**: Existing ML approaches for HRA [17–21] calculate generic error probabilities retrospectively. Control room supervisors require task-specific, real-time prediction of imminent success or failure to enable timely intervention during scenarios.
3. **Comparative Decision Support Evidence**: Organizations implementing Industry 4.0 initiatives need quantitative evidence on which support tools provide value under what conditions [23,24]. Current research evaluates tools individually, lacking a systematic comparison across cumulative configurations and complexity levels.

This study addresses these gaps by: (1) identifying behavioural metrics accessible from DCS operational data without intrusive sensors, (2) developing real-time predictive models (Logistic Regression and Bayesian Networks) for task-specific alarm response outcomes, and (3) systematically comparing four support configurations (baseline, alarm rationalisation, digital procedures, AI decision support) across three complexity levels with 92 participants. The findings demonstrate which behavioural metrics best predict operator success, which support tools are most effective under varying conditions, and how practitioners can leverage existing DCS infrastructure for continuous operator performance assessment.

Against this background, this article aims to address the following questions:

1. Is it possible to gain useful insight on the dynamic of human machine interaction during alarm response conditions that can be used to predict the probability of a successful resolution for the alarmed state?
2. What operator behavioural metrics are the best predictors for the successful recovery of a plant-alarmed state?
3. How do alarm rationalisation, digital (on-screen) procedures, and AI-based decision support systems (DSS) impact operator performance in managing plant-alarmed conditions across varying levels of task complexity?

1.4. Contributions and paper organisation

This paper briefly presents the performance and behavioural outcomes from an experimental study that compare four HITL configurations using the operational data derived by running a Design of Experiment involving 92 participants managing three alarm response scenarios using the Human Machine Interface (HMI) developed for the simulator of a formaldehyde production plant (REF). The analysis proceeds in two main phases.

- **Phase 1: Statistical Comparison of Support Configurations** (Section 5) evaluates how each support technology (alarm rationalisation, digital procedures, and AI decision support) impacts individual behavioural and performance variables through group-wise statistical comparisons across three scenarios. This addresses Research Question 3 by determining which support configurations improve specific operator outcomes.
- **Phase 2: Predictive Modelling of Operator Error** (Section 6) develops predictive models for operator error from the operational data collected. Two complementary approaches, stepwise Logistic Regression (LR) and Bayesian Networks (BNs), identify which behavioural metrics most strongly predict error probability (Research Questions 1 and 2). The models then estimate success rates for each support configuration, allowing comparison of their effectiveness in preventing operator errors across different complexity levels.

LR is used to provide an initial insight into the predictive performance of operator error. This approach is complemented by using BNs, a method selected for its interpretability and transparency through graphical representation of causal relationships, making it a particularly suitable choice for industrial applications where model explainability is crucial for stakeholder acceptance and regulatory compliance.

Initially, LR is used to identify and test the significance of various predictor variables. LR is a statistical method well-suited for binary classification tasks. It provides clear and interpretable results that show the impact of each predictor variable on the outcome. The simplicity and efficiency of LR make it a starting point for the analysis, especially when dealing with large datasets. Using stepwise logistic regression, we iteratively select the most significant variables, ensuring that the final model is both streamlined and robust. This step is crucial because it helps to understand the relative importance of different predictors and provides a foundation for further analysis. The method is then used to predict error with inputs from the behavioural data and subsequently with subjective measures.

Following the initial analysis with LR, BNs are used to model the complex dependencies among the identified significant variables. BNs are particularly effective in capturing and quantifying the conditional dependencies between variables, making them ideal for understanding the intricate interactions within operational data. Moreover, in addition to their interpretability, BNs offer a significant advantage when it comes to handling missing data, a common challenge in real-world applications. Their probabilistic framework allows for inference even with incomplete datasets, enabling predictions based on available information. This flexibility makes BNs well-suited for online deployment in control rooms, where real-time assessment of error likelihood is crucial. The model can adapt to varying levels of data completeness, providing continuous risk assessment and potentially enhancing operator performance and system safety.

2. Material and methods

2.1. Formaldehyde simulator

The case study examines the production of formaldehyde through the partial oxidation of methanol with air. A secondary reaction, which fully oxidises methanol to carbon monoxide, reduces the formaldehyde yield. This study and simulator, originally developed by [25] at Politecnico di Torino, Italy, and later adapted by the authors, details a plant consisting of six main sections and has a total of eighty different alarms. Each plant section can be accessed as a detailed mimic display by clicking the corresponding button at the bottom of the main screen (Fig. 1). For example, Fig. 2 shows the tank section mimic. Each mimic contains: (1) control interfaces (e.g., nitrogen flow adjustment and nitrogen system selection for the tank section), (2) a process flow diagram for that specific section displayed centrally with all associated alarms, and (3) key monitoring graphs (e.g., pressure trends for the tank section). The simulator is not publicly available.

As shown in Fig. 1, the plant layout begins with the Tank Section, where a methanol storage tank, equipped with alarms for parameter deviations, is located. The Methanol Section follows, where liquid methanol is converted to gas via a pump and heater before mixing with compressed gas from the Compressor Section. This mixture then passes through a heat exchanger (REC2), heating it to approximately 200°C before entering the reactor. The Reactor Section contains the reactor, and the heat exchanger section includes the heat exchangers REC1, REC2, and REC3. REC1 and REC3 cool the reactor's output, and the heat recovered from REC1 and REC3 is used for water boiling and cooling, respectively.

The reactor, located in the Reactor Section, works in conjunction with the heat exchangers in the Heat Exchanger Section. REC1 and REC3 cool the reactor's product, with the recovered heat used to boil and cool water. REC3 specifically cools the product to around 67°C before it enters the absorber, where it is absorbed by water flowing countercurrently. This absorber is located in the Absorber Section.

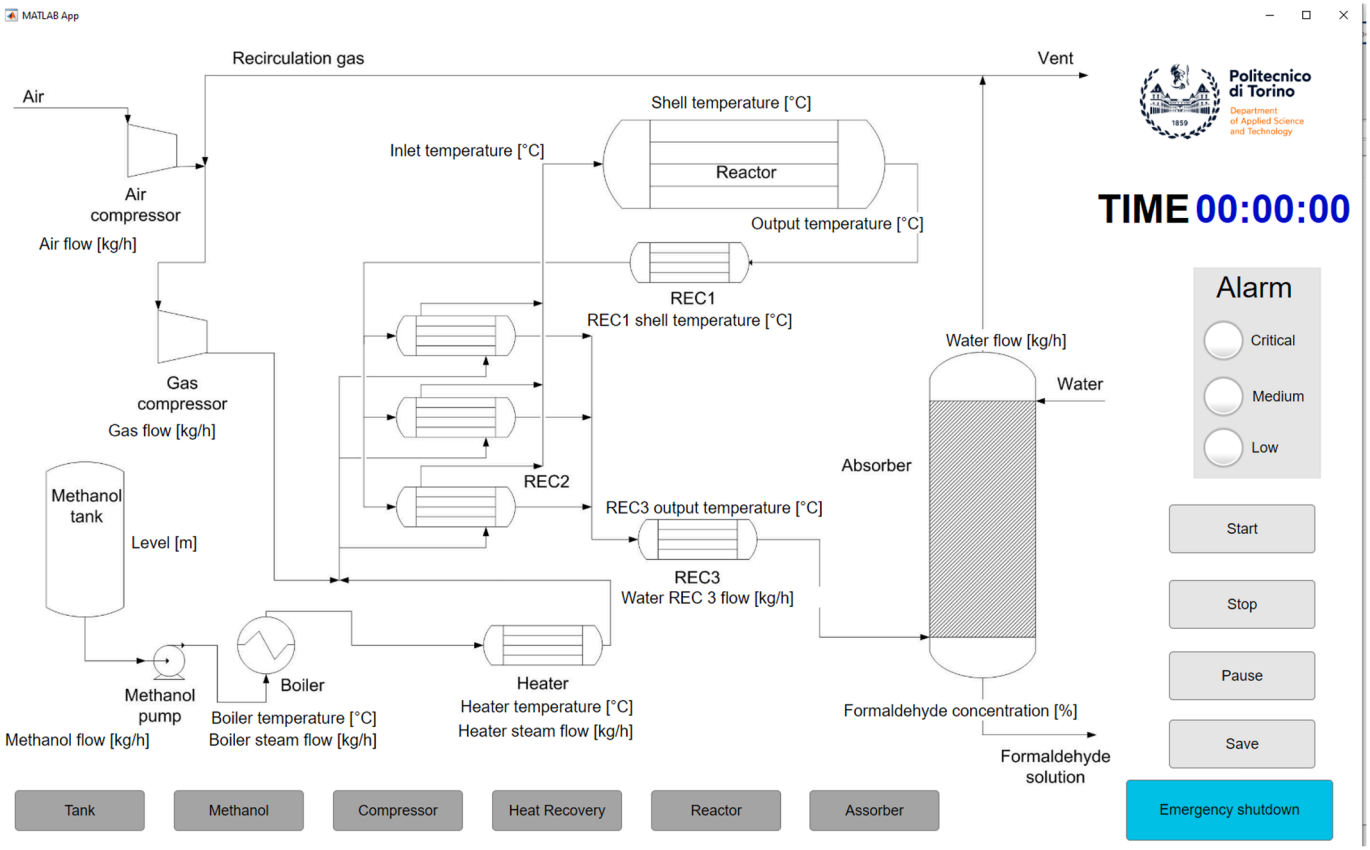


Fig. 1. Main screen of the simulator [26].

Several potential accident scenarios could occur in this plant, such as overheating of the reactor or absorber if the temperatures from the mixture or product exceed expectations. Attempting to control this temperature by increasing the water input into the absorber could result in a loss of the volume of the target product.

2.2. Participants

The study was designed to capture conditions and process events similar to those in a formaldehyde production plant as closely as possible. Ninety-two participants voluntarily participated in this study [26]. They consisted mainly of chemical engineering master's students and some staff familiar with the plant process. Participants were recruited over a two-month period, and group assignment was performed iteratively to maintain equal distribution across the four groups, reaching 23 participants per group at the end. Additionally, to account for potential fatigue effects, the timing of experimental sessions (morning vs. afternoon) was balanced across groups, as participants were students with other ongoing coursework that could influence their performance. Their mean age was 25 years (standard deviation = 5.4), ranging from 21 to 61 years, with training and experience in Chemical Engineering and, in some cases, knowledge of control room operations. Participants rated their level of expertise in both chemical process and control room operations prior to the test.

Each group had a unique configuration to explore the effectiveness of different technological tools. The four-group experimental design was specifically structured to assess the incremental impact of each support technology on operator performance, directly addressing Research Question 3. Each group represents a progressive enhancement of operator support. The characteristics of each group are detailed below and summarised in Table 1.

Table 1
Characteristics of the Groups/type of support.

	Alarmrationalisation	ProcedureOn Screen	AISupport
G1	no	no	no
G2	yes	no	no
G3	yes	yes	no
G4	yes	yes	yes

1. **Group 1 - Baseline without Alarm Rationalisation**: This group operated under standard conditions without any alarm rationalisation system, serving as the control group.
2. **Group 2 - Introduction of Alarm Rationalisation**: Equipped with an alarm rationalisation system, this group aimed to assess the impact of filtering out non-critical alarms on reducing cognitive load and improving focus on critical issues.
3. **Group 3 - Transition to On-Screen Procedures**: This group shifted from paper-based methods to an on-screen interface for accessing procedures. The objective was to determine if digital access could enhance operational efficiency and response times compared to traditional methods.
4. **Group 4 - Integration of AI DSS**: Representing the most advanced level of support, this group incorporated an AI DSS to measure its incremental benefits in improving performance, reducing errors, and enhancing overall safety and efficiency. Details of the DSS development can be seen in [27].

The study was reviewed and approved by the Technological University Ethics Review Committee and the Collaborative Intelligence for Critical Safety Systems (CISC) internal ethics committee and was performed with appropriate consent obtained from participants.

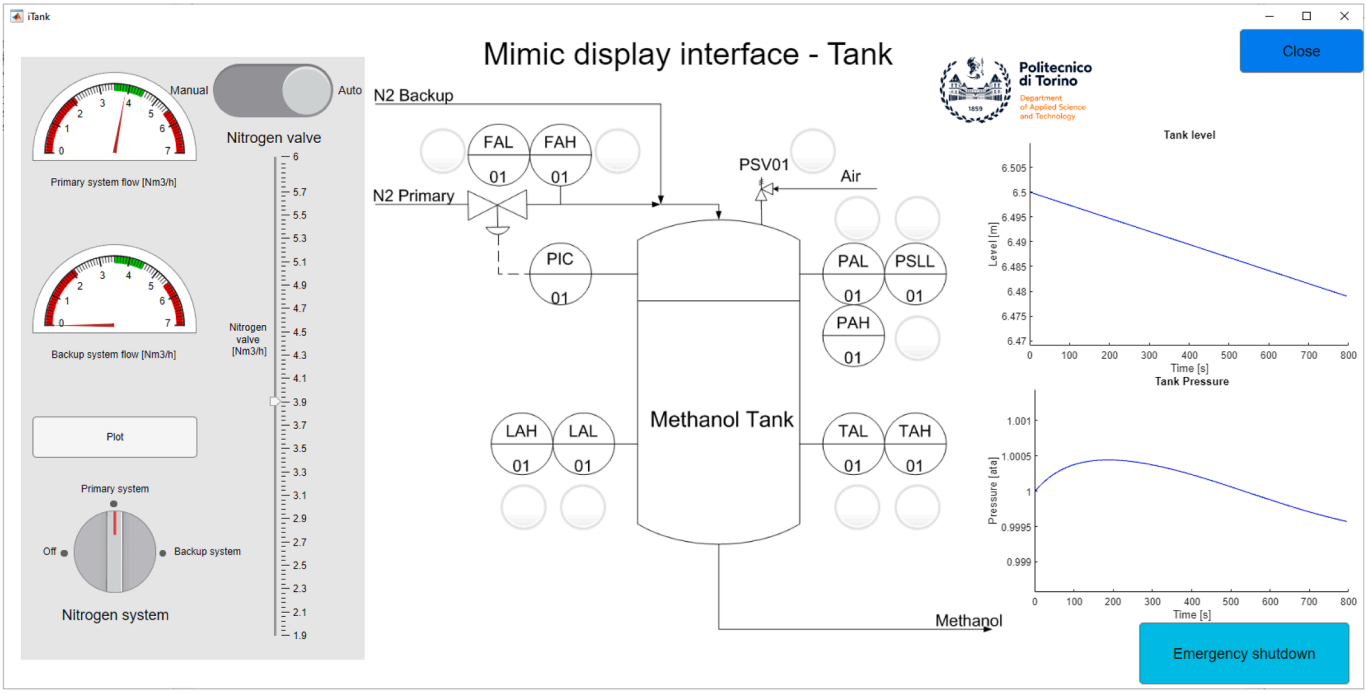


Fig. 2. Tank mimic of the simulator [26].

2.3. Scenarios/task complexity

We designed three scenarios to evaluate different human-in-the-loop (HITL) configurations in a simulated control room environment. These scenarios aimed to test the robustness of the system and the operator's ability to respond effectively under varying conditions. Each scenario simulates a specific failure or challenge that an operator might encounter in a plant environment.

1. **Pressure Indicator Control Failure (Scenario 1)**: The automatic pressure management system in the tank fails, requiring the operator to manually adjust the nitrogen inflow to maintain pressure. The scenario becomes more challenging when the nitrogen flow stops, leading to a pressure drop as methanol continues to be pumped into the plant.
2. **Nitrogen Valve Primary Source Failure (Scenario 2)**: The primary nitrogen supply system fails, and the operator must switch to a backup system. The backup system starts slowly, necessitating careful regulation of pump power to maintain tank pressure.
3. **Temperature Indicator Control Failure in the Heat Recovery Section (Scenario 3)**: Initially, the operator tries to solve the problem by manually adjusting the cooling water flow set point in the absorber. When this fails, the operator consults the supervisor for advice. The supervisor indicates that the issue cannot be resolved from the control room and requires a field operator's intervention. Meanwhile, the control room operator must manage the reactor's temperature until the end of the scenario (18 min). The main challenge in this scenario is preventing the reactor from overheating, necessitating close monitoring and adjustment of the reactor's cooling water temperature. This scenario corresponds to a more challenging situation with alarm overflow.

Each scenario was designed to evaluate the effectiveness of the DSS in aiding operators during complex and high-pressure situations. These scenarios, as summarised in Table 2, provide a comprehensive test of the system's utility in enhancing operator performance and decision-making under various challenging conditions.

3. Metrics extraction from simulator logs

From the raw simulator logs collected during the experiment, we systematically extracted two categories of metrics to characterise operator performance and behaviour across the 92 participants and three scenarios. Performance-based metrics quantify task outcomes (e.g., recovery time, accuracy, consequences), while behavioural metrics capture observable operator-system interactions during task execution (e.g., alarm acknowledgement patterns, mimic navigation, procedure consultation). These measures form the basis for both the comparative analysis of support configurations (Section 5) and the development of predictive models for operator error (Section 6).

Table 3 provides comprehensive definitions for all extracted metrics. An important feature of this extraction approach is scenario-specific adaptation: Scenarios 1 and 2 share similar tank-related metrics reflecting their common focus on pressure management (pressure recovery, nitrogen valve adjustment), while Scenario 3 employs distinct reactor-focused metrics appropriate for temperature control tasks (survival time rather than recovery time, reactor temperature management). This adaptation ensures that metrics meaningfully capture operator performance for each fault type's specific operational demands.

The following subsections detail the extraction methodology for performance-based metrics and behavioural metrics. Explicit references to raw simulator log variables are provided throughout to ensure transparency and reproducibility, enabling future studies to replicate or extend the extraction procedure using the same similar and/or data.

3.1. Performance and system

Performance-based metrics obtained in this study include recovery time, accuracy, overall performance, and consequence. The number of alarms activated is a system feature, but also tells the outcomes of operators' actions - correct or incorrect. Table 3 shows the definition of these measures for each scenario in the experimental study. As observed, the extraction of some of the metrics for scenario 2 was similar to scenario 1 and varied with scenario 3, given the difference between the scenario goals. This was because scenario 3 was related to an entirely different

Table 2
Scenario characteristics.

	Scenario 1	Scenario 2	Scenario 3
Description	Pressure IndicatorControl Failure	Nitrogen Valve Primary SourceFailure	Temperature Indicator Control Failure
Complexity Level	Low	Medium	High
Alarm Flood	No	No	Yes

section of the plant, the reactor and absorber sections, while scenarios 1 and 2 were related to the Tank section.

Recovery time for scenarios 1 and 2 is extracted based on the feature '$All2_1$' (Pressure Alarm Low). In Scenario 3, the recovery time does not count as recovery is meant for the field operator and not the control room operator, given that an issue on the heat recovery (REC3) cooling water system was intentionally induced and can only be fixed on the field (see Fig. 1). The measure was derived from the raw data variable 'TMAXREACTORE', which is the reactor temperature. Survival time, however, is used in place of response time for scenario 3. It means the operators' ability to maintain a no-shutdown or reactor overheating situation until the end of the test at the 18th min. The assumption was that at this 18th min, the field operator should have resolved the issue with the REC3 cooling system.

Accuracy was derived based on the variables '$FN2serb1O_1$' (Primary Nitrogen Flow), '$MWpopOld_1$' (Manual Pump power value), and '$Toutrec2c_1$' (Temperature Out REC1) for scenarios 1, 2, and 3, respectively. It refers to the mean squared error (mse) between their adjusted value and the mean of the range as written in the operating or intervention procedure.

In deriving the overall performance class, the individual group distributions of recovery time was considered. Those classified as 'optimal' fell within the 25th percentile of the study population, 'Good' below the 50th percentile and others 'poor'. The consequence is derived from the raw features ' $PSERB_1$', '$TMAXREATORE_1$' and '$emmerO_1$', leading to one of the consequences; Safe - rank 1, Impurity of air in the tank atmosphere - rank 2, Air in the tank atmosphere - rank 3, Shutdown - rank 4, Implosion - rank 5 (for scenarios 1 and 2 only), Reactor Overheated - rank 5 (for scenario 3 only).

Finally, the number of activated alarms was recorded by summing all alarm activations during the scenario. Each alarm in the plant was recorded with a feature starting with 'All + the number', for example, $All2_1$ for pressure alarm low (PAL01). When an alarm is activated the value changes from 0 to 1. The feature extraction code checks for every occurrence of 1 during the scenario, also taking into account any other activation after a previous deactivation and sums this up.

A description of other features is shown in Table 3. An example plot showing the extraction of the reaction time and response time for scenario 1 using one of the participant's outcomes is shown in Figs. 3 and 4, respectively.

3.2. Behavioural feature extraction

The metrics counted as behavioural in this study include reaction time, response time, no procedures opened, no mimics, silenced alarms, and acknowledgement of alarms. These show the operators' behaviour during the task, which is influenced by different underlying factors. These metrics reflect operator situational awareness during the task. Excessive mimic navigation beyond task requirements, high alarm acknowledgement counts, or unnecessary procedure consultation are indicative of poor critical alarm perception. Similar operator-system interaction metrics have been used by [6] in classifying workload. The reaction time in 3 is the time it takes to switch the Nitrogen valve button from Auto to Manual, or in scenario 3, to switch the REC3 cooling water flow system from manual to automatic. It was calculated based on the raw feature '$MNitsel_1$' (Manual Nitrogen Valve Switch) for scenarios 1 and 2, and $REC3WMO_1$ (Heat Recovery 3 cooling water system Switch) for scenario 3.

Response time, Fig. 4, is the time it takes to adjust the nitrogen valve scale to the correct value. For Scenario 2, it is the time it takes to switch from primary to backup and the time of the last activity before the temperature spikes for Scenario 3. It is based on the features '$MmanNit_1$' (Manual Nitrogen Valve), '$Nitsel_1$' (Primary to Backup System Control and '$Toutrec2c_1$' (Reactor Cooling System) for scenarios 1, 2, and 3, respectively. The number of procedures is based on the variable '$proclis_1$' (procedures). The record on the number of procedures was only possible for Group 3 and Group 4 because of the ease with recording the operator system interaction on the display for these groups. The number of mimics/display opened, number of alarms silenced and no of alarms acknowledged were recorded for all groups based on the features '*intop*', '*sAll*', and '*AckAll*', respectively. The detailed description of the simulator log variables from which the metrics were derived is published in [28]

4. Predictive modelling methods

The following algorithms and data processing are used to build a human reliability model for each group and scenario.

4.1. Step-wise logistic regression analysis

Step-wise LR is a method used for building a predictive model by selecting the most significant variables incrementally until certain criteria, like the model fit or statistical significance, are met. This approach is particularly advantageous when dealing with a large set of potential predictors, as it helps to identify the most relevant ones while excluding those that do not contribute significantly to the model's predictive power.

The logistic regression model predicts the probability of a binary outcome based on one or more predictor variables. The logistic function used in this model is defined as:

$$P(Y = 1 \mid X) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 X_1 + \beta_2 X_2 + \ldots + \beta_n X_n)}}$$

where Y is the binary dependent variable (e.g., error occurrence), $X_1, X_2, \ldots, X_n$ are the independent variables (predictors), and $\beta_0, \beta_1, \ldots, \beta_n$ are the coefficients of the model.

The step-wise approach involves the following steps:

1. **Initialisation**: Start with an empty model or a model with a single predictor.
2. **Forward Selection**: Add predictors to the model one by one. At each step, add the predictor that improves the model the most, usually based on the Akaike Information Criterion (AIC) or Bayesian Information Criterion (BIC).
3. **Backwards Elimination**: After adding a new predictor, remove any predictor that does not significantly improve the model. This is also done based on AIC or BIC values.
4. **Iteration**: Repeat steps 2 and 3 until no further predictors can be added or removed based on the chosen criterion.

The goal of step-wise logistic regression is to achieve a balance between model complexity and predictive accuracy, thereby avoiding overfitting by including only those predictors that significantly contribute to the model (as observed using the p-values) and to the accuracy score. In our study, we applied step-wise logistic regression to

Table 3
Behavioural, system and performance metrics extracted from simulator logs during the experiment; scenario 1, 2 and 3.

No.	Metrics	S1	S2	S3
1	Recovery Time (performance)	The time it takes to recover the pressure alarm PAL01. This time is calculated from the initiation of the fault at 60 seconds after the start of the experiment until the time PAL01 is cleared. Same for everyone.	Same as S1.	For scenario 3, the recovery time is not considered since the operator has to maintain the temperature of the reactor stable until the end of the scenario.
2	Reaction Time (Behavioural)	The time it takes to switch the Nitrogen valve button from Auto to Manual. This time is calculated from the initiation of the fault at 60 seconds after the start of the experiment until the time 'manual' is activated.	Same as S1.	We have in this case the time it takes to switch the cooling water system of FAL11 (REC3WMO_1) from Auto to Manual.
3	Response Time (Behavioural)	The time it takes to adjust the nitrogen valve scale to the correct value. This time is calculated from the initiation of the fault at 60 seconds after the start of the experiment until the last adjustment action on the nitrogen valve scale.	The time it takes to switch from primary to backup. If not switched at all then label 'error'. This time is calculated from the initiation of the fault at 60 seconds after the start of the experiment.	It is the time of last activity before the spike of trend in the temperature cooling system (Toutrec2c_1). In scenario 3 it is rather referred to as the survival time (expected to survive up to 1080 seconds, which is the task duration).
4	Accuracy (performance)	refers to the difference between their adjusted value on the nitrogen valve scale and the mean of the range given in the procedure (Mean Squared Error).	The difference between their adjusted value on the pump power valve scale (MWpopOld_1) and the mean of the range given in the procedure (MSE).	The difference between their adjusted value on the temperature cooling system scale (Toutrec2c_1) and the mean of the range given in the procedure (MSE).
5	Overall performance (performance)	Optimal: Recovery time $\leq$ (25th percentile). Good: 25th percentile <Recovery time $\leq$ (50th percentile). Poor: Recovery time >50th percentile.	Same as S1	Optimal: Recovery time $\geq$ (900) and the consequence is 'Safe'. Good: 850 $\leq$ (Recovery time) < 900. Poor: All other cases
6	Consequence (performance)	• If PSV01 is activated, then Consequence == Safe – (1) • If PSV01 activated, that is Pressure <0.98000) and Pressure >= 0.975, then Consequence == Impurity of air in the tank atmosphere – (2) Else Consequence == Air in the tank atmosphere – (3) • If PSLL, that is Pressure <= 0.97000 activated, then Consequence == Shutdown – (4) • OR If 'emmerO_1, that is Emergency Shutdown == 1, then Consequence == Shutdown – (4) • If all systems failed, that is Pressure <= 0.97000) and 'emmerO_1' == 1 then Consequence == Implosion – (5)	Same as S1.	If TMAXREATORE_1, that is Reactor Temperature >= 400 then Consequence == Reactor Overheated - Score: (5)
7	Alarms silenced (Behavioural)	No. of alarms silenced during the scenario.	Same as S1.	Same as S1.
8	Acknowledgement (Behavioural)	No. of alarms acknowledged during the scenario.	Same as S1.	Same as S1.
9	Mimics opened (Behavioural)	No. of mimics opened throughout the scenario.	Same as S1.	Same as S1.
10	No. of procedures (Behavioural)	No. of procedures opened throughout the scenario.	Same as S1.	Same as S1.
11	No. of alarms (System)	No. of alarms activated.	Same as S1.	Same as S1.
12	Error (Performance)	If at the end of the scenario, the pressure is increasing or has reached a normal level and shutdown is not reached	Same as S1.	If the high high temperature alarm is reached in the reactor.

identify key factors influencing the likelihood of human error. By iteratively adding and removing variables, we derived a parsimonious model that effectively predicts error occurrence based on these behavioural metrics.

Prior to model development, categorical variables required appropriate encoding to ensure meaningful analysis. The Group and Scenario variables were encoded as ordinal (1–4 for Group; 1–3 for Scenario). This ordinal encoding is justified by the inherent progression in both variables: Group represents the cumulative addition of support features, while Scenario reflects progressive task complexity. This encoding allows the model to capture the directional relationship between increasing support/complexity and error probability.

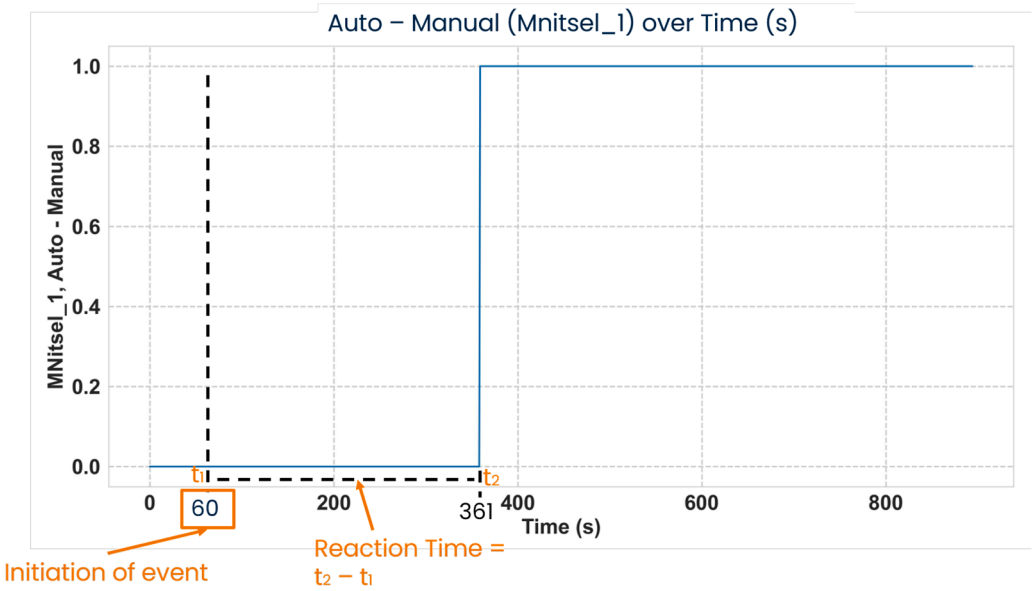


Fig. 3. Reaction time example from participant 10, scenario 1.

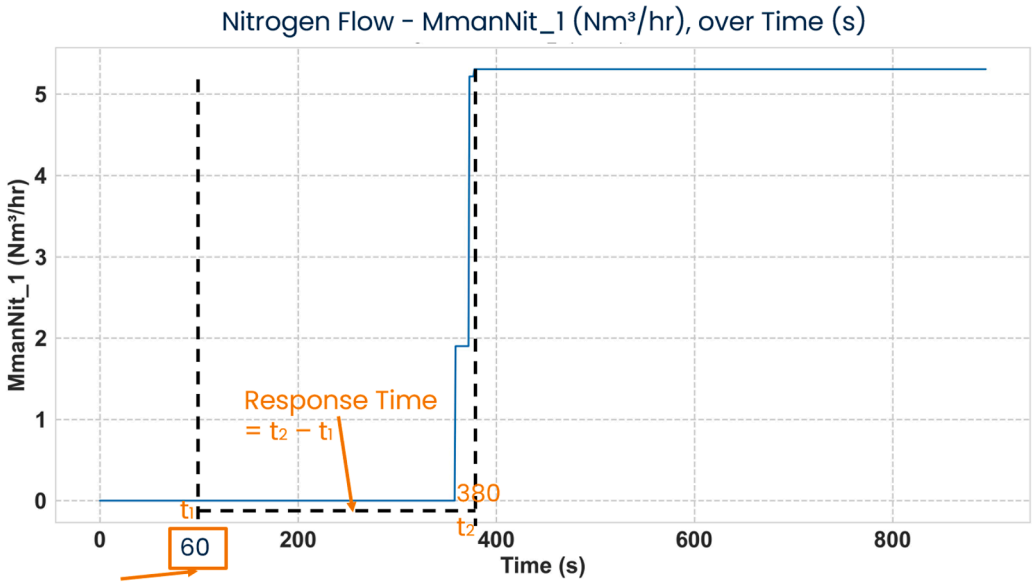


Fig. 4. Response Time example from participant 10, scenario 1.

The results from the step-wise logistic regression analysis provide insights into the relative importance of different predictors and help to refine our understanding of the factors contributing to human errors in control room environments.

4.2. Bayesian networks

In this study, we employ Naive Bayes (NB) classifiers and Tree-Augmented Naive Bayes (TAN) models to predict operational outcomes. These models are valuable tools for understanding and modelling the complexities of operational data.

Bayesian networks. A BN is a probabilistic graphical model that represents a set of variables and their conditional dependencies through a directed acyclic graph (DAG). In this graph, nodes represent random variables, while directed edges indicate probabilistic dependencies between these variables. This structure allows BNs to model (causal) interactions and predict events effectively. Each node is associated with a conditional probability distribution that quantifies the effect of the parent nodes on the node itself. This factorisation simplifies the joint probability distribution, making complex problems more manageable.

A discrete BN $\mathcal{N} = (\mathcal{V}, \mathcal{E}, \mathcal{P})$ is formally defined as follows [29]:

- $\mathcal{V} = \{v_1, v_2, \ldots, v_n\}$ is a set of nodes, each representing a discrete random variable X_v.
- $\mathcal{E}$ is a set of directed edges (arcs) between the nodes, forming a DAG. Each edge $(v_i, v_j) \in \mathcal{E}$ indicates a direct probabilistic influence from node v_i to node v_j.
- $\mathcal{P}$ is a set of conditional probability distributions, one for each variable X_v given its parents $X_{pa(v)}$ in the graph $\mathcal{G}$. Specifically, for each $v \in \mathcal{V}$, there is a conditional probability distribution $P(X_v \mid X_{pa(v)})$.

A BN encodes the joint probability distribution over a set of random variables $\mathcal{X} = \{X_1, X_2, \ldots, X_n\}$. The joint probability distribution $P(\mathcal{X})$ factors according to the structure of the DAG, leveraging the conditional

independencies implied by the graph. This factorisation is given by:

$$P(\mathcal{X}) = \prod_{v \in \mathcal{V}} P(X_v \mid X_{pa(v)}) \tag{1}$$

This equation illustrates that the joint probability distribution is decomposed into a product of conditional probabilities, each corresponding to a node given its parents in the DAG. This factorisation significantly reduces the complexity of representing the joint distribution, especially for large networks.

Naive bayes. The NB classifier is a simple yet powerful probabilistic model used for classification tasks. It assumes conditional independence between feature variables given the class variable C, which simplifies the computation of the joint probability distribution. Despite this simplification, the NB is often effective and is robust to noise [30]. Let $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ represent an instance to be classified, and C be the class variable. The NB classifier computes the probability of the instance as follows:

$$P(x_1, x_2, \ldots, x_n \mid C) = \prod_{i=1}^{n} P(x_i \mid C)$$

This independence assumption allows the model to scale linearly with the number of features, making it computationally efficient and suitable for large datasets.

Tree-augmented naive bayes. The TAN model enhances the NB classifier by introducing dependencies between feature variables. This method improves the model's expressiveness while retaining computational efficiency. The TAN model adds a tree structure over the feature variables, maintaining a balance between simplicity and capturing essential dependencies [30].

Let $\mathbf{X} = (X_1, X_2, \ldots, X_n)$ represent the set of feature variables, with X_1 as the root variable of the tree structure. The TAN model is built through the following steps:

- Calculate the conditional mutual information $I(X_i, X_j|C)$ for all pairs of feature variables.
- Construct a maximum-weight spanning tree T_S for the complete graph over $\mathbf{X} \setminus \{C\}$, using the mutual information as edge weights.
- Direct the edges in T_S by selecting a root in $\mathbf{X} \setminus \{X_1\}$ and orienting edges away from it.
- Construct the graph corresponding to a NB model with X_1 as the root of T_S and add the directed edges of T_S to it.

By incorporating these dependencies, TAN provides a more accurate representation of the relationships between variables compared to the standard NB classifier. Importantly, the directed edges in the TAN structure represent conditional probability factorisation for computational efficiency, not causal relationships; the tree structure is determined algorithmically to optimise predictive performance.

Entropy minimisation algorithm. In this study, we employed an entropy-based discretisation algorithm to categorise continuous variables into discrete intervals. This method, introduced by Fayyad and Irani [31], recursively partitions the range of a continuous attribute, selecting cut points that maximise the information gain ratio. The algorithm aims to find optimal thresholds that preserve the class-attribute interdependence while reducing the data's complexity. This approach is particularly useful in preparing data for BN analysis, as it allows for the efficient handling of continuous variables while maintaining their predictive power. The entropy discretisation method has been shown to be effective in various machine learning applications, often outperforming other discretisation techniques in terms of maintaining classification accuracy [32].

5. Statistical analysis

The following sections present the results of statistical analyses comparing groups across all measured variables to assess the incremental impact of each support technology. The comparison structure is designed to isolate individual technology effects:

- **Group 1 vs. Group 2:** Assesses the impact of alarm rationalisation by comparing baseline performance (G1) against alarm rationalisation (G2).
- **Group 2 vs. Group 3:** Evaluates the added value of digital (on-screen) procedures beyond alarm rationalisation by comparing G2 (alarm rationalisation + paper procedures) against G3 (alarm rationalisation + digital procedures).
- **Group 3 vs. Group 4:** Examines the incremental benefit of AI-based decision support by comparing G3 (alarm rationalisation + digital procedures) against G4 (alarm rationalisation + digital procedures + DSS).

Figs. 5a through 10b present comprehensive visualisations of all measured variables across the four groups and three scenarios. These graphical representations complement the statistical tests by illustrating the relative magnitudes and distributions of each variable. While the statistical analyses (presented in Tables 4 through 6) identify significant differences between groups, the box plots and bar charts allow readers to assess the practical magnitude of these differences and understand the variability within each group. For continuous variables (accuracy, alarm counts, response times, etc.), box plots display the median, interquartile range, and outliers, providing insight into data distribution characteristics. For categorical variables (consequence, overall performance, error rate), bar charts show the proportion of outcomes within each category. Together, these visualisations and statistical tests provide a complete picture of how each support technology impacts operator behaviour and performance.

In total, 87 participant datasets were included in the analysis: 22 from Group 1, 23 from Group 2, 21 from Group 3, and 21 from Group 4. Five participants were excluded due to incomplete data or technical issues during the experiment. All statistical analyses were performed using the Python programming language (version 3.10) with the scipy.stats and statsmodels libraries. The source code of the statistical tests and plots can be found in https://github.com/Jomietk/Support-Group-wise-Comparisons-across-Scenarios

Statistical test

In this study, we conducted a comprehensive statistical analysis to examine the effects of experimental conditions across three distinct scenarios on various operational variables. The normality of data distributions was evaluated using the Shapiro-Wilk test [33], while the Levene test [34] was utilised to assess the equality of variances among groups. Based on the results of these preliminary tests, the appropriate statistical tests were applied: the Student's t-test [35] was used for data exhibiting normal distribution and equal variances, the Welch's t-test [36] for data with normal distribution but unequal variances, and the Wilcoxon Rank-Sum Test [37] for data not adhering to normality assumptions.

Comparison between Group 1 and Group 2

In this section, we undertake a comparative analysis of operational variables between two distinct groups to assess the impact of alarm prioritisation on operator performance. Group 1 operates without the benefit of alarm prioritisation, whereas Group 2 utilises an alarm prioritisation system. Through this comparison, we aim to determine the effectiveness of alarm prioritisation in enhancing operational efficiency and decision-making processes among operators. This analysis is pivotal in determining whether the implementation of prioritised alarms can

significantly influence operator response times, accuracy, and overall performance in managing operational tasks. The result of the statistical analysis can be seen in Table 4.

Scenario 1:

- **Acknowledgement** demonstrated a significant difference; G2 had a higher acknowledgement rate ($p = 0.015$).
- **No. of alarms** revealed significant differences; G2 had a lower number of alarms ($p = 0.033$).
- The rest of the variables showed no significant differences.

Scenario 2:

- All tested variables show no significant differences.

Scenario 3:

- **Accuracy** showed significant differences with G1 demonstrating higher accuracy (p-values: 0.008).
- **Alarms silenced** and **Acknowledgement** showed significant differences with G2 having a higher rate of acknowledged and silenced alarms (p-values: 0.004, 0.007).
- **Recovery time**, **response time** showed significant differences with G2 having a higher recovery and response time (p-values: 0.000, 0.002).
- **Consequence** showing borderline statistical significant difference (p-value 0.054) with G2 having better consequences.
- **Error rate** showed significant differences with G2 demonstrating better performance (p-values: 0.034)
- The rest of the variables showed no significant differences.

The comparison between Group 1 and Group 2 highlights the influence of alarm prioritisation on operator performance. In Scenarios 1 and 2, where the number of alarms was relatively low, the differences between the groups were minimal. However, Group 2 consistently showed a higher acknowledgement rate and fewer alarms in Scenario 1, suggesting better alarm management.

In Scenario 3, significant differences emerged, with Group 2 exhibiting higher acknowledgement and silencing rates of alarms, better overall performance, and better outcomes in terms of consequences. Despite Group 2 having higher response times, these findings suggest that alarm prioritisation can enhance operator performance and efficiency, particularly in more alarm-intensive scenarios. This underscores the potential benefits of implementing alarm prioritisation systems to improve operational tasks, though the specific context and number of alarms play a crucial role in the observed outcomes.

Comparison between Group 2 and Group 3

In this analysis, we compare the performance impacts of transitioning from paper-based to digital procedure management between two distinct experimental groups. Group 2 retains the use of paper-based procedures alongside alarm prioritisation, whereas Group 3 employs an on-screen procedure display, maintaining the same alarm prioritisation system. This comparison aims to explore the effectiveness of digital procedures in enhancing operator interaction, reducing response times, and improving the accuracy and efficiency of decision-making processes. The results from this section will provide insights into whether the digitalisation of procedures significantly benefits operators in terms of operational performance and task management. The result of the statistical analysis can be seen in Table 5.

Scenario 1 (S1):

- **Response time** showed significant differences with G3 having a lower response time (p-values: 0.029).
- Other variables showed no significant differences.

Scenario 2 (S2):

- Variables showed no significant differences.

Scenario 3 (S3):

- **Acknowledgement** showed significant differences with G3, demonstrating a lower alarm acknowledgement (p-values: 0.005).
- Other variables showed no significant differences.

Scenario 1 revealed a nearly significant reduction in the number of alarms for Group 3 ($p = 0.058$), suggesting that digital procedures may help in reducing alarm clutter, thereby potentially decreasing the cognitive load on operators. Significantly lower response times in Group 3 ($p = 0.029$) further underscore the efficiency brought about by digital procedures, which likely facilitate quicker access to necessary information and streamline operator interactions.

Scenario 2 revealed no significant statistical differences.

In Scenario 3, Group3's lower alarm acknowledgement rate and a similar number of alarms suggest less efficient management of alarms (p-values: 0.005).

Overall, transitioning from paper-based to digital procedures appears to have some positive impact. The findings suggest that digital procedures improve the response time and reduce the number of alarms in low-complexity scenarios. The limited number of statistically significant differences across the various metrics, especially in complex scenarios, suggests that the impact of digitalising procedures may be less profound than initially anticipated.

Comparison between Group 3 and Group 4

This section delves into the comparative analysis of operational variables between two technologically advanced groups to assess the added impact of AI-based decision support. Group 3 operates with alarm prioritisation and digital procedures, while Group 4 enhances this setup with AI-driven DSS tools. The DSS detects the fault early and tells the operators about the potential consequences of the fault and the steps to recover. Those steps are similar to the traditional procedure but drastically reduce and adapt to the current situation. By examining these configurations, we aim to identify the incremental benefits provided by the DSS in terms of enhancing operational safety, reducing workload, and improving decision-making accuracy. This comparison is crucial for understanding if the integration of the DSS can meaningfully elevate operator performance and optimise process management under complex operational scenarios. The result of the statistical analysis can be seen in Table 6.

Scenario 1:

- **Acknowledgement**, **Alarms silenced** and **No. of alarms** showed significant differences, with G4 having a lower rate of silenced, acknowledged alarms and also a lower number of alarms (p-values: 0.026, 0.01, 0.000).
- **Overall performance** showed significant differences, with G4 exhibiting superior performance (p-values: 0.004).
- **Reaction time** and **Recovery time** showed significant differences, with G4 exhibiting lower reaction and recovery time (p-values: 0.000, 0.000).

Table 4
Statistical test result of the comparison between Group 1 and Group 2.

Variable	Scenario	Test Used	Shapiro-Wilk G1	Shapiro-Wilk G2	Levene Test	p-value
Accuracy	S1	Wilcoxon Rank-Sum	0.000	0.000	0.807	0.751
Accuracy	S2	Wilcoxon Rank-Sum	0.000	0.000	0.331	0.364
Accuracy	S3	Wilcoxon Rank-Sum	0.000	0.000	0.263	**0.008**
Acknowledgement	S1	Wilcoxon Rank-Sum	0.000	0.000	0.441	**0.015**
Acknowledgement	S2	Wilcoxon Rank-Sum	0.015	0.008	0.833	0.794
Acknowledgement	S3	Wilcoxon Rank-Sum	0.000	0.008	0.348	**0.007**
Alarms silenced	S1	Wilcoxon Rank-Sum	0.000	0.000	0.156	0.442
Alarms silenced	S2	Wilcoxon Rank-Sum	0.000	0.000	0.761	0.23
Alarms silenced	S3	Wilcoxon Rank-Sum	0.000	0.206	0.689	**0.004**
Consequence	S1	Chi-Squared				0.469
Consequence	S2	Chi-Squared				0.985
Consequence	S3	Chi-Squared				0.054
Mimics opened	S1	Wilcoxon Rank-Sum	0.000	0.003	0.916	0.385
Mimics opened	S2	Wilcoxon Rank-Sum	0.000	0.008	0.911	0.96
Mimics opened	S3	Wilcoxon Rank-Sum	0.066	0.000	0.768	0.713
No. of alarms	S1	Wilcoxon Rank-Sum	0.000	0.000	0.302	**0.033**
No. of alarms	S2	Wilcoxon Rank-Sum	0.000	0.000	0.593	0.567
No. of alarms	S3	t-test	0.054	0.329	0.929	0.584
Overall performance	S1	Chi-Squared				0.566
Overall performance	S2	Chi-Squared				0.648
Reaction time	S1	Wilcoxon Rank-Sum	0.000	0.002	0.868	0.136
Reaction time	S2	Wilcoxon Rank-Sum	0.009	0.026	0.389	0.481
Reaction time	S3	t-test	0.465	0.135	0.439	0.636
Recovery time	S1	Wilcoxon Rank-Sum	0.004	0.656	0.493	0.162
Recovery time	S2	t-test	0.946	0.665	0.841	0.352
Response time	S1	Wilcoxon Rank-Sum	0.000	0.022	0.861	0.622
Response time	S2	t-test	0.064	0.221	0.741	0.302
Response time	S3	Wilcoxon Rank-Sum	0.012	0.002	0.881	**0.002**
Error rate	S1	Chi-Squared				1
Error rate	S2	Chi-Squared				0.806
Error rate	S3	Chi-Squared				**0.034**

Table 5
Statistical test result of the comparison between Group 2 and Group 3.

Variable	Scenario	Test Used	Shapiro-Wilk G1	Shapiro-Wilk G2	Levene Test	p-value
Accuracy	S1	Wilcoxon Rank-Sum	0.000	0.000	0.912	0.162
Accuracy	S2	Wilcoxon Rank-Sum	0.000	0.000	0.313	0.601
Accuracy	S3	Wilcoxon Rank-Sum	0.000	0.000	0.557	0.758
Acknowledgement	S1	Wilcoxon Rank-Sum	0.000	0.004	0.415	0.96
Acknowledgement	S2	Wilcoxon Rank-Sum	0.008	0.006	0.885	0.855
Acknowledgement	S3	Wilcoxon Rank-Sum	0.008	0.000	0.169	**0.005**
Alarms silenced	S1	Wilcoxon Rank-Sum	0.000	0.000	0.329	0.945
Alarms silenced	S2	Wilcoxon Rank-Sum	0.000	0.000	0.379	0.595
Alarms silenced	S3	Wilcoxon Rank-Sum	0.206	0.012	0.841	0.675
Consequence	S1	Chi-Squared				1
Consequence	S2	Chi-Squared				0.383
Consequence	S3	Chi-Squared				0.664
Mimics opened	S1	Wilcoxon Rank-Sum	0.003	0.018	0.54	0.576
Mimics opened	S2	Wilcoxon Rank-Sum	0.008	0.003	0.359	0.963
Mimics opened	S3	Wilcoxon Rank-Sum	0.000	0.000	0.9	0.866
No. of alarms	S1	Wilcoxon Rank-Sum	0.000	0.000	0.183	0.058
No. of alarms	S2	Wilcoxon Rank-Sum	0.000	0.000	0.259	0.606
No. of alarms	S3	t-test	0.329	0.374	0.571	0.691
Overall performance	S1	Chi-Squared				0.221
Overall performance	S2	Chi-Squared				0.12
Reaction time	S1	Wilcoxon Rank-Sum	0.002	0.000	0.117	0.117
Reaction time	S2	Wilcoxon Rank-Sum	0.026	0.002	0.18	0.725
Reaction time	S3	Wilcoxon Rank-Sum	0.135	0.04	0.283	0.085
Recovery time	S1	Wilcoxon Rank-Sum	0.656	0.019	0.106	0.125
Recovery time	S2	t-test	0.665	0.19	0.1	0.466
Response time	S1	Wilcoxon Rank-Sum	0.022	0.000	0.451	**0.029**
Response time	S2	Wilcoxon Rank-Sum	0.221	0.055	0.05	0.617
Response time	S3	Wilcoxon Rank-Sum	0.002	0.015	0.603	0.239
Error rate	S1	Chi-Squared				1
Error rate	S2	Chi-Squared				0.339
Error rate	S3	Chi-Squared				0.555

- **No of procedures** showed significant differences, with G4 opening fewer procedures (p-values: 0.000).

Scenario 2:

- **Accuracy** and **No of alarms** showed significant differences with G4 displaying higher accuracy and fewer alarms (p-values: 0.001, 0.008).
- **Overall performance** also showed significant differences with G4 having a superior Overall performance ($p = 0.000$).
- **Reaction time**, **Recovery time**, and **Response time** showed significant differences, with G4 exhibiting lower reaction and recovery time (p-values: 0.000).

Scenario 3:

- **Acknowledgement** indicated G4 had a higher acknowledgement rate ($p = 0.018$).

The comparative analysis of Group 3 and Group 4 across various operational scenarios highlights the tangible benefits of integrating AI-based DSS in enhancing operational efficiency and safety. The data clearly demonstrates that Group 4, equipped with AI DSS, consistently outperformed Group 3, which utilised only alarm prioritisation and digital procedures.

The improved accuracy observed in Scenario 2 for G4 suggests that the DSS helps operators make more precise decisions. The lower number of acknowledged and silenced alarms in Scenario 1, combined with the overall reduction in alarm counts, implies that the DSS enables early detection and resolution of issues, thereby preventing alarms from escalating. This early intervention is further supported by the reduced reaction and response times in Scenarios 1 and 2, indicating that the DSS provides timely and relevant information, allowing operators to act more quickly and effectively.

The consistently lower recovery times in scenarios 1 and 2 for G4 highlight the DSS's role in facilitating faster resolution of problems, thus minimising downtime and enhancing operational efficiency. Additionally, the fewer procedures opened in Scenario 1 suggest that the DSS offers more streamlined and tailored guidance, reducing the need for extensive procedure consultation.

The improved consequences in Scenario 2 and enhanced overall performance in Scenarios 1 and 2 demonstrate some benefits of the DSS in assisting operators. These results suggest that advanced decision support tools can be valuable in control room environments, potentially enhancing decision-making and improving operational performance in certain situations. However, the findings from Scenario 3 reveal limitations of the DSS. In this complex scenario, Group 4 only showed a higher rate of alarm acknowledgement, with no significant impact on other variables. This indicates that during highly challenging situations, the DSS's effectiveness may be limited, highlighting the importance of not overrelying on such systems and the need for further development to address complex operational scenarios.

5.1. Summary of statistical findings

The statistical analysis revealed context-dependent effectiveness of support technologies. Alarm rationalisation demonstrated the clearest benefit, with Group 2 showing significantly lower error rates than Group 1, specifically in Scenario 3 ($p = 0.034$), where alarm flooding occurred. This finding aligns with cognitive load theory: when alarm volume overwhelms operator capacity during floods, filtering non-critical alarms proves essential. Digital procedures showed very few differences. Group 3 achieved faster response times in simple scenarios (S1: $p = 0.029$), and the advantages diminished in complex situations. In S3, Group 3 actually showed lower alarm acknowledgement than Group 2 ($p = 0.005$), suggesting potential cognitive overhead from managing multiple on-screen displays during high-workload conditions.

The AI-based DSS (Group 4) excelled in simple-to-moderate complexity scenarios, significantly improving multiple performance dimensions in S1 and S2. However, its advantages largely disappeared in S3's extreme complexity, where only alarm acknowledgement remained significantly higher ($p = 0.018$). This complexity-dependent effectiveness, where the most technologically advanced configuration did not achieve the best performance in the most challenging scenarios, challenges assumptions that more technology uniformly improves outcomes.

These patterns set the foundation for predictive modelling to identify which operational metrics most reliably forecast operator error across these varying contexts (Table 7).

6. Error prediction

In this section, we develop and compare two predictive models, stepwise Logistic regression and Bayesian network, explicitly based on the behavioural metrics and the scenario and group variables from the operational data to forecast potential outcomes: Success or Failure. This enables the identification of key patterns and trends that may influence outcomes. The predictors considered included only the behavioural variables from previously discussed metrics such as the "Reaction Time", "Response Time", "Alarms Acknowledged", "Mimics Opened", "Number of Alarms", and the "Number of Procedures". These have been selected because of the possibility of extracting them in a more practical control room setting. The "Scenario" (also seen as the Task Complexity) and "Group" (also seen as the type of support configuration) were also included.

6.1. Prediction - LR

The LR model was specified using main effects only, without interaction terms. With $N = 87$ distributed across four groups and three scenarios, and a limited number of error events, considering interaction terms would add a significant number of additional parameters, risking overfitting and unstable coefficient estimates. The LR model, therefore, serves as a first assessment tool to evaluate the individual contribution of each predictor, with the more complex conditional dependencies between variables subsequently captured by the BN analysis.

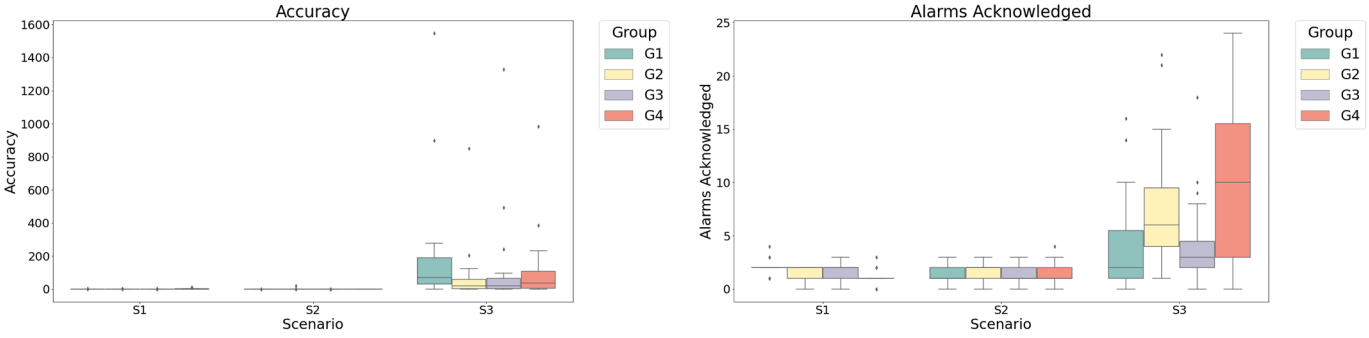
6.1.1. Step 1

The first step in the subsection involved predicting the outcome [Error: Success, Failure] based on all variables, including the type of group and scenario. Predictions were made in Python language using Jupyter Notebook. The logistic regression model outputs several statistical measures that quantify the relationship between predictor variables and the outcome (Error). The Coefficient represents the log-odds change in the outcome for a one-unit increase in the predictor variable, with positive values indicating increased error probability and negative values indicating decreased error probability. The Standard Error (Std Err) measures the precision of the coefficient estimate. The z-value is calculated by dividing the coefficient by its standard error, providing a standardised measure of the predictor's effect size. The P-value ($P > |z|$) indicates statistical significance, where values below 0.05 suggest the predictor has a statistically significant relationship with the outcome. Additionally, the model's overall performance is assessed through several metrics: Accuracy represents the proportion of correct predictions, Precision indicates the proportion of predicted errors that were actual errors, Recall measures the proportion of actual errors correctly identified, F1 Score provides a balanced measure combining precision and recall, and ROC-AUC Score evaluates the model's ability to discriminate between success and failure across all classification thresholds.

Using the logistic regression technique, the best variables with respect to the highest coefficient impact are Scenario, No of alarms,

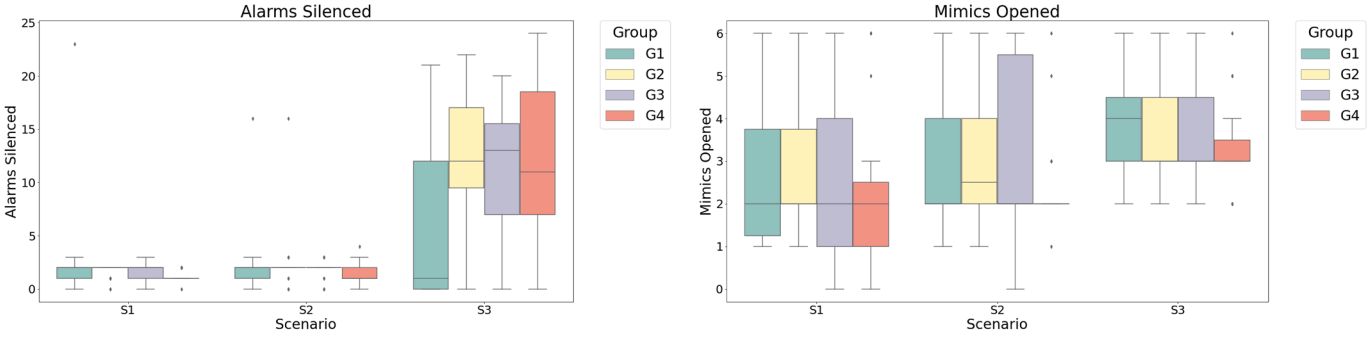
Table 6
Statistical test result of the comparison between Group 3 and Group 4.

Variable	Scenario	Test Used	Shapiro-Wilk G1	Shapiro-Wilk G2	Levene Test	p-value
Accuracy	S1	Wilcoxon Rank-Sum	0.000	0.000	0.029	0.202
Accuracy	S2	Wilcoxon Rank-Sum	0.000	0.000	0.000	**0.001**
Accuracy	S3	Wilcoxon Rank-Sum	0.000	0.000	0.92	0.629
Acknowledgement	S1	Wilcoxon Rank-Sum	0.004	0	0.097	**0.026**
Acknowledgement	S2	Wilcoxon Rank-Sum	0.006	0.006	0.701	0.512
Acknowledgement	S3	Wilcoxon Rank-Sum	0.000	0.094	0.002	**0.018**
Alarms silenced	S1	Wilcoxon Rank-Sum	0.000	0.000	0.152	**0.01**
Alarms silenced	S2	Wilcoxon Rank-Sum	0.000	0.000	0.59	0.107
Alarms silenced	S3	Wilcoxon Rank-Sum	0.012	0.085	0.253	0.732
Consequence	S1	Chi-Squared Test				1
Consequence	S2	Chi-Squared Test				**0.004**
Consequence	S3	Chi-Squared Test				0.276
Mimics opened	S1	Wilcoxon Rank-Sum	0.018	0.001	0.213	0.44
Mimics opened	S2	Wilcoxon Rank-Sum	0.003	0.000	0.006	0.201
Mimics opened	S3	Wilcoxon Rank-Sum	0.000	0.000	0.252	0.268
No. of alarms	S1	Wilcoxon Rank-Sum	0.000	0.000	0.581	**0.000**
No. of alarms	S2	Wilcoxon Rank-Sum Test	0.000	0.000	0.017	**0.008**
No. of alarms	S3	t-test	0.374	0.493	0.365	0.386
Overall performance	S1	Chi-Squared				**0.004**
Overall performance	S2	Chi-Squared				**0.000**
Reaction time	S1	Wilcoxon Rank-Sum	0.000	0.000	0.063	**0.000**
Reaction time	S2	Wilcoxon Rank-Sum	0.002	0.000	0.568	**0.000**
Reaction time	S3	Wilcoxon Rank-Sum	0.04	0.000	0.000	0.153
Recovery time	S1	Wilcoxon Rank-Sum	0.019	0.000	0.836	**0.000**
Recovery time	S2	Wilcoxon Rank-Sum	0.19	0.000	0.54	**0.000**
Response time	S1	Wilcoxon Rank-Sum	0.000	0.141	0.063	0.075
Response time	S2	Wilcoxon Rank-Sum	0.055	0.000	0.324	**0.000**
Response time	S3	Wilcoxon Rank-Sum	0.015	0.157	0.762	0.06
No. of procedures	S1	Wilcoxon Rank-Sum	0.000	0.000	0.358	**0.000**
No. of procedures	S2	Wilcoxon Rank-Sum	0.000	0.000	0.296	0.535
No. of procedures	S3	Wilcoxon Rank-Sum	0.015	0.026	0.192	0.136
Error rate	S1	Chi-Squared				1
Error rate	S2	Chi-Squared			1	
Error rate	S3	Chi-Squared				0.176



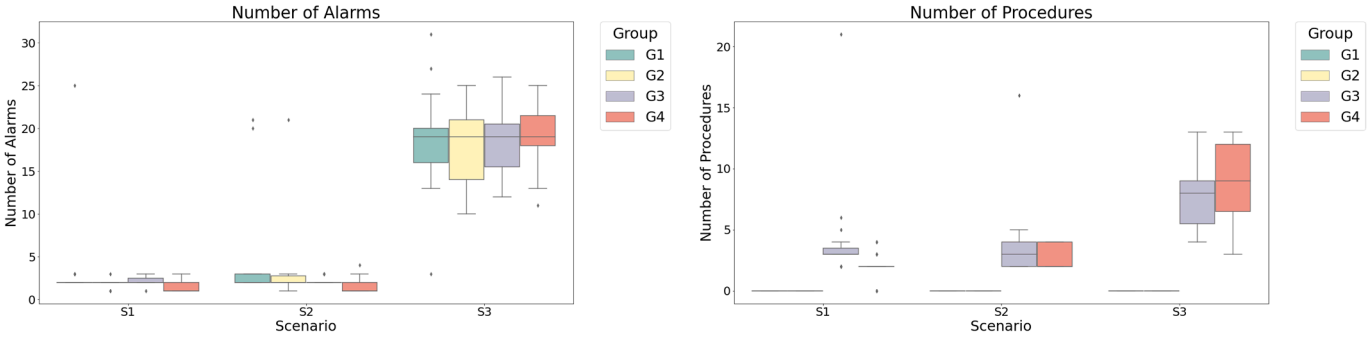
(a) Accuracy Boxplot (b) Alarms acknowledged Boxplot

Fig. 5. Accuracy and alarms acknowledged distribution.



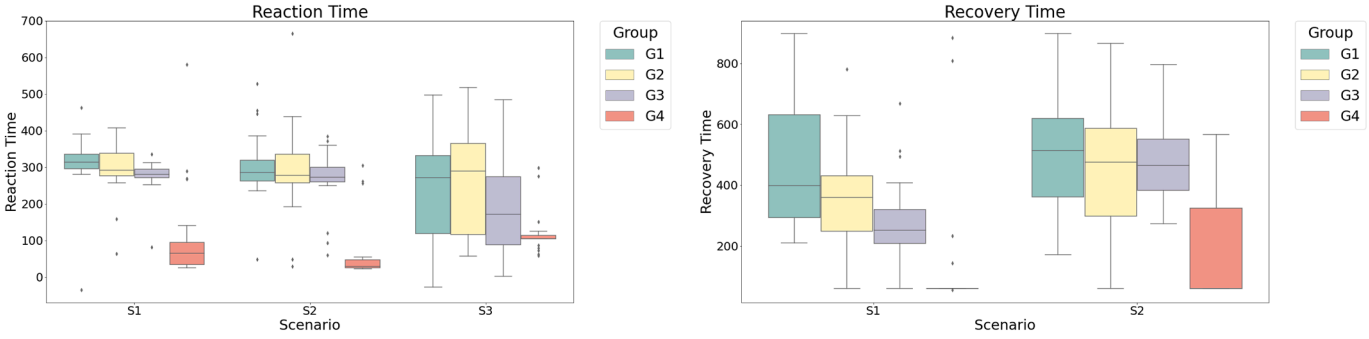
(a) Alarms silenced Boxplot (b) Mimics opened Boxplot

Fig. 6. Alarms silenced and mimics opened distribution.



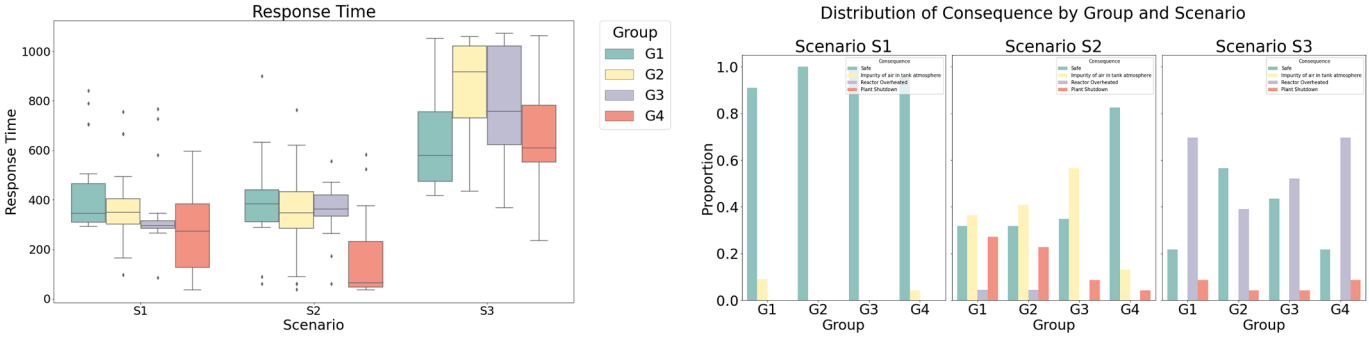
(a) Number of Alarms Boxplot (b) Procedure open Boxplot

Fig. 7. Number of alarms and procedures distribution.



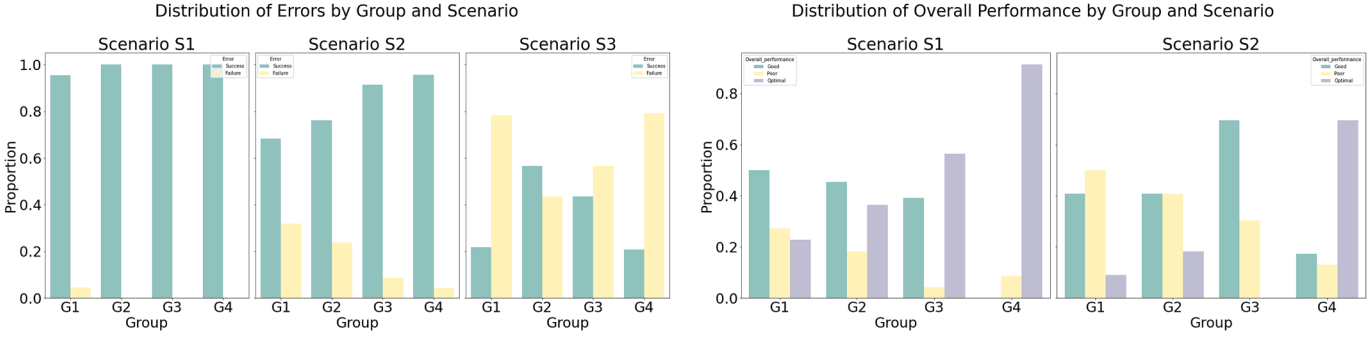
(a) Reaction Time Boxplot (b) Recovery time Boxplot

Fig. 8. Reaction and recovery times distribution.



(a) Response time Boxplot (b) Consequence Bar plot

Fig. 9. Response time and consequence distribution.



(a) Error rate Bar plot (b) Overall performance Bar plot

Fig. 10. Error rate and overall performance distributions.

Table 7
Comparison of variable impact and statistical significance across groups and scenarios. When comparing group G_i with G_{i+1}, a "+" indicates a statistically significant higher value in G_{i+1}, while a "-" denotes a lower value. A superscript ✓ signifies a favourable outcome for G_{i+1} relative to G_i, and × indicates an adverse effect. The +/- sign indicates a difference in the distribution.

Variable	Scenario	Group G1 vs G2	Group G2 vs G3	Group G3 vs G4
Accuracy	S1			
	S2			$+^{\times}$
	S3	$-^{\checkmark}$		
Alarms Acknowledged	S1		$+^{\checkmark}$	$-^{\times}$
	S2			
	S3	$+^{\checkmark}$		$+^{\checkmark}$
Alarms Silenced	S1			$-^{\times}$
	S2			
	S3	$+^{\checkmark}$		
Mimics Opened	S1-S3			
Number of Alarms	S1	$-^{\checkmark}$		$-^{\checkmark}$
	S2			$-^{\checkmark}$
	S3			
Reaction Time	S1			$-^{\checkmark}$
	S2			$-^{\checkmark}$
	S3			
Recovery Time	S1			$-^{\checkmark}$
	S2			$-^{\checkmark}$
Response Time	S1		$-^{\checkmark}$	
	S2			$-^{\checkmark}$
	S3	$+^{\checkmark}$		
Number of procedure	S1			$-^{\checkmark}$
	S2			
	S3			
Consequence	S1			
	S2			$+^{\checkmark}$
	S3			
Overall performance	S1			$+^{\checkmark}$
	S2		+/−	$+^{\checkmark}$
Error rate	S1			
	S2			
	S3	$-^{\checkmark}$		

Table 8
Coefficients and statistical significance of predictors using LR. All groups and scenarios combined and without feature selection. *Accuracy: 86%, Precision: 76.9%, Recall: 71%, F1 Score: 74%, ROC-AUC Score: 92%* .

Predictor	Coefficient	Std Err	z-value	$P > \|z\|$
Group	−2.1600	0.551	−3.923	0.000
Scenario	3.8938	1.182	3.295	0.001
Reaction_time	−1.3770	0.877	−1.570	0.116
Response_time	−2.5546	1.079	−2.368	0.018
Acknowledgements	−1.2167	1.123	−1.083	0.279
Mimics_opened	−2.7103	1.028	−2.637	0.008
No_of_alarms	3.4684	1.294	2.681	0.007

Table 9
Coefficients and statistical significance of the predictors using a Step-wise Logistic Regression Algorithm. *Accuracy: 86%, Precision: 76.9%, Recall: 71%, F1 Score: 74%, ROC-AUC: 92.6%.*

Predictor	Coefficient	Std Err	z-value	$P > \|z\|$
Group	−2.1600	0.551	−3.923	0.000
Scenario	−3.4045	1.037	−3.283	0.001
Reaction_time	1.4152	0.916	1.544	0.123
Mimics_opened	2.0829	0.822	2.535	0.011
Response_time	1.9138	0.968	1.977	0.048
No_of_alarms	−2.3328	1.202	−1.940	0.052

acknowledgement, reaction time, group, response time, and mimics opened (see Table 8).

The confusion matrix and derived metrics indicate that the model performs quite well overall with an Accuracy of 86%, Precision: 76.9%, Recall: 71%, F1 Score: 74%, and ROC–AUC Score: 92%. Overall, the high values of precision, recall, F1, and ROC–AUC score demonstrate that the model is effective at distinguishing between the classes, especially given the high recall, which indicates strong performance in identifying actual positive cases. The slightly lower precision compared to recall suggests there are some false positives, but this is balanced by the high recall rate.

However, the relevant variables should only be considered for prediction, as this has the advantage of providing a better model. To achieve this, a step-wise logistic regression is applied.

6.1.2. Performing a step-wise logistic regression

After performing a step-wise logistic regression with a forward-feature selection, the key variables according to the highest coefficient impact are 'Scenario', 'No. of alarms', 'Group', 'Mimics opened' and 'Response time'. Using this approach produced a slightly better ROC–AUC score of 93% compared to the previous (Table 9).

Both outcomes in steps 1 and 2 indicate good performance, but using the step-wise logistic regressions can streamline the number of variables included and potentially improve performance.

6.1.3. Outcome per group and scenario with LR

Given that the type of scenario/event and group impacts the outcomes and behaviour of the variables significantly, it would be important to see how each group in each scenario perform as shown in Table 14. To calculate the probability of failure for each group and scenario, the same step-wise logistic model was employed as before, using the same selected features: number of alarms, response time, mimics opened, Group and Scenario. For each data point, the probability of failure was calculated using the logistic function:

$$p(failure) = \frac{1}{1 + e^{-(\beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3 + \beta_4 x_4)}} \quad (2)$$

Table 10
Probability of success for each group in scenarios 2 and 3 according to the step-wise logistic regression model.

Scenario	Group 1	Group 2	Group 3	Group 4
S1	0.94	0.96	0.96	0.97
S2	0.79	0.84	0.87	0.89
S3	0.31	0.38	0.43	0.45

Table 11
Coefficients and statistical significance of additional predictors. *Accuracy: 84%, Precision: 71.00%, Recall: 71%, F1 Score: 71% and ROC–AUC: 91.7%, at threshold of 0.5* .

Predictor	Coefficient	Std Err	z-value	$P > \|z\|$
Scenario	3.8900	0.964	4.036	0.000
Response_time	−2.7665	0.998	−2.771	0.006
TLX_Index	3.4879	1.313	2.657	0.008
SART_Index	−7.6300	1.151	−6.627	0.000

where β_i are the model coefficients and x_i are the standardized feature values. Then the probabilities are aggregated by group and scenario, calculating the mean probability for each combination. Table 10 presents the resulting probabilities of failure for each group across the three scenarios. These results indicate that Scenario S3 consistently presents the highest risk of failure across all groups, with probabilities ranging from 0.55 to 0.69. Group_1 shows the highest overall probability of failure, particularly in Scenario S3 (0.69), while Group_4 demonstrates the lowest overall probability of failure across all scenarios. Also, Group_3 and Group_4 had close outcomes in terms of failure probability across the three scenarios.

6.2. Predicting with subjective measures

As an extra step towards exploring the importance of classical performance shaping factors on alarm response outcomes, some of the subjective variables measured during the experiment, such as training, familiarity, situational awareness and workload, were included. The situational awareness techniques used for the questionnaire-based data collection are the situational awareness rating technique (SART) and the situation presence assessment method (SPAM), while that of workload is the NASA Task Load Index (NASA-TLX). When included in the dataset and run for the combined group and scenario comparison using the step-wise logistic regression technique, the result in Table 11 was realised. Some behavioural variables were not significant for prediction following the inclusion of these subjective variables. Also, training and familiarity were not significant for the prediction.

Incorporating the new subjective variables, as detailed in Table 11, identified SART Index, Response time, Scenario, and TLX Index as key predictors. Variables such as the number of alarms, reaction time, group, mimics opened, acknowledgement, SPAM Index, familiarity, and training were not chosen by the step-wise LR model. Nevertheless, it remains evident that Scenario and response time consistently proved to be significant indicators of error in both earlier and recent predictions. Moreover, the importance of situational awareness and workload for correct alarm response was confirmed by the significance of SART and NASA TLX for the error prediction as shown in Table 11. These findings highlight the necessity of considering both behavioural and cognitive measures when evaluating operator performance.

6.3. Prediction - Bayesian network

In this study, we extended our analysis to encompass both the NB model and the TAN model for the predictive assessment of operational outcomes. This dual approach enables us to leverage the intrinsic advantages of Bayesian network methodologies in understanding and modelling the complexities inherent in operational data. We use HUGIN EXPERT software to build and assess the Bayesian network. The operational variables considered in our research include reaction time, total procedure count, alarm silencing, alarm acknowledgement, mimic opening, accuracy, response time, total alarm count, and classifications by group and scenario. We use the scenario as a complexity variable and the group as different levels of support. These variables are critical in constructing a comprehensive framework for our predictive analysis.

Table 12
Average on 1000 folds of the confusion matrix for TAN model performance.

	Predicted Negative	Predicted Positive
Actual Negative	55.54	3.46
Actual Positive	5.10	17.89

Table 13
Confusion matrix for NB model performance.

	Predicted Negative	Predicted Positive
Actual Negative	54.02	4.98
Actual Positive	5.21	17.79

To validate the robustness and applicability of our models across varied operational settings, we employed a Stratified Shuffle Split cross-validation strategy with 1000 folds. Additionally, we utilised an entropy-based discretisation technique for the effective categorisation of continuous variables into discrete intervals. Given the small size of the dataset and the variability introduced by entropy discretisation on the training set of each fold, the performance fluctuates. Therefore, a large number of folds were used to ensure the stability and reliability of the results.

The performance of the models was evaluated using a confusion matrix and several key metrics. The confusion matrix indicates the distribution of predictions as shown in Tables 12 and 13:

A threshold of 50% was chosen for the classification of control room operator error. The model achieved a precision of 0.838, indicating that 83.8% of the predicted positive instances were true positives. The accuracy of 0.895 shows that 89.5% of all predictions were correct, while a recall of 0.778 demonstrates the model's effectiveness in identifying 77.8% of actual positive instances. The F1 score of 0.807 balances precision and recall, confirming the model's solid performance in predicting human errors. Also, the ROC–AUC is 0.934. This level of performance suggests that the TAN model is effective in supporting decision-making processes in the control room by accurately predicting potential human errors based on metrics like "Recovery Time," "Reaction Time," "Response Time," "Accuracy," "Alarms Acknowledged," "Mimics Opened," and "Number of Alarms."

In comparison, the NB model attained a precision of 0.781, accuracy of 0.876, recall of 0.773, an F1 score of 0.777 and ROC–AUC of 0.921, indicating that the NB model has solid but slightly lower performance compared to the TAN model in predicting human errors. The confusion matrix for the NB model is as follows:

The TAN model structure (Fig. 11) reveals relatively simple conditional dependencies among predictor variables. "Complexity" (Scenario) dominates as the tree root, connecting to nearly all behavioural metrics except "Acknowledgements", which instead connects to "Level of Support" (Group). This indicates that alarm acknowledgement behaviour is more strongly associated with support configuration type, while most other behavioural variables are primarily influenced by scenario complexity.

In Fig. 11, we can also see the strength of the relation between the variables by edge width. The widths of the links are proportional to the symmetric mutual information. We can notably see the strong link between the complexity of the scenario and the number of alarms, and between error probability and the number of alarms, reflecting the increased number of alarms depending on the scenario.

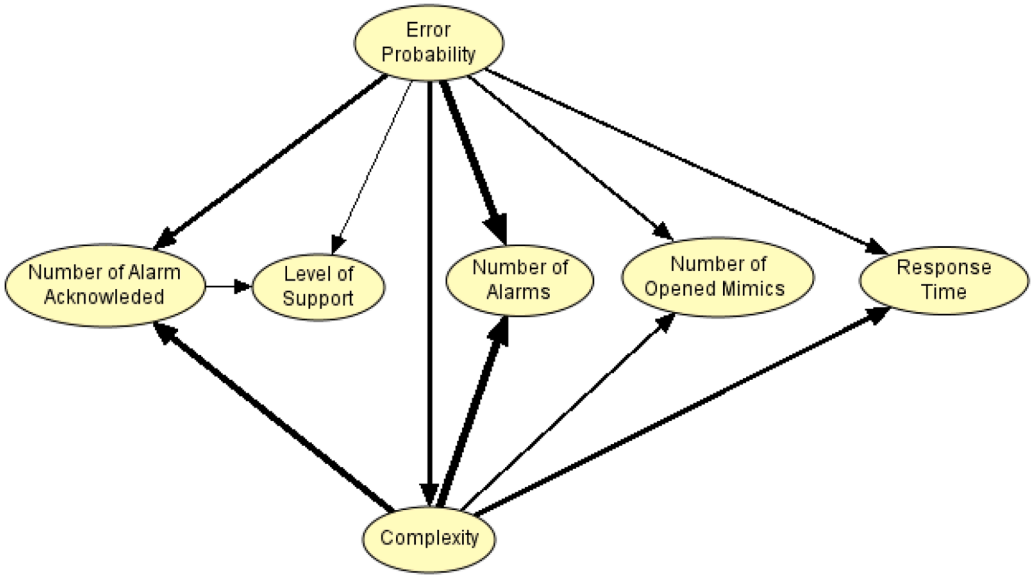


Fig. 11. TAN model structure for predicting operator error probability. "Error Probability" (top centre, class variable) connects to all predictor variables. Predictor variables form a maximum-weight spanning tree rooted at "Complexity" (Scenario), with edges pointing away from the root. This structure indicates that scenario complexity primarily drives most behavioural metrics, while support configuration type behaviour is more strongly associated with alarm acknowledgement. Edge thickness indicates mutual information strength.

In terms of performance, both models demonstrate strong predictive capabilities. However, the slight performance edge of the TAN model underscores the importance of considering variable dependencies in probabilistic models. This comparison also highlights that the simpler NB model remains a viable and efficient tool for similar predictive tasks, especially when ease of implementation and computational efficiency are prioritised.

6.4. BN model analysis

The TAN model, having demonstrated superior performance in our comparative analysis, warrants a more in-depth examination. This section delves into the intricacies of the TAN model, leveraging mutual information analysis to uncover the relative importance of various factors in predicting operator errors. We'll also explore how the model estimates error probabilities across different scenarios and operator groups, providing crucial insights into the effectiveness of DSS in complex control room environments. Through this analysis, we aim to shed light on the key drivers of operator performance.

6.4.1. Mutual information

In our study, mutual information (MI) was utilised to assess the dependency between the operational variables and the target variable 'Error'. The MI values indicate the reduction in uncertainty about the 'Error' variable given the knowledge of the predictor variable. Higher MI values denote a stronger relationship between the predictor and the target variable.

The analysis, as depicted in Fig. 12, highlights *No. of alarms* with the highest MI score of 0.26. This indicates that the number of alarms is a critical factor influencing the error variable, implying a strong dependency and suggesting that a higher number of alarms is significantly associated with errors in operations.

Scenario follows with an MI score of 0.20, suggesting that different operational scenarios contribute significantly to the prediction of errors. This emphasises the need to consider scenario-specific factors when analysing operational performance and predicting errors.

Acknowledgements and *mimics opened* also show notable MI values (0.08 and 0.05, respectively). These variables indicate that the rate of alarm acknowledgements and the number of mimic displays opened during operations have a meaningful impact on the error variable, though to a lesser extent compared to *No. of alarms* and *Scenario*.

On the other hand, *Response time* and *Group* present lower MI scores (0.04 and 0.01, respectively). The low score for *Group* suggests a minimal direct impact on the error variable within the context of this study. Similarly, the relatively low MI for *Response time* indicates that while response time affects the error variable, its influence is less significant compared to the number of alarms and scenario conditions.

It is important to note that *Reaction time* is not present in the MI analysis. This absence is due to the application of the entropy-based discretisation algorithm. The algorithm determines optimal cut points for continuous variables by maximising the class-attribute interdependence. In this case, the algorithm could not find any threshold for Reaction time that would significantly improve the prediction of the error variable. This indicates that *Reaction time* has no discernible connection with the error variable, reinforcing its lack of predictive power within the context of our study.

In summary, the mutual information analysis identifies *No. of alarms* and *Scenario* as the most significant predictors of the error variable. These insights are crucial for developing targeted strategies to minimise errors in operations. By focusing on these key variables, it is possible to enhance operational performance and reduce the incidence of errors.

6.4.2. Outcome per group and scenario

Table 14 presents the probability of success for each group across three different scenarios (S1, S2, and S3). The success probabilities provide insight into the effectiveness of different operational setups and technologies in achieving desired outcomes under varying conditions.

Scenario 1 (S1): All groups demonstrated a high probability of success, with Group 1, Group 2, and Group 3 each achieving a 0.99 probability of success, and Group 4 slightly lower at 0.98. This indicates that under less challenging conditions, the presence of alarm prioritisation, digital procedures, or AI decision support does not significantly affect the probability of success, as all groups perform at a near-optimal level.

Scenario 2 (S2): The differences between groups become more apparent in this scenario. Group 2, which uses alarm prioritisation, showed a substantial improvement with a probability of success at 0.87, compared to Group 1's 0.69. Group 3, with digital procedures, had a success probability of 0.86, slightly lower than Group 2 but still significantly

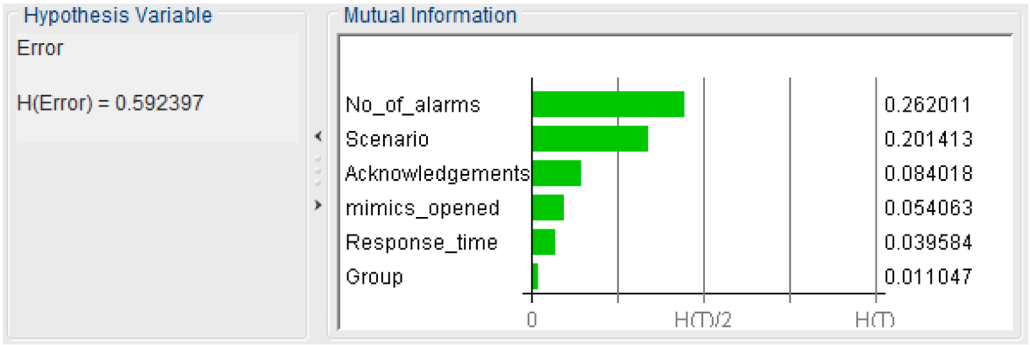


Fig. 12. Mutual information between the error and the feature variables. The entropy of the error H(error) can be seen on the right.

Table 14
Probability of success for each group across scenarios for each group and scenario according to the TAN model built.

Scenario	Group 1	Group 2	Group 3	Group 4
S1	0.99	0.99	0.99	0.98
S2	0.69	0.87	0.86	0.91
S3	0.22	0.52	0.36	0.31

Table 15
Performance comparison of BN and LR models.

Metric	TAN Model	Step-wise LR Model
Accuracy	89.5%	86%
Precision	83.8%	76.9%
Recall	77.8%	71%
F1 Score	80.7%	74%
ROC-AUC	93.4	93%

higher than Group 1. Group 4, which incorporates AI decision support, achieved the highest success probability at 0.91. These results suggest that alarm prioritisation, digital procedures, and AI decision support all contribute to improved performance in moderately challenging conditions, with AI providing the most significant benefit.

Scenario 3 (S3): This scenario presents the most challenging conditions, where the probability of success drops considerably across all groups. Group 1, without any advanced systems, has the lowest success probability at 0.22. Group 2's success probability improves to 0.52, indicating that alarm prioritisation offers a significant advantage. Group 3's probability is 0.36, showing that the use of the digital procedure increased the error rate compared to the paper-based procedure. Group 4, with AI support, has a success probability of 0.31, which, while better than Group 1, is lower than both Groups 2 and 3. This suggests that under highly challenging conditions, alarm prioritisation with paper-based procedures provides the greatest benefit, and the addition of AI support does not perform as well as expected, potentially due to the complexity of the scenario or the need for further refinement of AI tools.

Overall, the data indicate that while all advanced systems (alarm prioritisation, digital procedures, and AI decision support) improve success probabilities, their effectiveness varies depending on the scenario's complexity. The DSS group performed better in moderate complexity, and the paper-based procedure with alarm rationalisation performed better in complex scenarios.

7. Discussion

This study examined control room operator performance through two complementary analytical lenses: statistical comparison of support configurations and predictive modelling of operator error. Together, these approaches provide insights into which support technologies improve performance and why certain operational metrics predict successful alarm response.

7.1. Integration of support configuration and predictive modelling findings

The convergence between statistical analysis and predictive models strengthens confidence in our key findings. Alarm rationalisation's significant impact in reducing S3 error rates (G1 vs G2: $p = 0.034$) is reflected in both predictive models, which show substantial probability differences: BN predicted 22% success for G1 versus 52% for G2 in S3, while LR showed 31% versus 38% improvement. This consistency across analytical methods confirms alarm rationalisation's critical value during alarm flooding conditions.

However, the limited overall predictive importance of "Group" (support type), ranked last in TAN's mutual information analysis ($MI = 0.01$) and third in LR despite statistical significance, reveals an important nuance. While specific group comparisons showed significant differences for individual variables, support type does not consistently predict error across all scenarios. This apparent contradiction resolves when considering context-dependency: alarm rationalisation matters intensely during alarm floods (S3) but provides little advantage when alarm rates are manageable (S1, S2). The use of DSS shows similar context-dependent patterns with a positive impact in simple and moderate complexity scenarios. Thus, "Group" exhibits low overall mutual information with error despite scenario-specific statistical significance.

The predictive models identified the number of alarms and scenario complexity as dominant error predictors, with behavioural metrics like mimics opened, and response time contributing secondary predictive value. This relates to the finding of [6] that the most important features to predict operator workload were the number of alarms and the amount of operator–system interaction.

7.2. Comparison of Bayesian network and logistic regression approaches

While both methods demonstrated strong predictive capabilities, they offered unique insights and diverge in some aspect.

Both the TAN model and the step-wise LR model showed high predictive accuracy.

Table 15 presents a comparative analysis of the performance metrics between the TAN model and the Step-wise LR model. Both models demonstrate strong predictive capabilities, with the TAN model slightly outperforming the LR model across all metrics. The TAN model achieves higher accuracy (89.5% vs. 86%), precision (83.8% vs. 76.9%), recall (77.8% vs. 71%), and F1 score (80.7% vs. 74%). The ROC–AUC scores are nearly identical, with the TAN model at 93.4 and the LR model at 93%. These results indicate that while both models are effective in predicting operator errors, the BN approach (TAN) provides marginally better overall performance, particularly in terms of precision and recall. This suggests that the TAN model may be slightly more reliable in identifying control room operator error and minimising false positives.

Table 16
Ranked key predictors identified by BN and LR models.

TAN Mutual Information Analysis	Step-wise LR
1. No. of alarms	1. Scenario
2. Scenario	2. No. of alarms
3. Acknowledgements	3. Group
4. Mimics opened	4. Mimics opened
5. Response time	5. Response time
6. Group	6. Reaction time

The two methods identified similar key predictors, but with some notable differences. Table 16 summarises the most important variables identified by each method.

Both of the models place the Number of alarms and scenario/task complexity as the main factors influencing control room operator error. The LR model placed more importance on the 'Group/type of support' variable, which was less significant in the BN analysis. Conversely, the BN model emphasised 'Acknowledgements', which was not selected in the final LR model. Also, neither model included reaction time as an important predictor.

7.3. Unexpected findings and theoretical implications

Several findings challenge conventional assumptions about technology and operator performance:

1. More technology ≠ better performance in high complexity. Group4's AI-based DSS, despite superior performance in S1 and S2, did not achieve consistently better outcomes in S3. The two predictive models diverge on this point: the BN predicted G2 (alarm rationalisation + paper procedures) outperforming G4 with 52% versus 31% success probability, while LR suggested a marginal advantage for G4 (45% versus 38%). Critically, however, the statistical analysis revealed G4's S3 advantages limited to alarm acknowledgement only ($p = 0.018$), with no significant improvement in error rate or overall performance. Taken together, these results suggest that under high complexity alarm flooding conditions, the incremental benefit of AI decision support over alarm rationalisation alone is at best marginal and model-dependent, contradicting assumptions that more advanced automation uniformly improves outcomes.

Post-experiment observations revealed heterogeneous DSS usage strategies in S3: some operators followed DSS recommendations strictly, others cross-referenced with procedures (increasing cognitive load), and some minimised DSS consultation under high workload. Additionally, the DSS's 10-s temperature adjustment recommendations in S3 are better suited to automated rather than manual implementation, overwhelming operators attempting continuous manual control. These factors suggest that interface design and human-automation interaction principles become limiting factors under extreme stress, highlighting that technology effectiveness depends critically on matching automation recommendations to human supervisory control capabilities.

2. Digital procedures showed inconsistent benefits. The transition from paper (G2) to digital procedures (G3) produced the most ambiguous results. Statistical analysis showed faster S1 response times ($p = 0.029$) but lower S3 alarm acknowledgement ($p = 0.005$). Predictive models diverged on net impact: LR suggested modest improvement (+3-5% success in S2-S3), while BN indicated degradation (-1% in S2, -16% in S3).

This inconsistency likely reflects that digital procedure effectiveness depends heavily on implementation quality, features like context-sensitivity, seamless alarm system integration, and navigation design, which may not have been optimally realised in this study's implementation. Simply digitising paper content without a strong human factors-informed interface design appears insufficient. This aligns with broader findings that digitalisation benefits require more than technology adoption; they demand thoughtful human-centred design.

7.4. Practical implications for control room design

These findings carry several actionable implications:

Alarm rationalisation investment is justified, especially for alarm flood scenarios. The clear benefits during S3's alarm flooding, both in statistical comparisons and predictive models, support prioritising alarm rationalisation implementation. The effectiveness appears specifically when needed: during information overload conditions, where filtering reduces extraneous cognitive load.

Digital procedure implementation requires careful human factors design. The mixed results suggest that merely transferring paper procedures to screens provides limited benefit. Effective digital procedures likely require advanced features: context-aware guidance, real-time process data integration, intelligent navigation, and seamless alarm system coordination. Organisations considering digitalisation should carefully invest in human factors evaluation and iterative design.

AI decision support systems need context-appropriate design and training strategies. While DSS enhanced performance in simple-moderate scenarios, limitations in S3 indicate important design principles: (1) DSS recommendations should match human supervisory control capabilities rather than mimicking automated control loops, (2) operators require training on when and how to use DSS across complexity levels, (3) trust calibration matters, undertrust prevented some operators from leveraging DSS benefits, and (4) interface design must account for competing attentional demands during extreme stress.

Real-time error prediction systems are feasible with accessible metrics. Both models achieved strong performance (≥ 86% accuracy, ≥ 93% ROC-AUC) using only behaviourally-derived metrics accessible from standard control system historians: alarm counts, acknowledgement patterns, mimic navigation, and response times. This enables practical deployment without intrusive psychophysiological measurement equipment. Control rooms could implement continuous error probability estimation to trigger support interventions, additional operator backup, supervisory consultation, or temporary task simplification when error risk exceeds thresholds.

7.5. Alignment with cognitive load theory and human-automation interaction

The pattern of results aligns well with established theoretical frameworks. Cognitive load theory explains alarm rationalisation's scenario-specific effectiveness: it reduces extraneous load during information overload (S3) but provides minimal benefit when intrinsic load is manageable (S1, S2). Similarly, digital procedures' mixed results may reflect competing effects, potentially reducing load through enhanced navigation but potentially increasing load through split-attention between multiple screens, with net effects depending on implementation quality and scenario demands.

Human-automation interaction theory illuminates DSS limitations in S3. Appropriate reliance on automation requires calibrated trust, a clear understanding of automation capabilities and limitations, and interface transparency, all of which become difficult to maintain during high-workload alarm floods. The heterogeneous usage strategies and some operators' undertrust reflect this challenge. Moreover, the mismatch between DSS's 10-s update recommendations and human manual control capabilities exemplifies poor function allocation: assigning tasks to humans that exceed their temporal processing abilities while underutilising their adaptive problem-solving strengths.

7.6. Methodological contributions

This study demonstrates value in combining statistical group comparisons with predictive modelling. Statistical analysis identified which support configurations improved which specific performance dimensions across scenarios. Predictive modelling revealed which operational

metrics most reliably forecast error and how support configurations ultimately affect error probability. Together, they provide both mechanism-level understanding and outcome prediction.

The use of both LR and Bayesian Networks proved complementary. LR provided transparent coefficient interpretation valuable for stakeholder communication, while BN revealed conditional dependency structures and enabled probabilistic inference with missing data. Their convergence on key predictors (alarms, scenarios) strengthened confidence, while their divergence on secondary predictors (Group importance) highlighted different aspects of predictor influence, direct linear effects versus conditional dependencies.

The study also validates behavioural metrics from control system historians as viable alternatives to psychophysiological measurements for real-time error prediction, offering practical advantages for operational deployment: non-intrusiveness, no special equipment, automatic logging, and retrospective analysis capability.

7.7. Limitations and future research directions

Several limitations warrant consideration. The simulated environment, while realistic, cannot fully capture operational complexities of genuine industrial settings, including organisational factors, shift patterns, fatigue, and long-term system familiarity. The student participant population, though possessing fundamental process knowledge, differs from experienced operators with developed heuristics and procedural expertise. Future studies with professional control room operators in operational settings would strengthen external validity.

The relatively modest sample size ($n = 23$ per group) provides adequate power for primary comparisons but may limit detection of smaller effects or complex interactions. The digital procedure implementation may not have realised full potential benefits; more advanced implementations with context-awareness and intelligent integration warrant investigation. Similarly, DSS interface refinement could address S3 limitations, particularly by better matching recommendations to human supervisory control capabilities.

A further limitation concerns the specification of the Logistic Regression model. As interaction terms were excluded in favour of parsimony and to avoid overfitting, the LR model captures only the average effect of support configuration across all complexity levels. Since the effect of the Group variable is heterogeneous across scenarios, most notably, alarm rationalisation providing significant benefit during alarm flooding in Scenario 3 but a minimal advantage in simpler conditions, this likely leads to a partial underestimation of its predictive importance in the LR model.

A limitation of this study is that model performance was evaluated using complete scenario data collected after scenario completion. While this demonstrates which metrics predict outcomes and validates accuracy (TAN: 89.5%, LR: 86%), it does not evaluate the models' capacity for early error detection during ongoing scenarios. However, the Bayesian Network framework employed here is inherently suited for incremental prediction. As behavioural metrics accumulate progressively, the BN can update error probability estimates using its probabilistic inference capability for partially observed data. The current study establishes that these metrics predict outcomes; future research must determine when, during scenarios, predictions become sufficiently reliable to trigger interventions. Specifically, future work should answer: (1) At what time points do error probabilities stabilise enough to support intervention decisions? (2) Which metrics provide the earliest reliable signals? (3) What probability thresholds warrant different interventions (guidance, supervisor alert, task reallocation)? For example, if error probability exceeds 0.70 at the 8-min mark in a 15-min scenario, supervisory assistance could prevent failure through timely intervention. This temporal analysis is critical for translating predictive models into operational decision support systems that enable prospective error prevention rather than retrospective classification.

8. Conclusion

This study addressed three fundamental questions about control room operator performance during alarm response scenarios: whether operational data can predict alarm response success, which behavioural metrics are most predictive, and how different support configurations impact performance across varying task complexity levels. Our findings establish that real-time behavioural metrics accessible from standard control system historians, alarm counts, acknowledgement patterns, mimic navigation, and response times can reliably predict operator error. Both Logistic Regression and Bayesian Network models achieved strong predictive performance, demonstrating practical feasibility for continuous error probability monitoring without intrusive measurement equipment. The number of alarms and scenario complexity emerged as dominant predictors across both analytical approaches, with secondary contributions from acknowledgement patterns, mimic navigation, and response times. The comparative evaluation of support configurations revealed context-dependent effectiveness that challenges simplistic assumptions about technology and performance. Alarm rationalisation provided clear value during information overload conditions. Digital procedures showed benefits in response efficiency for simple scenarios but presented implementation challenges requiring careful human factors design. AI-based decision support enhanced performance in moderate complexity scenarios yet exhibited limitations during extreme complexity, highlighting critical considerations for human-automation interaction design, appropriate function allocation, and operator training strategies. Methodologically, this work demonstrates the value of combining statistical group comparisons with predictive modelling, using complementary analytical techniques (LR and BN) to reveal both direct predictor effects and complex conditional dependencies. The validation of historian-derived behavioural metrics as viable alternatives to psychophysiological measurements addresses a significant practical barrier to operational deployment of real-time performance monitoring systems. Future research should focus on temporal aspects of error prediction, determining when during scenarios predictions become sufficiently reliable to trigger interventions, and validation with experienced operators in operational industrial settings.

Declaration of generative AI and AI-assisted technologies in the writing process

During the preparation of this work, the authors used Claude AI from Anthropic to clarify long sentences and correct spelling. After using this tool/service, the authors reviewed and edited the content as needed and take full responsibility for the content of the published article.

CRediT authorship contribution statement

Chidera W. Amazu: Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Investigation, Formal analysis, Data curation, Conceptualization; **Joseph Mietkiewicz:** Writing – review & editing, Writing – original draft, Visualization, Validation, Methodology, Formal analysis, Data curation, Conceptualization; **Ammar N. Abbas:** Writing – review & editing, Writing – original draft, Formal analysis, Data curation; **Gabriele Baldissone:** Writing – review & editing, Software, Resources, Investigation; **Davide Fissore:** Writing – review & editing, Supervision; **Micaela Demichela:** Writing – review & editing, Supervision, Resources, Project administration; **Anders L. Madsen:** Writing – review & editing, Supervision; **Maria Chiara Leva:** Writing – review & editing, Writing – original draft, Supervision, Methodology, Funding acquisition, Conceptualization.

Consent and ethics statement

Participants were briefed prior to the study, read a detailed description of what the study entailed via an information sheet, and signed

the necessary consent form before participating. These documents, together with the ethics application, were approved by the Internal Ethical Committee of the Collaborative Intelligence for Safety-Critical Systems, following a first approval by the Ethics Review Committee of the Technological University of Dublin, Ireland, with approval number REC-20-52.

Funder information and acknowledgements

This work has been carried out within the Collaborative Intelligence for Critical Safety Systems (CISC) project. The CISC project has received funding from the European Union's Horizon 2020 Research and Innovation Programme under Marie Sklodowska-Curie grant agreement no. 955901. The authors thank Rob Turner and Adrian Kelly at Yokogawa and EPRI Europe for contributing to the development of the human-system interfaces. Finally, thanks to the participants for the time they dedicated to this study and their feedback.

Data availability statement

Preliminary data supporting this study's findings are available on GitHub at https://github.com/CISC-LIVE-LAB-3/dataset.

Declaration of competing interest

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

References

[1] Simonson RJ, Keebler JR, Blickensderfer EL, Besuijen R. Impact of alarm management and automation on abnormal operations: a human-in-the-loop simulation study. Appl Ergon 2022;100:103670. https://doi.org/10.1016/j.apergo.2021.103670

[2] Gao Q, Wang Y, Song F, Li Z, Dong X. Mental workload measurement for emergency operating procedures in digital nuclear power plants. Ergonomics 2013;56(7):1070–85. https://doi.org/10.1080/00140139.2013.790483

[3] Hinss MF, Brock AM, Roy RN. Cognitive effects of prolonged continuous human-machine interaction: the case for mental state-based adaptive interfaces. Front Neuroergon 2022;3. https://doi.org/10.3389/fnrgo.2022.935092

[4] Gertman DI, Blackman HS, Marble JL, Smith C, Boring RL, O'Reilly P. The SPAR-h human reliability analysis method. American nuclear society 4th international topical meeting on nuclear plant instrumentation, control and human machine interface technology 2004;:17–24.

[5] Iqbal MU, Srinivasan B, Srinivasan R. Multi-class classification of control room operators' cognitive workload using the fusion of eye-tracking and electroencephalography. Comput Chem Eng 2024;181. https://doi.org/10.1016/j.compchemeng.2023.108526

[6] Braarud PO, Bodal T, Hulsund JE, Louka MN, Nihlwing C, Nystad E, et al. An investigation of speech features, plant system alarms, and operator–system interaction for the classification of operator cognitive workload during dynamic work. Hum Factors 2021;63(5):736–56. https://doi.org/10.1177/0018720820961730

[7] Rahman H, Ahmed MU, Barua S, Begum S. Non-contact-based driver's cognitive load classification using physiological and vehicular parameters. Biomed Signal Process Control 2020;55:101634. https://doi.org/10.1016/j.bspc.2019.101634

[8] Chen F, Ruiz N, Choi E, Epps J, Khawaja MA, Taib R, et al. Multimodal behavior and interaction as indicators of cognitive load. ACM Trans Interact Intell Syst 2012;2(4). https://doi.org/10.1145/2395123.2395127

[9] Schmidt B, Borrison R, Gärtler M, Maczey S, Kotriwala A. Practical aspects for exploration and analysis of manual interventions in process plants. In: 2021 IEEE 19th international conference on industrial informatics (INDIN). 2021, p. 1–6. https://doi.org/10.1109/INDIN45523.2021.9557368

[10] Crompton J. Chapter 5 - data management from the DCS to the historian. In: Bangert P, editor. Machine learning and data science in the oil and gas industry. Gulf Professional Publishing; 2021, p. 83–110. https://doi.org/10.1016/B978-0-12-820714-7.00005-4

[11] Jung W, Park J, Kim Y, Choi SY, Kim S. HuREX – a framework of HRA data collection from simulators in nuclear power plants. Reliab Eng Syst Saf 2020;194:106235. . https://doi.org/10.1016/j.ress.2018.05.011

[12] Park J, Boring RL, Ulrich TA, Lew R, Lee S, Park B, et al. A framework to collect human reliability analysis data for nuclear power plants using a simplified simulator and student operators. Reliab Eng Syst Saf 2022;221:108326. https://doi.org/10.1016/j.ress.2022.108326

[13] Srinivasan R, Iqbal MU, Srinivasan B. Multi-class classification of control room operators' cognitive workload using the fusion of eye-tracking and electroencephalography. Comput Chem Eng 2024;181:108526. https://doi.org/10.1016/j.compchemeng.2023.108526

[14] Iqbal MU, Srinivasan B, Srinivasan R. Dynamic assessment of control room operator's cognitive workload using electroencephalography (EEG). Comput Chem Eng 2020;141:106993. . https://doi.org/10.1016/j.compchemeng.2019.106993

[15] Ikuma LH, Harvey C, Taylor CF, Handal C. A guide for assessing control room operator performance using speed and accuracy, perceived workload, situation awareness, and eye tracking. J Loss Prev Process Ind 2014;32:454–65. https://doi.org/10.1016/j.jlp.2014.11.004

[16] Schmidt B, Borrison R, Gärtler M, Maczey S, Kotriwala A. Practical aspects for exploration and analysis of manual interventions in process plants. In: 2021 IEEE 19th international conference on industrial informatics (INDIN). IEEE; 2021, p. 1–6. https://doi.org/10.1109/INDIN45523.2021.9557368

[17] Garg V, Kumar A, Garg S, Vashisht V, Anand D. Machine learning based predictive model to enhance human reliability analysis for risk assessment of nuclear reactors. Int J Syst Assur Eng Manage 2025;https://doi.org/10.1007/s13198-025-02960-9

[18] Groth KM, Smith R, Moradi R. A hybrid algorithm for developing third generation HRA methods using simulator data, causal models, and cognitive science. Reliab Eng Syst Saf 2019;191:106370. https://doi.org/10.1016/j.ress.2019.106370

[19] Morais C, Yung KF, Johnson K, Moura R, Beer M, Patelli E. Identification of human errors and influencing factors: a machine learning approach. Saf Sci 2022;151:105741. https://doi.org/10.1016/j.ssci.2021.105741

[20] Liu J, Zou Y, Wang W, Zhang L, Liu X, Ding Q, et al. Analysis of dependencies among performance shaping factors in human reliability analysis based on a system dynamics approach. Reliab Eng Syst Saf 2021;215:107890. https://doi.org/10.1016/j.ress.2021.107890

[21] Liu J, Zou Y, Wang W, Zio E, Yuan C, Wang T, et al. A Bayesian belief network framework for nuclear power plant human reliability analysis accounting for dependencies among performance shaping factors. Reliab Eng Syst Saf 2022;228:108766. https://doi.org/10.1016/j.ress.2022.108766

[22] Kim Y, Park J. Incorporating prior knowledge with simulation data to estimate PSF multipliers using Bayesian logistic regression. Reliab Eng Syst Saf 2019;189:210–17. https://doi.org/10.1016/j.ress.2019.04.022

[23] Dorgo G, Palazoglu A, Abonyi J. Alarm rationalization and dynamic risk analyses for rare abnormal events. Comput Chem Eng 2024;183:108616. https://doi.org/10.1016/j.compchemeng.2024.108616

[24] Qing T, Liu Z, Tang Y, Hu H, Zhang L, Chen S. Effects of automation for emergency operating procedures on human performance in a nuclear power plant. Health Phys 2021;121(3):261–70. . https://doi.org/10.1097/HP.0000000000001445

[25] Demichela M, Baldissone G, Camuncoli G. Risk-based decision making for the management of change in process plants: benefits of integrating probabilistic and phenomenological analysis. Ind Eng Chem Res 2017;56(50):14873–87. https://doi.org/10.1021/acs.iecr.7b03059

[26] Amazu CW, Mietkiewicz J, Abbas AN, Briwa H, Alonso-Perez A, Baldissone G, et al. Exploring the influence of human system interfaces: introducing support tools and an experimental study. Int J Hum–Comput Interact 2024;0(0):1–18. https://doi.org/10.1080/10447318.2024.2376354

[27] Mietkiewicz J, Abbas AN, Amazu CW, Baldissone G, Madsen AL, Demichela M, et al. Enhancing control room operator decision making. Processes 2024;12(2):328. https://doi.org/10.3390/pr12020328

[28] Amazu CW, Mietkiewicz J, Abbas AN, Briwa H, Alonso Perez A, Baldissone G, et al. Experiment data: human-in-the-loop decision support in process control rooms. Data Brief 2024;53:110170. . https://doi.org/10.1016/j.dib.2024.110170

[29] Jensen FV, Nielsen TD. Bayesian networks and decision graphs. Springer New York, NY; 2007. https://doi.org/10.1007/978-0-387-68282-2

[30] Kjærulff UB, Madsen AL. Bayesian networks and influence diagrams: a guide to construction and analysis. Springer New York, NY; 2013. . https://doi.org/10.1007/978-1-4614-5104-4

[31] Fayyad UM, Irani KB. Multi-interval discretization of continuous-valued attributes for classification learning. In: IJCAI; vol. 93. Citeseer; 1993, p. 1022–9.

[32] Liu H, Hussain F, Tan CL, Dash M. Discretization: an enabling technique. Data Min Knowl Discov 2002;6:393–423. . https://doi.org/10.1023/A:1016304305535

[33] Shapiro SS, Wilk MB. An analysis of variance test for normality (Complete samples). Biometrika 1965;52(3/4):591–611. https://doi.org/10.2307/2333709

[34] Levene H. Robust tests for equality of variances. In: Contributions to probability and statistics: essays in honor of harold hotelling. Stanford University Press; 1960, p. 278–92. https://doi.org/10.2307/2984135

[35] Student. The probable error of a mean. Biometrika 1908;6(1):1–25. https://doi.org/10.2307/2331554

[36] Welch BL. The generalization of 'student's' problem when several different population variances are involved. Biometrika 1947;34(1-2):28–35. https://doi.org/10.2307/2332510

[37] Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat 1947;18(1):50–60.

Chidera Winifred Amazu, a PhD student and research assistant at Politecnico di Torino, is part of the Collaborative Intelligence for Safety-Critical Systems network. Her research centres on the HMI for situational awareness, decision support of control room operators, and safety analysis. She studies operator behaviour and cognition during human-machine interaction.

Joseph Mietkiewicz, PhD, working at TU Dublin, Ireland, contributes to the Collaboration Intelligence for Critical Safety Systems project. Specialising in interpretable machine learning and Bayesian networks, he focuses on developing decision support systems for control room operators.

Ammar N. Abbas, PhD in Deep Reinforcement Learning, focuses on optimal decision-making in safety-critical systems (detecting anomalies and prescriptive maintenance, optimising the product quality, process scheduling) He utilises Reinforcement Learning as an online method, integrating human expertise with explainable and interpretable AI.

Gabriele Baldissone, is currently a Research Technician and Lecturer at Politecnico di Torino, Italy. He educates undergraduate and Master's students in advanced control and environmental safety techniques, as well as advanced technologies for risk-based decision-making. His expertise spans risk analysis, assessment, modelling, quantitative analysis, reliability, data mining, and process safety.

Davide Fissore, Full Professor of Process Control and Food Processing Technologies at Politecnico di Torino, specialises in process modelling and development of advanced model-based tools for process monitoring and control. His recent research extends to pharmaceutical engineering, focusing on applying the 'Quality by Design' concept to optimise, monitor and control pharmaceutical freeze-drying processes.

Micaela Demichela, Full Professor at Politecnico di Torino, is actively involved in teaching and research and holds positions on the Board of Directors for the ESReDa Association and the Scientific Board for the R3C Inter-Department Centre. Her expertise lies in risk management, occupational safety, probabilistic risk assessment, etc.

Anders L. Madsen Anders L Madsen, Full Professor at the Department of Computer Science at Aalborg University (part-time) and Chief Executive Officer at HUGIN EXPERT A/S, contributes to the Collaboration Intelligence for Critical Safety Systems project. His expertise is in algorithms for Bayesian networks and applications of Bayesian networks.

Maria Chiara Leva, co-chair of the Human Factors Technical Committee for the European Safety and Reliability Association, previously chaired the Irish Ergonomics Society and co-chaired the Symposium on Human Mental Workload. She co-founded Tosca Solutions and currently lectures at TU Dublin. Her expertise spans Human Factors and Safety Management Systems.