Article | Received 17 February 2025; Accepted 6 June 2025; Published 3 July 2025
https://doi.org/10.55092/sc20250016

A new hybrid inference model for human performance reliability prediction: a case study of construction workers

Yamo Cao, Zeren Jin and Yuguang Fu*

School of Civil and Environmental Engineering, Nanyang Technological University, Singapore

* Correspondence author; E-mail: yuguang.fu@ntu.edu.sg.

Highlights:

- Proposes a hybrid BN-SOM model for human reliability prediction.
- Utilizes a fuzzy CHRS to quantify worker performance conditions and reduce subjective bias.
- Offers a structured framework integrating expert evaluations and worker self-reports for validation.
- Demonstrates superior predictive accuracy and specificity over traditional CREAM models.

Abstract: Human performance reliability is crucial in the construction industry, characterized by complex socio-technical systems and a high incidence of workplace accidents. Traditional human performance models often rely on expert experience, complicating effective validation. This study proposes a comprehensive framework for collecting worker self-reports and expert evaluations, providing a robust approach to validate the Cognitive Reliability and Error Analysis Method (CREAM) model. In particular, Common Performance Conditions (CPCs) are quantified based on workers' self-assessment data, utilizing a fuzzy-based Contextual Human Reliability Score (CHRS). Expert evaluations serve as the ground truth, providing the criteria for human performance classification. Furthermore, a novel hybrid inference model is designed and built on the data collection framework to predict human performance reliability among construction workers. This model integrates Bayesian Networks (BNs) and Self-Organizing Maps (SOM) to address complex and nonlinear relationships between CPCs. A case study is conducted on a construction site to validate the proposed model, demonstrating its ability to generate reliable predictions of human performance failures. The results show that conventional CREAM models fail to predict human performance failures within the context of our dataset. In comparison, the proposed hybrid inference model exhibits significant improvements, particularly in terms of accuracy and specificity. This hybrid inference model offers valuable insight into human reliability and contributes to enhancing safety and operational efficiency in the construction industry.

Keywords: workplace safety and health; human reliability analysis; CREAM; fuzzy theory; Bayesian network; self-organizing maps

1. Introduction

Human factors have become increasingly important in ensuring the safety of complex socio-technical systems in recent years [1]. To assess the influence of human factors on system failures, Human Reliability Analysis (HRA) has been developed as a dynamic and multifaceted field with qualitative and quantitative approaches [2]. Due to its extensive human involvement, the construction industry is recognized as one of the most hazardous sectors, exhibiting a higher mortality rate than other industries [3]. Human error, closely related to human reliability, causes most workplace accidents, leading to significant physical and financial losses [4]. HRA offers a framework for probabilistic risk assessments to quantify human error probabilities. Additionally, HRA identifies human failure events, offering insight into the causes, interventions, and consequences of errors.

Over the past several decades, numerous HRA methods have been proposed to assess human performance and error potential in complex systems. These techniques are divided into two generations. First-generation HRA methods, developed in the 1970s and 1980s, rely on expert judgment and quantitative data to estimate Human Error Probabilities (HEPs). These methods remain prevalent in industries, with examples like the Technique for Human Error Rate Prediction (THERP) [5,6] and the Human Error Assessment and Reduction Technique (HEART) [5]. In contrast, second-generation HRA methods aim to overcome the limitations of their predecessors by incorporating cognitive modeling and considering the situational and psychological factors that contribute to human errors. Prominent examples of these methods include the Cognitive Reliability and Error Analysis Method (CREAM) [7] and the Systematic Human Error Reduction and Prediction Approach (SHERPA) [8]. Developed by Erik Hollnagel in 1998, CREAM, a leading second-generation method, highlights cognitive processes in decision-making and contextual factors, making it valuable for analyzing human performance in dynamic environments. Nonetheless, the original CREAM model has limitations: (1) the basic CREAM model's human error rates have a wide interval, which is unsuitable for precise quantitative failure analysis [9]; (2) the assessment of Common Performance Condition (CPC) levels in the original CREAM relies heavily on expert judgment, but lacks specific standardized rules for evaluation [10]. (3) The original CREAM assumes that each CPC equally affects HEP [11], oversimplifying real-world CPC interactions.

To address the limitations of the original CREAM model, subsequent research has focused on two main fronts: enhancing CPC quantification under uncertainty, and modeling CPC interdependencies more rigorously. First, capturing expert judgments on CPCs has proven challenging, so many studies have adopted uncertainty-handling methods [9,12–15]. Fuzzy set theory has been widely used to translate qualitative expert scores into continuous membership values [13,15], while extensions such as fuzzy D-numbers further accommodate the ambiguity inherent in expert assessments [14]. These approaches yield more nuanced CPC representations, improving both accuracy and adaptability across diverse contexts. Second, understanding how CPCs jointly influence human performance has led to the use of structured multi-criteria methods. Techniques like the Analytic Hierarchy Process (AHP) [16], Fuzzy AHP [13], and DEMATEL [15] quantify the relative importance and interdependencies of CPCs, thereby capturing their combined effects on reliability. Recently, advanced probabilistic and dynamic models have been integrated with CREAM to provide comprehensive, quantitative analyses. Bayesian Networks (BNs)

map CPCs to control modes with explicit uncertainty propagation [17], while fault-tree and event-tree analyses [18,19], Markov models [13], and Petri nets [20] offer alternative frameworks for modeling human reliability under varying scenarios. Together, these hybrid methods extend CREAM's reach, enabling more precise and context-sensitive human performance assessments [21].

Building on these advances in uncertainty quantification and CPC interdependence analysis, a growing body of work has adopted a fully hybridized CREAM framework—combining fuzzy logic, BNs, and the original CREAM control-mode logic—to leverage the strengths of each component. Zhang and Tan [22], introduced trapezoidal fuzzy membership functions to translate expert linguistic judgments on CPCs into continuous inputs for a BN, achieving smoother uncertainty propagation and tighter human error–probability estimates in LNG-terminal power-supply operations. Ung [18] extended this by embedding fuzzy CPC scoring into a BN for high-temperature molten-metal tasks, enabling real-time HEP updates as process conditions shifted. Fan *et al.* [23] demonstrated that in an LNG bunkering context, fuzzy–BN integration can incorporate site-specific sensor and operational data—something pure CREAM cannot—thereby delivering faster point estimates under dynamic safety constraints. Shirali *et al.* [24], showed how a petrochemical control-room case study benefits from a hybrid that supports interactive "what-if" querying of CPC scenarios, thanks to the BN's graphical interface. Sezer *et al.* [25] further illustrated that fuzzy–BN models can be readily adapted across maritime loading operations, preserving interpretability of each CPC→control-mode link while improving predictive performance. Among these hybrid methodologies, the BN component stands out for its intuitive graphical structure, its ability to fuse empirical data, theoretical insights, and expert judgments, and its robust probabilistic treatment of uncertainty in human-error modelling [26]. Fuzzy logic plays a pivotal role in evaluating CPCs by capturing the inherent vagueness and ambiguity of context assessments, thus yielding more nuanced quantifications of Performance-Shaping Factors (PSFs) [18]. However, current fuzzy–BN–CREAM hybrids suffer three main drawbacks: No independent ground-truth validation, relying solely on expert-elicited CPC scores; rigid network topology, inherited directly from CREAM's fixed CPC to mode rules and unable to adapt to data-specific interdependencies; expert-driven conditional probability tables (CPTs), which—even when fuzzified—remain disconnected from empirical performance data.

Supervisor and manager evaluations offer a relatively more objective alternative to worker self-reports, as they are conducted by trained third-party observers who do not rate their own behaviour. Comparative studies show that supervisor assessments are systematically more discriminating and less prone to self-enhancement biases than trainees' self-ratings [27,28]. Psychometric analyses further reveal that self-appraisals tend to exhibit greater leniency and lower variability compared to supervisor ratings, evidencing pervasive self-report biases in safety behaviour measures [29]. Investigations of method bias in safety research demonstrate how impression management can inflate self-reported safety constructs, whereas supervisor observations help mitigate these distortions [30].

To fill the gap in prior CREAM research—where few studies have validated their models against an independent ground truth—we introduce a unified framework that brings together expert evaluations, worker self-assessments, and structured field observations. At its core is the Contextual Human Reliability Score (CHRS), which applies fuzzy logic and rank-standardization to harmonize these diverse inputs and curb subjective bias. Over three weeks of on-site monitoring, we

gathered detailed CPC and behavioral data to benchmark our baseline model. When we observed that the original approach failed to capture the nonlinear interdependencies among CPCs, we enhanced it with unsupervised clustering: a Self-Organizing Map (SOM) front end learns CPC clusters directly from the data, and these empirically derived patterns then feed into a Bayesian Network for transparent probabilistic inference. In addition to yielding reliable control-mode classifications, the SOM–BN pipeline produces data-driven CPTs that replace purely theoretical or expert-elicited parameters. By clustering CPCs first, the BN captures their complex interdependencies, yet each directed edge still corresponds to a CPC-cluster to control-mode relationship, preserving full interpretability. These empirically grounded CPTs not only explain the observed joint distribution of clusters and control modes but also support targeted "what-if" analyses and actionable insights for reducing human error in construction environments.

2. Methodology

This section presents a comprehensive methodology to predict the varying levels of reliability of worker performance in a specific area. The methodology consists of two components: a generic framework for evaluating the human performance reliability within a given context, and an enhanced BN-based hybrid inference model for predicting workers' performance. Figure 1 illustrates the proposed method's analysis process, focusing on three key phases: data acquisition, data preprocessing, and the prediction phase. In Section 2.1, we describe the data collection and preprocessing process, aimed at creating input and ground truth data for model training. This process includes both objective and subjective data collection methods, along with a novel approach to preprocessing subjective evaluations to derive a context-sensitive, bias-free subjective score. Section 2.2 introduces the enhanced BN-based hybrid inference model, designed to improve the prediction of human performance reliability.

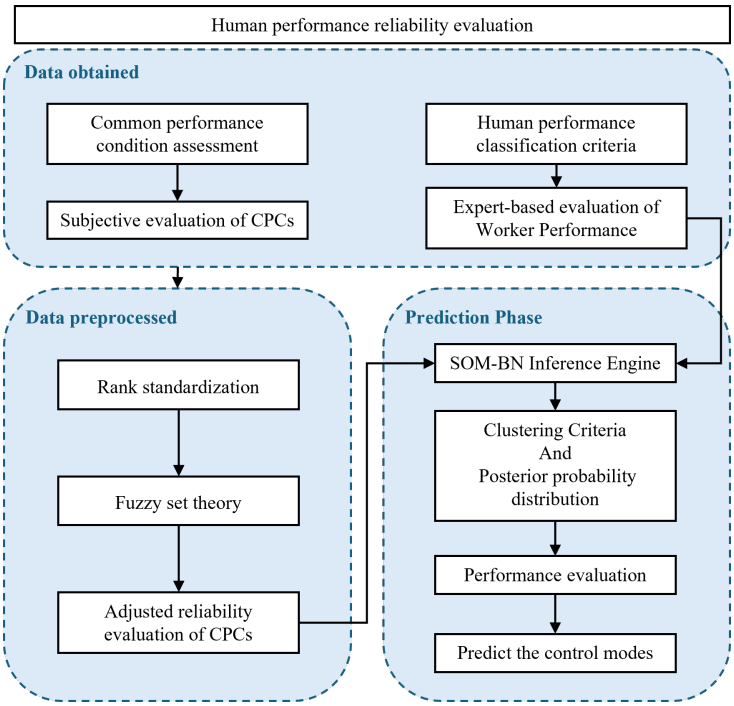


Figure 1. Overview of human performance reliability evaluation method.

2.1. Framework of worker data collection and processing

The purpose of this framework is to provide a comprehensive assessment of human performance reliability within a specific group of workers. This evaluation incorporates both subjective and objective assessments. Subjective assessment, which functions as input for the prediction phase, is expressed through a CHRS, derived from workers' self-reports based on CPCs and pre-processed using the standard rank set method and fuzzy logic. This assessment provides a context-sensitive, subjective measure of human performance, reflecting the unique characteristics and conditions of the data collection population. Objective assessment, which serves as the ground truth for the prediction phase, is represented through various levels of performance reliability. These levels are defined by control modes and are obtained from supervisors and managers of the worker group. This exper-based assessment provides a robust foundation for the evaluation of predictions generated through the CHRS and guarantees conformity with practical performance benchmarks. In this section, we will illustrate the application of this framework using the example of construction workers.

The subjective assessment framework involves three steps. Initially, it is essential to identify the scope of measurement while leveraging prior knowledge to refine the CHRS. A questionnaire is developed from the refined CHRS for self-reported evaluations. Finally, the collected self-reported data undergo preprocessing to derive the CHRS, ensuring that the assessment is accurate and reliable.

2.1.1. Common Performance Condition assessment applicable to construction

The original CREAM identifies nine factors known as Common Performance Conditions (CPCs) that influence human behavior reliability within a work environment. CPCs are based on a combination of theoretical insights and empirical observations that link specific conditions to human performance outcomes. They are grounded in the understanding that performance is context-dependent and influenced by a range of factors that can enhance or impair human reliability. CPCs provide a framework to classify error modes (how an action can go wrong) and their causes, addressing both the cognitive aspects of human behavior and the context in which performance occurs. Each CPC can influence performance differently depending on the task at hand. Therefore, the original CPCs can be adjusted to better suit specific sites or tasks [9]. When formulating CPCs, different industries tailor them based on specific operational environments and task requirements. Expert interviews, incident analysis, and simulation experiments are commonly used to identify high-risk CPCs. In the construction sector, risks arise from various sources, including work mix and management, task execution, and personal factors [3,31]. Following a literature review and a series of semi-structured interviews, we have refined the CPC framework into nine categories tailored to the construction context in Table 1.

The self-reported questionnaire survey results capture workers' perceptions of targeted scenarios relevant to the study. As previously outlined, the nine developed CPCs provide a foundational description of human performance, providing detailed guidelines for structuring the questionnaire [13,32]. The questionnaire consists of nine sections and is designed to correspond directly to the nine CPCs listed in Table 1. Respondents are asked to evaluate their level of agreement with each statement using a five-point Likert scale: "strongly agree" (5), "agree" (4), "generally agree" (3), "disagree" (2), and "strongly disagree" (1). Detailed descriptions of the questionnaire items pertinent to the construction industry are

provided in Appendix. To ensure the internal consistency of the scale, a reliability analysis is performed using the Cronbach's α. A Cronbach's α value of ≥ 0.7 is generally considered acceptable, while a value of ≥ 0.8 indicates high reliability. In this study, Cronbach's α is calculated using the Pingouin statistical package [33], yielding a coefficient of 0.81 with the 95% confidence interval ranging from 0.72 to 0.88. This result confirms the high reliability of the scale used in the survey.

Table 1. CPCs and their levels, and effects on human reliability. Source: authors, adapted from Hollnagel (1998) [7].

No.	CPCs	Description	CPC Levels	Effects
C1	Training and experience	The quality and effect of training, the knowledge, skills, and experience of the work-players	Adequate, high experience Adequate but limited Inadequate	Increase Neutral Decrease
C2	Physical environment	Physical factors of the work environment, such as weather, temperature, humidity, lighting, noise, precipitation, topography	Advantageous Acceptable Incompatible	Increase Neutral Decrease
C3	Organizational management	The process of organizing, planning, leading and organization of personnel during the construction process. The effect of organization, supervision and management.	Very efficient Acceptable Deficient	Increase Neutral Decrease
C4	Work characteristics	Work stress, task characteristics during construction	Satisfied Acceptable Deficient	Increase Neutral Decrease
C5	Available time	Construction duration, delivery lead time constraints. Dynamic time to respond to various contingencies.	Adequate Temporarily inadequate Continuously inadequate	Increase Neutral Decrease
C6	Contract information	Such as contract duration, salary situation, job security, security insurance	Satisfied Acceptable Unsatisfied	Increase Neutral Decrease
C7	Working environment conditions	Availability of ancillary facilities, level of amenity situation, such as tools, equipment, and materials	Satisfied Acceptable Unsatisfied	Increase Neutral Decrease
C8	Coplayers collaboration	The quality of the collaboration between workforces, including the level of trust, communication, and the general social climate among workforces	Efficient Inefficient Deficient	Increase Neutral Decrease
C9	Individual differences	Nationality, race, age, attributes, health status, and fitness for work	Favorable Neutral Unfavorable	Increase Neutral Decrease

The self-reported results derived from workers' subjective awareness and evaluation are inherently prone to bias and contain various forms of uncertainty. In addition, in real-world settings, particularly within homogeneous work environments characterized by uniform procedures and standards, consistent training programs, comparable experience levels, and routine tasks of low complexity, these conditions can lead to a clustering effect. This effect contributes to highly concentrated distributions of CPC scores among certain groups of workers. Furthermore, the use of limited rating scales in assessment methods can further exacerbate this clustering, as the scales may not adequately capture subtle differences in performance, leading to compressed and less differentiated scores. To tackle these challenges, we present a new index to assess human performance in this specific population. This index, termed the Contextual Human Reliability Score (CHRS), provides a refined, context-sensitive measure that more accurately reflects the unique characteristics and conditions of the data collection population.

2.1.2. Contextual Human Reliability Score

For each worker i and CPC j, multiple self-reported responses to questions related to CPC j are collected. These responses are quantified using a 5-point Likert scale, where the linguistic terms "strongly disagree," "disagree," "neither agree nor disagree," "agree," and "strongly agree" are assigned the numerical values of 1, 2, 3, 4, and 5, respectively.

Let q_{ijk} represent the quantified response of worker i to the k^{th} question associated with CPC j, $k = 1, 2, \ldots, Q_j$, where Q_j denotes the total number of questions related to CPC j. The final score for worker i with respect to CPC j is calculated as the arithmetic mean of the responses across all relevant questions. This can be formally expressed as:

$$x_{ij} = \frac{1}{Q_j} \sum_{k=1}^{Q_j} q_{ijk} \tag{1}$$

This approach assumes that each question is equally weighted when calculating the overall CPC score for each worker. Let m denote the number of workers and n denote the number of CPCs. Then a subjective evaluation matrix $X = [x_{ij}] \in \mathbb{R}^{m \times n}$ can be obtained, where each element x_{ij} represents the mean CPC score for worker i and CPC j.

To ensure comparability between different CPCs, scores x_{ij} are first ranked. This ranking process converts each score into a percentile rank within its respective CPC column. The ranking is achieved using the following formula:

$$s_{ij} = \frac{\text{rank}(x_{ij}) - 1}{m - 1} \tag{2}$$

where $\text{rank}(x_{ij})$ represents the rank of x_{ij} among the scores $x_{1j}, x_{2j}, \ldots, x_{mj}$, with the highest score receiving a rank of 1. The standardized rank s_{ij} thus ranges from 0 to 1, providing a normalized representation of each score's position within its CPC.

Each standardized rank s_{ij} is then converted into fuzzy membership values for the three levels of CPCs, *i.e.*, Decrease, Neutral, and Increase—using triangular fuzzy numbers (TFNs). TFNs are widely adopted in CREAM-based human reliability analyses due to their simplicity, computational efficiency, and ease of interpretation, especially when handling limited or uncertain data [34,35]. The membership functions for these levels are defined as follows:

(1) Decrease: This fuzzy set captures the degree to which a score is associated with a decrease in performance.

$$\mu_{\text{Decrease}}(s_{ij}) = \begin{cases} 1 & \text{if } s_{ij} \leq \alpha \\ \frac{\beta - s_{ij}}{\beta - \alpha} & \text{if } \alpha < s_{ij} \leq \beta \\ 0 & \text{if } s_{ij} > \beta \end{cases} \tag{3}$$

(2) Neutral: This fuzzy set represents scores that are neither high nor low, indicating a neutral performance level.

$$\mu_{\text{Neutral}}(s_{ij}) = \begin{cases} 0 & \text{if } s_{ij} \leq \gamma \text{ or } s_{ij} \geq \delta \\ \frac{s_{ij} - \gamma}{\varepsilon - \gamma} & \text{if } \gamma < s_{ij} \leq \varepsilon \\ \frac{\delta - s_{ij}}{\delta - \varepsilon} & \text{if } \varepsilon < s_{ij} \leq \delta \end{cases} \tag{4}$$

where γ, ε, and δ define the triangular fuzzy number parameters for Neutral.

(3) Increase: This fuzzy set represents the degree to which a score is associated with an increase in performance.

$$\mu_{\text{Increase}}(s_{ij}) = \begin{cases} 0 & \text{if } s_{ij} \leq \kappa \\ \frac{s_{ij} - \kappa}{\zeta - \kappa} & \text{if } \kappa < s_{ij} \leq \zeta \\ 1 & \text{if } s_{ij} > \zeta \end{cases} \tag{5}$$

where κ and ζ are the parameters defining the triangular fuzzy number.

Following Zhou *et al.* [36], the universes of discourse for the fuzzy sets corresponding to each CPC are defined as $[0, 0.5]$, $[0.1, 0.9]$, and $[0.5, 1]$ for the *Decrease*, *Neutral*, and *Increase* sets, respectively, with peaks at 0.1, 0.5, and 0.9 (Figure 2). This configuration ensures even coverage of the normalized range $[0, 1]$. To assess the overall effect, the weighted fuzzy membership values are aggregated. The elements of the CHRS are calculated using the following formula:

$$CR_i = \sum_k \mu_k(s_{ij}) \cdot \eta_k \tag{6}$$

where $\mu_k(s_{ij})$ represents the membership value of s_{ij} for the fuzzy set k, and η_k is the effect index assigned to the fuzzy set membership. The effect index η_k is defined as follows:

$$\eta_k = \begin{cases} +1 & \text{for "Increase"} \\ 0 & \text{for "Neutral"} \\ -1 & \text{for "Decrease"} \end{cases} \tag{7}$$

In the above equations, $\mu_k(s_{ij})$ reflects how strongly the standardized rank s_{ij} belongs to the fuzzy set k (Increase, Neutral, or Decrease), while η_k quantifies the effect associated with each fuzzy set. The aggregation of these weighted fuzzy membership values provides the final Cognitive Human Reliability score CR_i for each worker i, reflecting the overall performance effect in the context of the CPC evaluated.

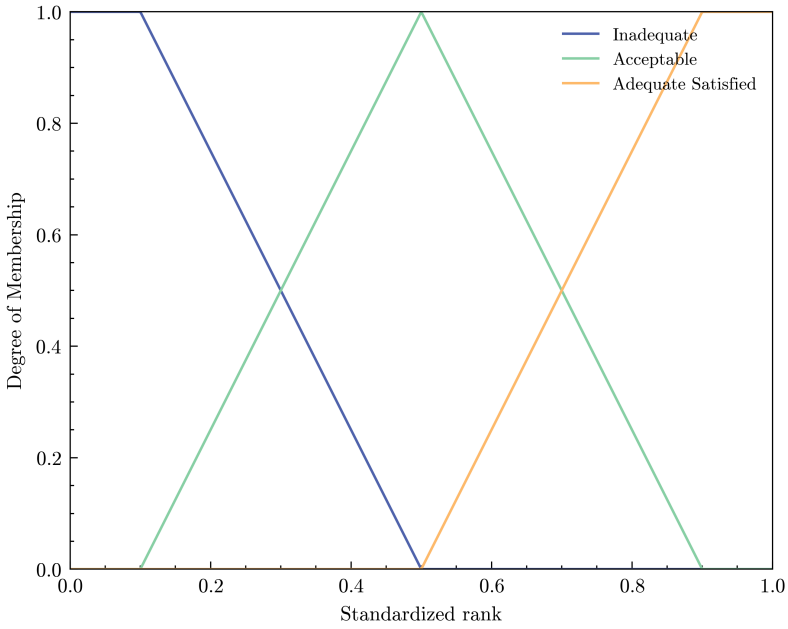


Figure 2. Membership functions for each CPC.

2.1.3. Control modes

The main task to obtain an expert-based objective assessment is to establish human performance classification criteria based on control modes. Human performance classification is based on the CREAM methodology, derived from the Contextual Control Mode (COCOM) framework. The ultimate goal of the original CREAM methodology is to determine the control modes that characterize human performance in various contexts. According to Hollnagel [7], the degree of control in a given situation can be described by four characteristic control modes: Strategic, Tactical, Opportunistic, and Scrambled. Each mode represents a different level of reliability and has an associated probability interval, as outlined in Table 2.

Table 2. Control modes and probability intervals (Hollnagel, 1998) [7].

Control Mode	Description	Probability Interval
Strategic Control	Strategic Control involves considering the global context with a broader time horizon and focusing on higher-level goals, resulting in a more efficient and robust approach.	(0.000005, 0.01)
Tactical Control	Tactical Control relies on planning based on established procedures or rules, addressing a more limited scope and sometimes limited in its adaptability.	(0.001, 0.1)
Opportunistic Control	Opportunistic Control determines the next action based on the prominent features of the current context with minimal planning or anticipation, often due to unclear contexts or time constraints.	(0.01, 0.5)
Scrambled Control	Scrambled Control is characterized by unpredictable or haphazard decision-making with little to no deliberate thinking involved, resulting in a random approach.	(0.1, 1.0)

To ensure the reliability and objectivity of the "ground-truth" control-mode labels, a rigorous observer protocol was implemented: four experts from Hwa Seng Builders Pte Ltd. (site managers and safety-team members) completed a standardized training program covering explicit definitions and classification criteria for control modes based on Hollnagel's CREAM framework [7], the daily observation schedule and use of a standardized behavior-checklist drawn from the Centre to Protect Workers' Rights (CPWR) Construction Solutions database [37] and aligned with Hussain *et al.*'s safety-training-evaluation procedures [38] which catalogues 26 performance-shaping hazards, and protocols to minimize common rater biases (e.g., halo and recency effects); next, each worker was observed independently by all four experts over a three-week period to mitigate individual subjectivity and to allow calculation of inter-rater reliability (Cohen's $\kappa > 0.75$, indicating substantial agreement), with periodic calibration sessions—using video-based scenarios and refresher workshops—ensuring consistency throughout the data-collection period.

2.2. *Enhanced BN-based hybrid inference model*

In the CREAM methodology, CPC scores are crucial to determine the expected control mode by evaluating whether CPCs enhance or diminish performance reliability. To construct a predictive model using this approach, two fundamental challenges must be addressed: (1) understanding the intricate relationships between CPCs and (2) mapping the combined CPC scores onto the corresponding control modes. As shown in Figure 3, the proposed hybrid inference framework comprises two main stages after data preprocessing and CHRS computation: unsupervised clustering and Bayesian network inference.

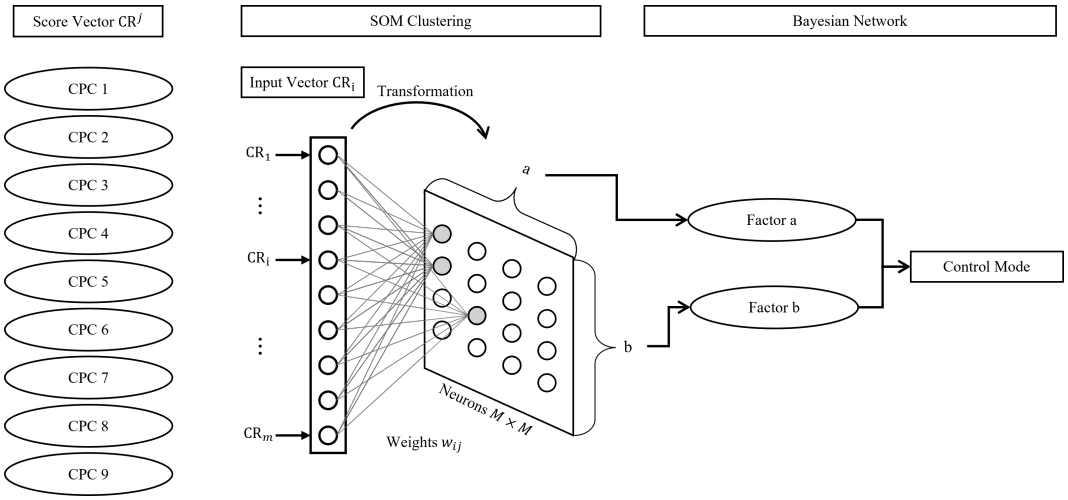


Figure 3. SOM-BN framework flowchart for control mode classification.

2.2.1. Unsupervised clustering

A significant challenge in HRA is the scarcity of objective data and the limited information on human performance at the lower level that can be collected from real-world events. Traditionally, the relationships between CPCs have been defined using empirical and theoretical models. However, our collected data and previous CREAM models, which relied solely on BN inference, have proven inadequate for precise predictions. This inadequacy suggests that the relationships between CPCs are more complex and nonlinear. The SOM [39] is an unsupervised machine learning technique that can be utilised for

exploratory data analysis and pattern recognition, and is capable of handling non-linear problems. SOM allows for nonlinear mapping of all CPCs, transforming high-dimensional data into a two-dimensional space while preserving the original data structure.

SOM operates by mapping high-dimensional input data onto a lower-dimensional (typically two-dimensional) grid. This mapping is accomplished by training a network of artificial neurons to cluster similar data points together. Each neuron's weight vector adapts during training to represent the data it encounters.

The input and output layer neurons define an SOM network. Before constructing the SOM, it is important to standardize the input features to a mean of zero and a variance of one. This prevents features with larger ranges from dominating the learning process. Once standardized, input neurons match the Score Vector features CR^j. Choosing the number of output neurons is complex and does not follow fixed rules. It crucially determines how compressed the experimental data will be, assessed by comparing the number of samples with the number of nodes. Balance is crucial: too many nodes cause underutilization with empty nodes, while too few may lose data structure, making the SOM like k-means clustering. In the proposed model, we suppose that the SOM output layer is designed as a two-dimensional grid with a 1:1 aspect ratio, composed of $M \times M$ neurons.

In the self-organization phase of the SOM algorithm, each input vector is processed to adjust the weight vectors of the Kohonen output map. The process involves calculating the distance between the input vector and the weight vectors of the output nodes. To determine how closely each output node's weight vector matches the input vector, a distance metric is computed. We use the Euclidean distance, which is given by:

$$\text{Distance} = \sqrt{\sum_{i=1}^{n} (CR_i - w_{ji})^2} \tag{8}$$

where CR_i represents the components of the input vector, and w_{ji} represents the components of the weight vector for the j-th output node. The output node with the smallest distance from the input vector is identified as the Best Matching Unit (BMU). The BMU and its neighbors are adjusted to minimize the distance between their weight vectors and the input vector. The learning rate, which decreases over time, controls the magnitude of adjustments, allowing early iterations to shape the global structure and later ones to refine local details.

The neighborhood function $h_j(t)$ defines the range of influence of the BMU on its surrounding nodes. Initially, the neighborhood size includes about half of the nodes in the Kohonen output map and decreases linearly to a size of one by the end of training. This function is defined by a Gaussian function:

$$h_j(t) = \exp\left(-\frac{d_j^2}{2\sigma(t)^2}\right) \tag{9}$$

where d_j is the distance between the BMU and the j-th node, and $\sigma(t)$ is the neighborhood radius.

After training, the SOM produces a grid with $M \times M$ nodes where each neuron represents a cluster of similar input vectors. The clusters identified by the SOM are used as inputs for the BN. These clusters are defined by two dimensions, which corresponded to two factors, labeled Combined Factor a and Combined Factor b.

2.2.2. Bayesian Network

A Bayesian Network (BN) is a probabilistic graphical model that encodes random variables and their conditional dependencies in a directed acyclic graph (DAG). In conventional CREAM-based HRA, the DAG structure is hand-crafted from Hollnagel's CPC→control-mode rules and CPTs are populated via expert judgment or fuzzy mappings. In this work, the BN topology is learned from data via the SOM output: each of the $M \times M$ SOM neurons defines a discrete "CPC cluster" parent node, so that multivariate CPC interactions are captured by grouping similar input vectors into a single variable.

Once the SOM–driven structure is fixed, CPTs are initialized uniformly and then updated via maximum-likelihood estimation on the combined dataset of SOM clusters and expert-rated control modes. This data-driven parameter learning replaces purely theoretical CPTs, ensuring that the network's quantitative relationships reflect the empirical joint distribution observed during field validation.

Exact inference employs the Variable Elimination algorithm, a general exact inference method for probabilistic graphical models that systematically marginalizes out non-query variables to compute posteriors. The core query of interest is the conditional probability of each control-mode given an observed SOM cluster:

$$P\big(\text{Control Mode} \mid \text{CPC Cluster} = k\big) \tag{10}$$

where CPC Cluster is a discrete random variable whose states $\{0, 1, \ldots, M^2 - 1\}$ index the $M \times M$ SOM neurons, each representing a cluster of similar CPC profiles.

Because the SOM grid induces a sparse BN structure, the treewidth remains small and intermediate factors stay manageable, yielding polynomial-time exact inference in practice. The final output is a set of data-driven CPTs and posterior distributions that (a) quantify $P(\text{Control Mode} = m \mid \text{CPC Cluster} = k)$ for each mode m, and (b) support downstream analyses such as sensitivity studies, scenario evaluation, and targeted safety recommendations.

2.2.3. Evaluation metric

To evaluate the performance of the SOM-BN model in predicting human performance potential, six key metrics are used: Accuracy, Precision, Recall, F1 score, Specificity, and the area under the receiver operating characteristic curve (AUC of ROC).

Accuracy measures the proportion of correctly predicted instances, indicating the model's overall effectiveness. Precision is the ratio of true positive predictions to all positive predictions, showing the model's ability to avoid false positives. Recall (sensitivity) captures the proportion of actual positive instances correctly identified, reflecting the model's capacity to detect true positives. The F1 score, the harmonic mean of Precision and Recall, provides a balanced performance metric. Specificity measures the proportion of true negatives correctly identified, important when false positives have high costs. The ROC curve plots the True Positive Rate (TPR) against the False Positive Rate (FPR), illustrating the trade-off between sensitivity and specificity. The AUC of the ROC curve quantifies overall performance, with values closer to 1 indicating better discrimination between positive and negative instances.

3. Case study

3.1. Dataset description

Participants were selected using a convenience sampling technique. A face-to-face questionnaire survey was conducted at the main campus construction sites of Nanyang Technological University with the support of Hwa Seng Builder Pte Ltd. Potential participants were construction workers who worked on construction sites during our visit and were invited to complete the questionnaire. A total of 49 workers participated in this study between April 2023 and May 2023. The demographic information of the participants, including age, gender, and work experience in the construction industry, was collected, shown in Table 3. All participants were male. More than 70% of the participants were between 26–45 years of age, with the largest age group being 36–45 years, representing 40.82% of the total. Moreover, 76.60% of the participants had at least one year of work experience in the construction industry, indicating that a majority had some level of experience in the field.

Table 3. Summary of the demographic information of participants (n = 49).

Description	Frequency	Percentage (%)
Age (years)	10	20.41
	15	30.61
	20	40.82
	4	8.16
Work experience (years)	11	23.40
	14	29.79
	11	23.40
	11	23.40

3.2. Subjective assessment

All 49 workers completed the CPCs self-reported questionnaire. After filtering out those who incorrectly answer the validation question, 35 valid responses are retained for analysis. The mean score for each CPC section is calculated, and the corresponding distributions are illustrated in Figure 4. The density plot reveals that certain CPCs exhibit skewed distributions, varying ranges, spreads, and central tendencies. Consequently, ranking the CPCs is warranted, as this approach standardizes the data by emphasizing the relative position of each data point within its distribution rather than its absolute value.

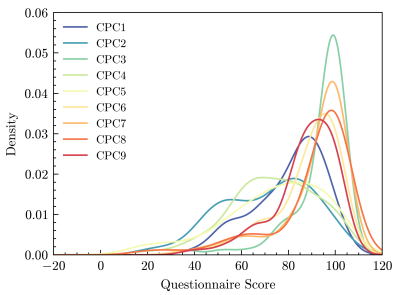


Figure 4. The distribution of mean Common Performance Conditions scores.

3.3. Expert-based assessment

Four experts from Hwa Seng Builders Pte Ltd. (site managers and safety-team members) conducted three weeks of on-site observations, assigning each worker to one of CREAM's four control modes. Each CPC effect was converted to -1 (Decrease), 0 (Neutral), or $+1$ (Increase), and the summed scores across nine CPCs yielded a CREAM score in the range $[-9, 9]$. This range was divided into four equal bins $[-9, -4.5]$, $(-4.5, 0]$, $(0, 4.5]$, and $(4.5, 9]$, producing an unbalanced distribution of 20%, 14.3%, 48.6%, and 17.1% across the Strategic, Tactical, Opportunistic, and Scrambled control modes, respectively.

3.4. Performances of the SOM-BN models

In a BN, the factors (or variables) are generally connected by directed edges that represent probabilistic dependencies. This allows for modeling complex relationships and dependencies between variables. However, the complexity of BN can lead to overfitting and model instability when working with small datasets. To address these challenges in our case study, we hypothesize that factors A and B could be considered independent. This assumption enables us to simplify the BN into a Naive Bayes model.

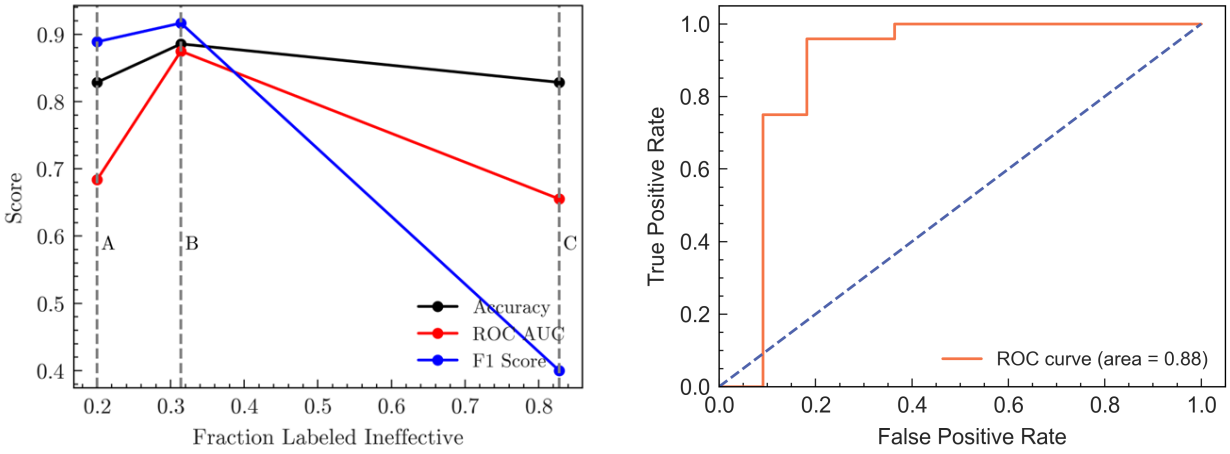
The evaluation metrics are calculated using leave-one-out cross-validation to ensure that each data point is used for both training and validation, providing a robust model performance assessment. Their average, referred to as the mean performance score, is used to assess the effectiveness of the SOM-BN models. Due to the limited dataset and such severe imbalance, to ensure sufficient samples per class, we opt to reduce the control modes from four to two to achieve more reliable and interpretable results in our case. A sensitivity analysis was conducted by evaluating three reclassification schemes:

- Scheme A: {Strategic, Tactical, Opportunistic} *vs.* {Scrambled}
- Scheme B: {Strategic, Tactical} *vs.* {Opportunistic, Scrambled}
- Scheme C: {Strategic} *vs.* {Tactical, Opportunistic, Scrambled}

For each scheme, the SOM-BN pipeline was retrained, and performance metrics including accuracy, ROC AUC, and F1-score were computed using leave-one-out cross-validation(Figure 5a). Scheme B, which grouped Strategic and Tactical as one class and Opportunistic and Scrambled as another, achieved the highest F1-score (0.92), with ROC AUC = 0.88 and accuracy = 0.89. Based on this analysis and to ensure sufficient sample sizes for each class, we adopted Scheme B and collapsed the four control modes into two categories:

- **Effective Control**: Strategic + Tactical
- **Ineffective Control**: Opportunistic + Scrambled

Performance evaluation of the SOM-BN model yields promising results. The ROC curve are shown in Figure 5b. The SOM-BN CREAM Model shows high accuracy (88.57%) with precision and recall both at 91.67%, effectively identifying true positives and minimizing false positives. Its F1 Score also reached 91.67%, showing a strong balance between precision and recall. The ROC curve (AUC = 0.88) further highlights the predictive capability of the model, while its specificity (81.82%) underscores its effectiveness in correctly identifying negative cases (Ineffective Control Mode).



(a) Reclassification Scheme Performance **(b)** ROC curve

Figure 5. SOM-BN model evaluation.

The SOM grid is composed of nodes, where each node represents a "cluster" of data points from the input space mapped during training. The grid size is a critical hyperparameter as it determines the number of prototype vectors used to represent the data. A practical guideline is to set the total number of nodes to approximately five times the square root of the sample size:

$$\text{nodes} \approx 5 \times \sqrt{N_{\text{samples}}} \tag{11}$$

For our $N = 35$ valid observations, this yields $5\sqrt{35} \approx 29.6$, which we round up to a 6×6 grid in order to preserve the data's topological relationships.

To avoid arbitrary training lengths, we employ a convergence criterion based on the change in quantization error QE_t after each epoch. Training is halted when

$$\left| QE_t - QE_{t-1} \right| < 10^{-6} \tag{12}$$

ensuring that additional iterations produce only negligible refinement. This approach minimizes computation while maintaining the integrity of the learned map.This procedure yielded stable convergence by approximately 120–150 epochs in our experiments.

Figure 6 presents the predicted probability of reliability across the SOM nodes generated by BN, demonstrating the classification confidence across the grid for both the efficient control mode and the inefficient control mode. The smooth gradient in color reflects the transition between high and low probabilities, emphasizing the nonlinear relationships captured by the SOM-BN model.

To interpret the classification of workers into control modes, we employ SHAP (SHapley Additive exPlanations), a tool that explains individual predictions by assigning each feature an importance value based on its contribution to the final prediction. SHAP values are derived from cooperative game theory, offering a unified approach to attribute a model's output to the different features involved, which makes it particularly suitable for understanding complex machine learning models such as SOM-BN. The results provide the worker relative performance in different CPCs, as well as their degree of impact on their behavior, and finally provide the suggestion on reinforcements to shift from ineffective control mode to effective control mode.

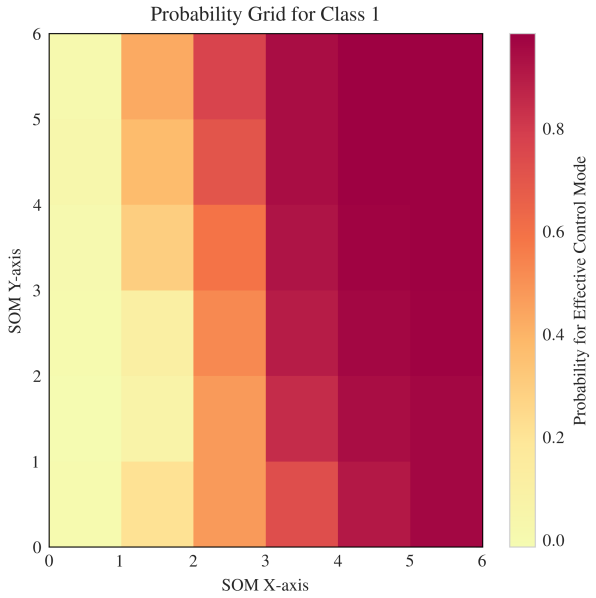
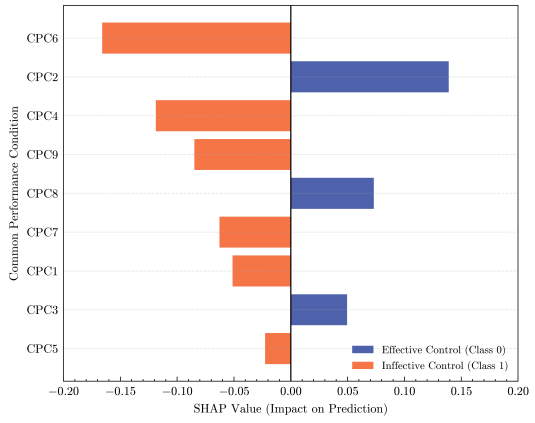


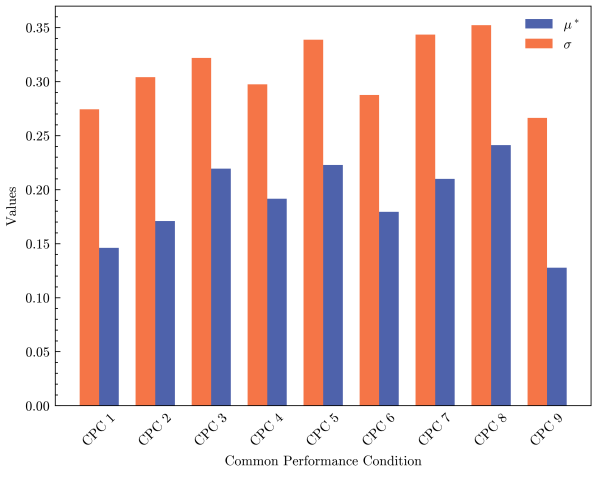
Figure 6. Probability distribution map of SOM nodes for Effective Control Mode classification.

Table 4 shows the results of Worker 1' questionnaire, and CHRS as well as the standardized CHRS. Figure 7a illustrates the SHAP values for a single instance (Worker 1) predicted to belong to the ineffective convergence mode (class 0) with a probability of 99%. The SHAP value contributions of each CPCs are visualized, where blue bars indicate a push towards Ineffective control mode, and orange bars indicate a push towards Effective control mode. The length of the bars reflects the magnitude of the contribution, with longer bars signifying greater influence on the prediction.

Although Worker 1 demonstrates relative strengths in collaboration, management efficiency, and environmental factors (CPC2, CPC3, and CPC8), the model predicts an Ineffective Control Mode. This is mainly due to major deficiencies in contractual conditions (CPC6) and work characteristics (CPC4). To improve safety and effectiveness for Worker 1, addressing issues related to contract security and reducing work stress could shift the prediction towards Effective Control Mode.



(a) Feature Contribution to Worker 1's Control Mode Prediction Using SHAP Values



(b) Sensitivity Analysis Results using Morris method

Figure 7. Local and global feature analysis.

Table 4. Assessment report for worker 1 (Ineffective Control Mode).

CPC	Score	CHRS (Raw)	CHRS (Standardized)
CPC1	80	−0.321429	−0.514792
CPC2	100	1.000000	1.428608
CPC3	100	0.464286	0.707015
CPC4	60	−0.785714	−1.196403
CPC5	76	−0.357143	−0.568723
CPC6	80	−0.785714	−1.193659
CPC7	100	0.571429	0.832340
CPC8	88	−0.750000	−1.306297
CPC9	80	−0.892857	−1.350793

3.5. Sensitive analysis

The sensitivity analysis performed using the Morris [40,41] method reveals significant insights into the factors influencing human performance in this research area. This analysis leverages two key metrics: μ^* (the mean of absolute elementary effects) and σ (the standard deviation of elementary effects), as originally proposed by Morris [41]. μ^* represents the average impact of a variable on the model output, effectively quantifying the overall influence of each factor on the system's behavior. σ reflects the variability of the feature's impact, indicating potential interactions or nonlinearities.

Figure 7b illustrates the results of the sensitivity analysis conducted to evaluate the impact of various parameters on the model output. The results indicate that Collaboration Quality (CPC8) not only has the highest standard deviation ($\sigma = 0.34$) but also the highest adjusted mean ($\mu^* = 0.24$), suggesting that it is highly variable across the dataset and has the most substantial influence on worker performance. The high σ indicates significant differences in the quality of collaboration between workers, likely due to varying work conditions or team dynamics. The high adjusted mean further confirms that Collaboration Quality plays a critical role in influencing performance consistently across workers. Given the importance of Collaboration Quality in this dataset, workers at this construction site experiencing low collaboration quality are more prone to errors, communication breakdowns, and delays, leading to reduced productivity and an increased risk of accidents.

3.6. Comparison with existing models for our dataset

This section presents a comparative analysis of the newly developed SOM-BN model using our collected data, in contrast to established methodologies, specifically the Fuzzy Inference CREAM method and the Hybrid Fuzzy-Bayesian CREAM model. The results are presented in Table 5. For this comparison, we select two representative studies [13,36] and adapt their methodologies to the construction domain through relevant modifications.

Since empirical models inherently do not require training, our analysis focuses on evaluating the predictive accuracy of these models. Additionally, we have enhanced our model by substituting the original input with the CHRS, which has been linearly transformed to a range of -1 to 1 to align with the

input requirements of the Fuzzy Inference and Hybrid Fuzzy-Bayesian CREAM Models. The results are shown in Table 5.

Table 5. Comparison of model performance metrics for Human Reliability Analysis.

Model	Performance Metrics			Specific Metrics		
	Accuracy	AUC	Recall	Precision	F1	Specificity
Fuzzy inference CREAM method [13]	69%	-	100%	69%	81%	0%
Hybrid Fuzzy-Bayesian CREAM Model [36]	71.43%	0.59	100%	70.59%	82.76%	9.09%
Fuzzy inference CREAM method with CHRS	80%	-	100%	77%	87%	36%
Hybrid Fuzzy-Bayesian CREAM Model with CHRS	77%	0.83	100%	75%	86%	27%
SOM-BN CREAM model	88.57%	0.88	91.67%	91.67%	91.67%	81.82%

The Fuzzy Inference CREAM Method and Hybrid Fuzzy-Bayesian CREAM Model achieve perfect recall (100%) by identifying all Effective Control, but they significantly struggle with specificity, failing to recognize Ineffective Control. Both models misclassify a large portion of negative cases as positive, leading to high false-positive rates. The Hybrid Fuzzy-Bayesian model is more accurate (71.43%) with a slight increase in specificity (9.09%) over the Fuzzy Inference method, but both poorly distinguish negative cases. Incorporating the CHRS enhances these models by improving their ability to identify both positive and negative cases. The Fuzzy Inference-CHRS model achieves 80% accuracy and 36% specificity, while the Hybrid Fuzzy-Bayesian-CHRS model records 77% accuracy and 27% specificity. Despite these gains, both models struggle with low specificity and moderate false-positive rates in identifying bad cases.

4. Conclusion

In our study, we established a standardized framework for collecting data that integrated subjective and objective assessments and introduced the CHRS, which integrates fuzzy logic and the standard rank set method. And we also developed a novel hybrid inference model to accurately predict human performance reliability among construction workers.

Previous studies have struggled with validation due to the reliance on subjective data and limited empirical support. Using our framework, we enhanced the validation process by incorporating expert evaluations as objective ground truth, ensuring that the predictions made using CHRS align with real-world performance evaluations by supervisors and managers. Our findings revealed that relying solely on a single BN structure informed by subjective assessments is insufficient for accurately classifying workers' performance. By incorporating unsupervised clustering techniques like the SOM we effectively captured the complex and nonlinear relationships between CPCs. The combination of SOM and BN significantly improved worker performance classification, particularly in cases involving intricate CPC interactions. Importantly, the interpretability of the model was demonstrated through the use of SHAP for local explainability, which provided clear insights into individual workers' strengths and weaknesses. Additionally, the use of Morris sensitivity analysis allowed us to perform a global assessment, identifying the most influential CPCs and highlighting areas for improvement in the construction environment.

Notwithstanding the aforementioned contributions, the study encountered several limitations. The small dataset, with only 35 valid responses, limited the analysis and required a simplification of the control

modes from four (Strategic, Tactical, Opportunistic, and Scrambled) to two (Effective and Ineffective). Although this reduction was essential for reliable analysis, it may have limited the model's predictive granularity. In particular, the CHRS framework relies on self-reported data, and the low validation rate of workers' self-reported questionnaires, *i.e.*, only 35 out of 49 responses were valid, can be attributed to various factors, including the generally low education level among workers, lack of incentives, and emotional or cultural influences. Future data collection efforts must address these issues. Furthermore, the study did not calculate HEP, a crucial component of many HRA models. Incorporating HEP in future research could yield a more comprehensive understanding of error likelihoods, enhancing the predictive accuracy of the model.

Future research should aim to overcome the limitations of this study by expanding the dataset to include a larger and more diverse worker population. Exploring advanced unsupervised learning techniques and refining SOM hyperparameters will further enhance classification performance. Finally, applying this framework to other industries or contexts could demonstrate its broader applicability, contributing to better human reliability assessments across diverse fields.

5. Supplementary data

The appendix provides a detailed overview of the self-reported questionnaire survey. Data supporting the findings of this study are available from the corresponding author upon reasonable request.

Acknowledgements

This research was supported by the Ministry of Education Tier 1 Grants, Singapore (No. RG146/23). We would also like to express our gratitude to Hwa Seng Builder Pte Ltd for their assistance.

Authors' contribution

Conceptualization, Cao Yamo and Fu Yuguang; methodology, Cao Yamo and Jin Zeren; software, Cao Yamo and Jin Zeren; validation, Cao Yamo; formal analysis, Cao Yamo; investigation, Cao Yamo and Jin Zeren; resources, Fu Yuguang; data curation, Cao Yamo; writing—original draft preparation, Cao Yamo; writing—review and editing, Cao Yamo, Jin Zeren and Fu Yuguang; visualization, Cao Yamo; supervision, Fu Yuguang; project administration, Fu Yuguang; funding acquisition, Fu Yuguang. All authors have read and agreed to the published version of the manuscript.

Conflicts of interests

The authors declare that they have no conflicts of interest.

Appendix–CPCs self-reported questionnaire

This appendix presents a detailed overview of the self-reported questionnaire survey, including the specifications and questions utilized in the assessment. As outlined previously, the survey is structured around nine Common Performance Conditions (CPCs), which serve as foundational descriptors of

human performance. These CPCs provide comprehensive guidelines for developing the questionnaire, ensuring its relevance to the construction industry. Each section of the questionnaire corresponds directly to one of the nine CPCs, allowing for a systematic evaluation of the factors influencing construction worker performance.

Table 6. CPCs self-reported questionnaire item specifications (Part 1).

CPC Category	Item No.	Question Description
C1: Training and experience	1.1	Is the training given adequate to do the work you are assigned to?
	1.2	Is the safety training sufficient in highlighting the hazards during work?
	1.3	How often are the safety drills? (Fire evacuation, first aid, accident drills.)
	1.4	How long have you been working as a construction worker?
	1.5	I consider myself an experienced construction worker.
C2: Physical environment	2.1	I often feel it is too hot to work.
	2.2	I often feel it is too stuffy to work.
	2.3	I frequently encounter adverse weather (heavy rain, storms, high temperatures) that delays work.
	2.4	Is nighttime work lighting excessively bright?
	2.5	Are you regularly exposed to loud construction noises?
C3: Organization management	3.1	How often do you observe safety officers conducting site inspections?
	3.2	Is the safety log consistently updated? If so, how frequently?
	3.3	Do managers/supervisors regularly attend site visits or safety briefings?
	3.4	Are organizational processes during construction clear and unambiguous?
	3.5	Rate the effectiveness of onsite organization, supervision, and management.

Table 7. CPCs self-reported questionnaire item specifications (part 2).

Number of CPCs	Questions' number	Description
C4: Work characteristics	4.1	The level of danger to you from the task assigned.
	4.2	How difficult is the assigned task for you?
	4.3	I often feel my workload makes me tired and overwhelmed.
	4.4	I need to give my full attention to my work to protect my personal safety.
	4.5	Please rate how difficult and stressful your work was.
C5: Time management	5.1	I often feel pressured to complete work within schedule.
	5.2	Is there enough dynamic time during construction to respond to unforeseen events?
	5.3	During the construction process, I am often required to work overtime to meet the deadline.
	5.4	I feel like the construction projects I'm on are always tight with deadlines and delivery dates.
	5.5	Please rate how reasonable the timing of the construction project you are working on is.
C6: Personal contract information	6.1	Is your job stable, do you frequently change the construction company you work for or position in the same company?
	6.2	I feel I was paid enough and satisfied with my salary.
	6.3	I am happy with my personal health and injury insurance.
	6.4	I feel that the labor contract I signed with the construction company is reasonable and stable, and effectively protects my legal rights.
	6.5	Are you satisfied with your current employment contract with the construction company?

Table 8. CPCs self-reported questionnaire item specifications (part 3).

Number of CPCs	Questions' number	Description
C7: Working environment	7.1	Are there adequate supporting facilities and services during the construction process, such as additional service rooms, water supply, food, snacks, containers for resting?
	7.2	Was the work equipment useful and able to handle most construction tasks?
	7.3	Was the work equipment in good condition during the construction process?
	7.4	How adequate is the personal protective equipment provided during work?
	7.5	How useful is the personal protective equipment provided during work?
C8: Collaboration with coworkers	8.1	My coworkers often remind me when there are surrounding hazards.
	8.2	My coworkers are friendly to me.
	8.3	I was able to communicate smoothly and without obstacles with other workers during the construction process.
	8.4	I can learn relevant experience of construction work from my coworkers. They will teach me how to do my job better.
	8.5	I really like the team I work for, and it makes me happy to work together with my coworkers.
C9: Individual differences	9.1	Is it easy for me to understand the safety instructions from supervisors or managers?
	9.2	I can fully understand why certain actions must be taken (for safety).
	9.3	Did you think that your age negatively affects the tasks that were given? (e.g., more night shifts, more dangerous tasks)
	9.4	What is your current health status?
	9.5	Does your medical history affect the tasks you were given?

Table 9. An example of questions and options of C1.

Questions in C1	Answer
1. Is the training given adequate to do the work you are assigned to? (Likert scale)	(5) very adequate and complete
	(4) relatively adequate
	(3) generally adequate
	(2) training is not perfect
	(1) training is very scarce
2. Is the safety training sufficient in highlighting the hazards during work? (Likert scale)	(5) very sufficient and useful
	(4) relatively sufficient
	(3) generally sufficient
	(2) works a bit but not very useful
	(1) not useful at all
3. How often are the safety drills? (Safety drills — fire evacuation drills, first aid drills, accident drills.) (Give a range)	(5) Once a month or less than it
	(4) Once every 3 months
	(3) Once every 6 months
	(2) Once a year
	(1) Only when I entered the company
4. How long have you been working as a construction worker?	Answer: please give a certain number or range, for example: less than a year, . . . , more than ten years, . . .
5. I think I am an experienced construction worker.	(5) Strongly agree
	(4) Agree
	(3) Neither agree nor disagree
	(2) Disagree
	(1) Strongly disagree

References

[1] Levine CS, Al-Douri A, Paglioni VP, Bensi M, Groth KM. Identifying human failure events for Human Reliability Analysis: a review of gaps and research opportunities. *Reliab. Eng. Syst. Saf.* 2024, 245:109967.

[2] Stanton NA, Salmon PM, Rafferty LA, Walker GH, Baber C, *et al.* *Human factors methods: a practical guide for engineering and design*, 2nd ed. London: CRC Press, 2017. p. 656.

[3] Choi J, Gu B, Chin S, Lee JS. Machine learning predictive model based on national data for fatal accidents of construction workers. *Autom. Constr.* 2020, 110:102974.

[4] Sinabariba MP, Ghifari M, Muslim E, Moch B. Analysis of human error risk with human reliability methods in construction projects. In *2nd International Conference on Industrial and Manufacturing Engineering*, Medan, Indonesia, September 3–4, 2020, p. 012079.

[5] Swain AD, Guttmann HE. Handbook of human-reliability analysis with emphasis on nuclear power plant applications. Final report. 1983. Available: https://www.osti.gov/biblio/5752058. (accessed on 3 May 2024).

[6] Abbassi R, Khan F, Garaniya V, Chai S, Chin C, *et al.* An integrated method for human error probability assessment during the maintenance of offshore facilities. *Process Saf. Environ. Prot.* 2015, 94:172–179.

[7] Hollnagel E. *Cognitive reliability and error analysis method (CREAM)*, 1st ed. Oxford: Elsevier, 1998.

[8] Stanton NA. Systematic human error reduction and prediction approach (SHERPA). In *Handbook of human factors and ergonomics methods*, 1st ed. London: CRC Press, 2004. pp. 394–403.

[9] Konstandinidou M, Nivolianitou Z, Kiranoudis C, Markatos N. A fuzzy modeling application of CREAM methodology for Human Reliability Analysis. *Reliab. Eng. Syst. Saf.* 2006, 91(6):706–716.

[10] Yang Z, Bonsall S, Wall A, Wang J, Usman M. A modified CREAM to human reliability quantification in marine engineering. *Ocean Eng.* 2013, 58:293–303.

[11] Wang N, Du X, Zhang M, Xu C, Lu X. An improved weighted fuzzy CREAM model for quantifying human reliability in subway construction: modeling, validation, and application. *Hum. Factors Ergon. Manuf. Serv. Ind.* 2020, 30(4):248–265.

[12] He X, Wang Y, Shen Z, Huang X. A simplified CREAM prospective quantification process and its application. *Reliab. Eng. Syst. Saf.* 2008, 93(2):298–306.

[13] Zhou Q, Wong YD, Xu H, Van Thai V, Loh HS, *et al.* An enhanced CREAM with stakeholder-graded protocols for tanker shipping safety application. *Saf. Sci.* 2017, 95:140–147.

[14] Shi H, Wang J, Zhang L, Liu H. New improved CREAM model for Human Reliability Analysis using a linguistic D number-based hybrid decision making approach. *Eng. Appl. Artif. Intell.* 2023, 120:105896.

[15] Li X, Guo Y, Ge F, Yang F. Human reliability assessment on building construction work at height: the case of scaffolding work. *Saf. Sci.* 2023, 159:106021.

[16] Marseguerra M, Zio E, Librizzi M. Human Reliability Analysis by fuzzy "CREAM". *Risk Anal.* 2007, 27(1):137–154.

[17] Kim MC, Seong PH, Hollnagel E. A probabilistic approach for determining the control mode in CREAM. *Reliab. Eng. Syst. Saf.* 2006, 91(2):191–199.

[18] Ung ST. Evaluation of human error contribution to oil tanker collision using fault tree analysis and modified fuzzy Bayesian Network based CREAM. *Ocean Eng.* 2019, 179:159–172.

[19] Abdelghany M, Ahmad W, Tahar S. Event tree reliability analysis of safety-critical systems using theorem proving. *IEEE Syst. J.* 2021, 16(2):2899–2910.

[20] Kouzehgar M, Badamchizadeh MA, Khanmohammadi S. Fuzzy petri nets for human behavior verification and validation. *arXiv* 2013, arXiv:1303.1247.

[21] Cai B, Kong X, Liu Y, Lin J, Yuan X, *et al.* Application of Bayesian Networks in reliability evaluation. *IEEE Trans. Ind. Inf.* 2018, 15(4):2146–2157.

[22] Zhang R, Tan H. An integrated human reliability based decision pool generating and decision making method for power supply system in LNG terminal. *Saf. Sci.* 2018, 101:86–97.

[23] Fan H, Enshaei H, Jayasinghe SG. Human error probability assessment for LNG bunkering based on fuzzy Bayesian Network-CREAM model. *J. Mar. Sci. Eng.* 2022, 10(3):333.

[24] Shirali G, Hosseinzadeh T, Angali KA, Kalhori SRN. Modifying a method for human reliability assessment based on CREAM-BN: a case study in control room of a petrochemical plant. *MethodsX* 2019, 6:300–315.

[25] Sezer SI, Elidolu G, Aydin M, Ahn SI, Akyuz E, *et al.* Analyzing human reliability for the operation of cargo oil pump using fuzzy CREAM extended Bayesian Network (BN). *Ocean Eng.* 2024, 299:117345.

[26] Wu B, Yip TL, Yan X, Soares CG. Review of techniques and challenges of human and organizational factors analysis in maritime transportation. *Reliab. Eng. Syst. Saf.* 2022, 219:108249.

[27] Lin C, Xu Q, Huang Y. An HFM–CREAM model for the assessment of human reliability and quantification. *Qual. Reliab. Eng. Int.* 2022, 38(5):2372–2387.

[28] Chen X, Liu X, Qin Y. An extended CREAM model based on Analytic Network Process under the interval type-2 fuzzy environment for Human Reliability Analysis in high-speed train operation. *Qual. Reliab. Eng. Int.* 2020, 37(1):284–308.

[29] Shi H, Wang J, Zhang L, Liu H. New improved CREAM model for Human Reliability Analysis using a linguistic D number-based hybrid decision making approach. *Eng. Appl. Artif. Intell.* 2023, 120:105896.

[30] Emroozi VB, Modares A, Roozkhosh P. A new model to optimize the human reliability based on CREAM and group decision making. *Qual. Reliab. Eng. Int.* 2023, 40(5):1079–1109.

[31] Chen H, Li H, Goh YM. A review of construction safety climate: definitions, factors, relationship with safety behavior and research agenda. *Saf. Sci.* 2021, 142:105391.

[32] Shirley RB, Smidts C, Zhao Y. Development of a quantitative Bayesian Network mapping objective factors to subjective performance shaping factor evaluations: an example using student operators in a digital nuclear power plant simulator. *Reliab. Eng. Syst. Saf.* 2020, 194:106416.

[33] Vallat R. Pingouin: statistics in Python. *J. Open Source Softw.* 2018, 3(31):1026.

[34] Yao K, Yan S, Tran CC. A fuzzy CREAM method for Human Reliability Analysis in digital main control room of nuclear power plants. *Nucl. Technol.* 2022, 208(4):761–774.

[35] Wang F. Preference degree of triangular fuzzy numbers and its application to multi-attribute group decision making. *Expert Syst. Appl.* 2021, 178:114982.

[36] Zhou Q, Wong YD, Loh HS, Yuen KF. A fuzzy and Bayesian Network CREAM model for Human Reliability Analysis—the case of tanker shipping. *Saf. Sci.* 2018, 105:149–157.

[37] Center for Construction Research and Training. Construction Solutions Database. 2016. Available: http://www.cpwrconstructionsolutions.org/work/19/. (accessed on 6 November 2023).

[38] Hussain R, Pedro A, Lee DY, Pham HC, Park CS. Impact of safety training and interventions on training-transfer: targeting migrant construction workers. *Int. J. Occup. Saf. Ergon.* 2020, 2(26):272–284.

[39] Kohonen T. Self-organized formation of topologically correct feature maps. *Biol. Cybern.* 1982, 43(1):59–69.

[40] Campolongo F, Cariboni J, Saltelli A. An effective screening design for sensitivity analysis of large models. *Environ. Model. Softw.* 2007, 22(10):1509–1518.

[41] Morris MD. Factorial sampling plans for preliminary computational experiments. *Technometrics* 1991, 33(2):161–174.