Developments in the Built Environment 20 (2024) 100570

Contents lists available at ScienceDirect

Developments in the Built Environment

journal homepage: www.sciencedirect.com/journal/developments-in-the-built-environment

Assessing cyber risks in construction projects: A machine learning-centric approach

Dongchi Yao [a,b,*], Borja García de Soto [a,b]

[a] S.M.A.R.T. Construction Research Group, Division of Engineering, New York University Abu Dhabi (NYUAD), Experimental Research Building, Saadiyat Island, P.O. Box 129188, Abu Dhabi, United Arab Emirates
[b] Department of Civil and Urban Engineering, Tandon School of Engineering, New York University (NYU), New York, 11201, United States

ARTICLE INFO

Keywords:
Cybersecurity
Risk assessment
Machine learning
Construction industry
Digital twin

ABSTRACT

The construction industry is undergoing digitalization, but it is increasingly vulnerable to cyber attacks due to its slow pace in developing effective cyber risk assessment tools. This study develops a Machine Learning (ML)-centric approach to assess common cyber risks for construction projects. This approach comprises three components: (1) For risk prediction, a simulated dataset is generated using Monte Carlo simulations, which is utilized for model training. A two-phase model development strategy is proposed to select the optimal model for each risk. (2) For risk factor analysis, ML feature analysis methods are adapted to identify risk factors that contribute significantly to risks of specific projects. (3) For the risk reduction strategy, a greedy optimization algorithm is proposed to efficiently address high-contributing risk factors. To demonstrate the applicability of the developed approach, a case study is conducted on a real construction project.

1. Introduction

The construction industry has transitioned into the digital era, known as Construction 4.0, featured by the application of digital tools like digital twins, Building Information Modeling (BIM), and robotics in physical processes, creating cyber-physical systems (CPS). While these advancements enhance efficiency and productivity, they also increase cybersecurity vulnerabilities, potentially leading to significant delays, financial losses, and reputational damage in construction projects. The construction sector has lagged behind other industries in cybersecurity, facing a sharp increase in cyber incidents over the past decade (Deloitte). As outlined in (Yao and García De Soto, 2024a), this increase is particularly noticeable in five types of cyber risks.

(1) **Ransomware**: In the construction industry, ransomware attacks encrypt critical project data like blueprints and financial records, demanding ransom for decryption. These attacks lead to project delays, financial losses, operational shutdowns, and damaged client relationships, jeopardizing business success.
(2) **Phishing**: Phishing schemes in construction deceive individuals into revealing sensitive information through misleading emails. The consequences include unauthorized access to financial and project data, resulting in significant financial losses and compromised project security.
(3) **Insider Attacks**: Insider threats in construction involve trusted employees who maliciously steal or sabotage key resources or leak proprietary information. This results in unexpected costs, project delays, and legal issues, highlighting the need for strict security measures and employee monitoring.
(4) **Data Breaches**: Data breaches in construction involve unauthorized access to sensitive digital data, leading to financial losses, legal issues, and reputational damage that can impact future business opportunities. Robust cybersecurity measures are essential to protect sensitive information.
(5) **Supply Chain Attacks**: Supply chain attacks in the construction industry compromise critical components through third-party relationships, leading to severe disruptions such as supplier insolvency, transportation issues, and the use of substandard materials. These disruptions delay projects, increase costs, and compromise project quality.

The manifestation of these cyber risks, as seen in the attacks on projects by Bouygues Construction (Barbaschow, 2023), Skender Construction (Thibault, 2024), Fast Brick Robots (FBR) (Chris, 2023),

* Corresponding author. Department of Civil and Urban Engineering, Tandon School of Engineering, New York Univeristy (NYU), New York, 11201, United States.
E-mail address: dongchi.yao@nyu.edu (D. Yao).

https://doi.org/10.1016/j.dibe.2024.100570
Received 5 June 2024; Received in revised form 23 October 2024; Accepted 1 November 2024
Available online 19 November 2024

Marous Brothers Construction (Sawyer and Rubenstone, 2021), Turner Construction (Stiles, 2021), and Bird Construction (Catharine, 2023), etc., underscores the imperative for the industry to strengthen its cybersecurity management in construction projects. Project managers equipped with predictive tools for forecasting cyber risks across project phases can proactively address these risks (Yao and García De Soto, 2024a). This foresight enables them to align mitigation strategies with the project's risk tolerance, whether reducing or completely averting risks. A dynamic risk assessment tool that accepts real-time project data and uses predictive models to evaluate cyber risk levels is preferred. It should identify key risk factors and propose tailored strategies for mitigation. The goal is not just to react to risks but to anticipate and neutralize them before they affect project success (Yao and García De Soto, 2024a), (Yao and García de Soto, 2022).

As mentioned in Section 2, cybersecurity research in the construction industry is relatively insufficient, with a lack of comprehensive and dynamic cyber risk assessment tools. However, machine learning (ML) techniques, already used in construction to assess risks like delays and financial or environmental issues, offer a promising solution for cyber risk. By automating large data analyses, ML enables the swift detection of abnormal patterns that manual reviews might miss, allowing for timely risk predictions. As more structured cybersecurity data becomes available, organizations can refine ML models and adjust security measures to keep pace with evolving threats and industry complexities.

This study aims to develop an ML-centric approach for cyber risk assessment in construction projects, with four key objectives: (1) Develop ML models to process project characteristics and predict cyber risk degrees, ranging from 0 to 1, representing the probability of risk occurrence. (2) Use ML feature analysis methods to identify risk factors of general importance across projects and those specific to individual projects. (3) Propose a greedy optimization algorithm to formulate risk reduction strategies, determining which risk factors to address and to what extent, to align with the project's risk tolerance. (4) Demonstrate the applicability of this approach with a real construction project case study. This study addresses a gap in the literature by developing a structured approach for cyber risk assessment at the project level. It pioneers research at the intersection of cybersecurity, ML, and construction management, thereby broadening the application scope and advancing integrated research fields. This approach lays the foundation for a tool that dynamically assesses cyber risks across project phases, enabling project managers to make informed decisions for preemptively reducing high-impact risks.

The paper is organized as follows: Section 2 reviews existing research on construction cybersecurity and ML applications for risk assessment in various industries. Section 3 outlines the methodology for developing the ML-centric approach. Section 4 presents a case study using a real construction project. Section 5 discusses the complexity of the cybersecurity landscape, key risk factors, the prospect of the developed models, and limitations with future research directions. Section 6 concludes the study.

2. Related works

2.1. Existing studies on construction cybersecurity

The construction industry's awareness of cybersecurity issues, though increasing, remains relatively underdeveloped compared to other sectors. This is reflected in the limited body of research dedicated to the topic. A 2023 scoping review by Pargoo and Ilbeigi (Salami Pargoo and Ilbeigi, 2023a) highlighted this concern, identifying only 45 studies focused on cybersecurity within the construction industry. These studies can be broadly categorized into three groups: general discussions, review papers, and specific solutions.

General discussions, such as those by Bello and Maurushat (2020), Sonkor and García de Soto (Sonkor and García de Soto, 2021a), (Sonkor and García de Soto, 2023), EI-Sayegh et al. (El-Sayegh et al., 2020), Mantha and García de Soto (Mantha and García de Soto, 2019), Yao and García de Soto (Yao and García de Soto, 2022), García de Soto et al., (García de Soto et al., 2022), Turk et al. (2022), have contributed by providing foundational insights into cybersecurity challenges specific to construction. These discussions have been important in elevating the profile of cybersecurity concerns, helping the sector to prioritize it as a vital issue. These foundational insights are instrumental as they delineate the unique vulnerabilities that our machine learning-centric approach aims to address.

Review papers, including those by Pargoo and Ilbeigi (Salami Pargoo and Ilbeigi, 2023a), Pärn and Edwards (Parn and Edwards, 2019), Sonkor and García de Soto (Sonkor and García de Soto, 2021b), and Goh et al. (2021), have contributed by synthesizing current research, identifying gaps, and shaping the future research agenda. These reviews not only highlight the theoretical framework within which construction cybersecurity operates but also underscore the urgent need for practical, implementable strategies—a gap this study seeks to fill using advanced machine learning techniques.

Studies focused on specific solutions have made important contributions by providing technical or methodological strategies to address cyber risks. These include blockchain technology (Parn and Edwards, 2019), (Shemov et al., 2020), (Sonkor and García de Soto, 2021c), cybersecurity frameworks (Turk et al., 2022), Common Vulnerability Scoring System (CVSS) (Mantha and García de Soto, 2021), (Mantha et al., 2020), ML and deep learning algorithms (Yao and García De Soto, 2024a), (Pan et al., 2019), (Sheikh et al., 2019), (Yao and García De Soto, 2024b), and risk assessment techniques (Mantha and García de Soto, 2019), (Mantha and García de Soto, 2021), (Mantha et al., 2024), (Mohamed Shibly and García de Soto, 2020). These contributions not only provide a toolkit of existing solutions that can be leveraged but also demonstrate the application of complex computational techniques in addressing cyber risks—an approach this study intends to extend and sophisticate.

2.2. Limitations of current cyber risk assessment methods in construction

Among the few risk assessment studies, Mantha and García de Soto (Mantha and García de Soto, 2019) used agent-based models to explore vulnerability spread among stakeholders but overlooked the likelihood of risk occurrences, which are influenced by project-specific factors. Their model, in its early and oversimplified phase, lacked generalizability. In another study (Mantha and García de Soto, 2021), they applied the CVSS to assess risks in construction networks, but the manual scoring process lacked adaptability across different projects. Later, Mantha et al. (2024) enhanced the agent-based model from (Mantha and García de Soto, 2019), proposing seven more detailed steps to assess vulnerability spread. However, the framework's implementation still required extensive manual work and relied on strong assumptions about agent interactions. Mohamed Shibly and García de Soto (Mohamed Shibly and García de Soto, 2020) applied attack tree-based modeling to assess risks, using a 3D printer tampering case study, but the manual approach made it cumbersome and lacked real-world validation. The scope of these studies is fragmented, ranging from project-level to equipment-level. The methods depend on manual processes and often fail to address the complexities of evolving cyber threats. This highlights the need for automated and adaptable solutions—areas where ML excels.

2.3. ML for assessing various construction risks

Building on the success of ML in assessing various construction risks, its potential for addressing cyber risks is becoming clear. For example, in delay risk assessments, Sanni-Anibire et al. (2022) demonstrated that Artificial Neural Networks (ANN) achieved high accuracy, while Gondia et al. (2020) and Fitzsimmons et al. (2022) showed decision trees and hybrid models outperforming traditional methods. These studies

highlight ML's ability to improve prediction accuracy, though real-time application and wider industry adoption remain challenging. In safety risk assessments, George et al. (2022) used ensemble models like Gradient Boosting Machine (GBM), while Liu and Tian (2019) integrated distributed ML algorithms with cloud theory for early-warning systems. Although these contributions are significant in identifying critical safety indicators, practical deployment in active construction sites requires more transparent and user-friendly systems for non-experts. Similarly, Poh et al. (2018) and Gondia et al. (2022) demonstrated ML's effectiveness in injury prediction, but broader adoption calls for more accessible tools and industry-wide training. In compliance risk, dispute resolution, and defect detection, Raliile et al. (Raliile and Haupt, 2020) explored unsupervised ML for compliance monitoring, while Anysz et al. (2021) and Fan (2020) demonstrated ML's effectiveness in predicting dispute outcomes and defect probabilities. These studies underscore the versatility of machine learning in construction risk management and demonstrate its potential to transform cybersecurity measures within the industry. By employing ML methodologies, which are readily transferable to cyber risk assessment, the industry can not only enhance predictive accuracy but also facilitate a proactive cybersecurity consciousness.

2.4. ML for managing risks in other industries

Beyond construction risks, ML has been widely integrated into cybersecurity management across various fields, particularly in IT-related areas. ML's role in enhancing cyber risk management—spanning identification, estimation, and monitoring—demonstrates its value in improving cybersecurity frameworks. For risk identification, Pang et al. (2016) used Support Vector Machine algorithms combined with ensemble learning to detect early software vulnerabilities, while P et al. (P et al., 2021) combined Bat Optimization Algorithm for Wrapper-based Feature Selection with Random Forest algorithms to identify Android malware. Similarly, Russell et al. (2018) applied Deep Representation Learning to analyze C and C++ code, improving vulnerability detection at a granular level. These examples show ML's ability to facilitate data analysis and proactively detect threats. For risk estimation, ML tools enable rapid and precise assessments. Matsika et al. (2016) developed an ML-based tool to estimate the risk of terrorist attacks on metro systems, while Jiao (2018) used a Genetic Algorithm-backed Neural Network to generate cyber risk scores, demonstrating ML's adaptability in creating detailed risk metrics. These tools transform large, complex data sets into actionable risk assessments, supporting swift decision-making. In the monitoring phase, ML is crucial for continuous risk management. Chung et al. (2016) used a Naïve Q-Learning algorithm for adaptive decision-making in cybersecurity systems, and Li et al. (2019) applied ensemble prediction algorithms with distributed streaming to monitor network traffic in cyber-physical systems, showcasing ML's ability to handle real-time threats. These developments in ML-driven cyber risk management highlight its significant potential for construction cybersecurity, marking a step forward in better detection, accurate risk assessments, and adaptive monitoring. As ML keeps improving and automating risk management across different sectors, its use in construction is expected to greatly enhance the protection of critical infrastructure.

In summary, several observations from existing studies inform our research: (1) the existing literature on cybersecurity within the construction industry is relatively insufficient compared to that in other industries; (2) the current cyber risk assessment methods applied in construction are typically manual, inflexible, and fragmented at the assessment level. In this context, ML techniques offer advantages due to their capabilities for automation and adaptability. (3) The successful application of ML for assessing various construction risks indicates its potential for assessing cyber risk in construction; and (4) its proven effectiveness in enhancing cybersecurity management in other sectors further supports its applicability for improving cyber risk assessments in construction. Given these insights, this study proposes an ML-centric approach to effectively assess and manage cybersecurity risks in construction projects, aiming to introduce more automatic, adaptive, and integrated risk assessment strategies.

3. Methodology

Fig. 1 presents the flowchart for developing the ML-centric approach. Step 1 outlines the feature sources for the ML models, derived from risk factors identified in (Yao et al., 2023). Step 2 involves generating data samples through Monte Carlo simulation and employing an ensemble labeling method that combines Fault Tree Analysis with criteria-based labeling to ensure comprehensive and objective labeling. Step 3 describes the two-phase model development strategy, which includes selecting the best model for each risk and determining the optimal weight combination for different labeling methods. Step 4 applies the ML feature analysis method to identify risk factors significantly contributing to specific project risks. Step 5 introduces a greedy optimization algorithm to efficiently formulate risk reduction strategies. The final approach results in a dynamic cyber risk assessment tool with three main modules: (1) risk degree prediction by the trained ML models (Steps 1–3); (2) risk factor analysis (Step 4); and (3) risk reduction strategy formulation (Step 5).

3.1. Risk factors as ML features

The features of the ML model are based directly on the risk factors associated with construction projects identified in (Yao et al., 2023), which are used to assess the five cyber risks. The study in (Yao et al., 2023) employed a rigorous and systematic process to ensure the comprehensiveness, accuracy, relevance, and suitability of the risk factors. This process involved a systematic literature review, expert evaluation via the Delphi method, and a detailed questionnaire survey, comprising a total of 7 steps: (1) Literature Review: A comprehensive review of 18 publications on construction risks and six cybersecurity sources identified key risk factors, establishing a foundation for analysis. (2) Defining Risk Factor Categories: Six aspects of a construction project were identified—basic information, project structure, cybersecurity scores, project context, IT factors, and OT factors—to provide a holistic view of the cybersecurity landscape. (3) Internal Identification and Evaluation: The literature for each category was reviewed through ten internal discussions, assessing factors based on relevance, clarity, and data collection feasibility. External expert input was also gathered, leading to 62 preliminary factors. (4) Questionnaire Survey: A detailed survey was developed with definitions for each of the 62 risk factors, divided into six sections. Experts provided feedback using a 1–5 scale for factor inclusion. (5) Expert Evaluation: Three experts (two in cybersecurity and one in construction) evaluated the factors for reasonability on a 1–5 scale, offering feedback over five months via online meetings, emails, and calls, enhancing validity. (6) Revising Risk Factors: Based on feedback, explanations were refined, and factors with average scores below 3 were removed, resulting in 32 final risk factors grouped into five categories: overall project information, project structure, IT factors, OT factors, and management/human factors. (7) Determining Scales for Risk Factors: A quantitative approach was used to define scales for each risk factor, adjusted through expert feedback, and transitioning some factors to numerical values to support precise risk evaluations in construction projects.

Key characteristics of these 32 factors are summarized below.

- The 32 risk factors encompass five facets of construction projects: (1) General Project Information; (2) Project Structure; (3) Information Technology (IT) factors; (4) Operational Technology (OT) factors; and (5) Management and Human Factors. They address both general vulnerabilities and those unique to the construction industry, ensuring a balanced risk assessment model.

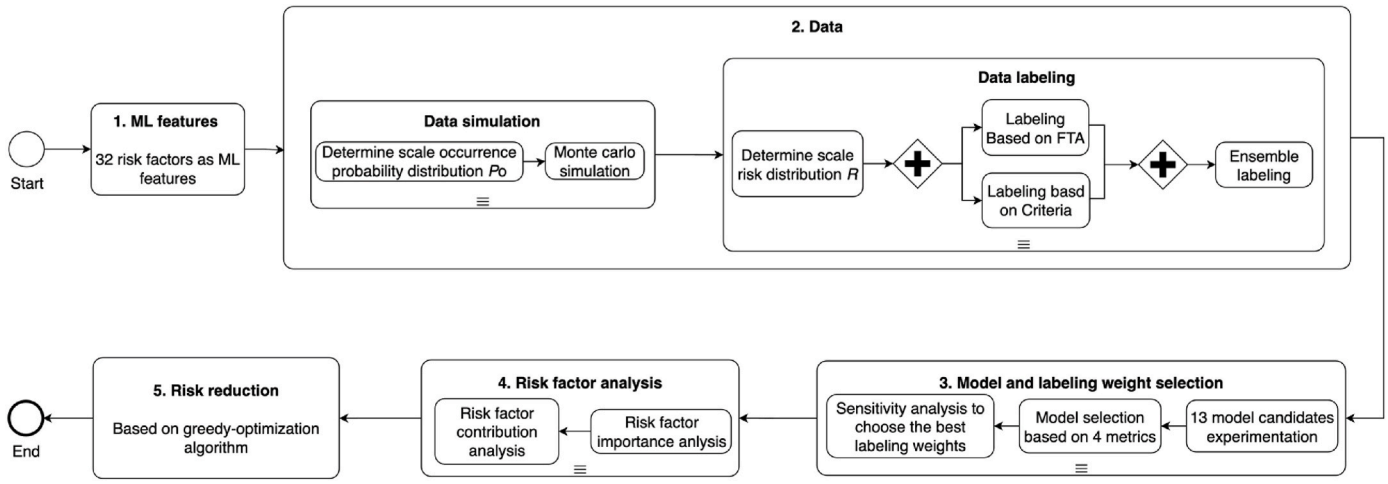


Fig. 1. The development flowchart of the cyber risk assessment approach.

- The projects are approached as a multi-layered network, with each layer consisting of diverse sub-teams and communication channels. From these, the authors identify Category 2 risk factors. This layer-specific approach enables the model to capture the dynamic nature of contemporary construction projects.
- Risk factors in Categories 1 and 2 provide an overview of the project, setting a context for the risk assessment model. Those in Categories 3, 4, and 5 are derived from specific project phases and use data directly from the companies involved in those phases. This combination of broad and detailed data enables the model to make more contextually precise predictions.
- The risk factors enable a quantitative risk assessment by categorizing each into distinct scales (levels, categories, or numerical values), requiring data collection to determine each factor's risk status. This approach facilitates detailed and objective risk decision-making. For example, the cybersecurity budget risk factor is divided into six scales:≤1%, 1%–2%, 2%–3%, 3%–4%, 4%–5%, and >5%. Similarly, project duration is divided into five intervals:≤3 months, 3–6 months, 6–12 months, 12–24 months, and >24 months. These proposed scales were refined through an iterative expert review process, including expanding or narrowing scales and transitioning some from categorical to numerical representations. The finalized scales are detailed in the Appendix and balance granularity with feasibility.

3.2. Data simulation and labeling

Early in the research, we encountered a shortage of existing datasets of construction project cases. Drawing inspiration from studies (Jacobsen, 2023), (Wolf et al., 2021), (Thambawita et al., 2022), we generated synthetic project samples for model training using Monte Carlo simulations (RiskAmp, 2012). We then adopted an ensemble labeling approach to label risk degrees to the synthetic samples. The data generation and labeling processes were rigorous, and the labeled dataset underwent evaluation to ensure its relevance to the construction industry.

3.2.1. Data simulation

Monte Carlo simulations can model real-world uncertainty and variability, enabling extensive scenario analysis and the creation of a comprehensive dataset for initial model training. The process is shown in Equation (1). To generate a project case *i*, probabilistic sampling is applied to all risk factors. Each risk factor has a unique probability distribution for its respective scales, denoted as $\boldsymbol{Po}_j$, indicating the likelihood of each scale's occurrence. We sampled from this probability distribution to select one scale for each risk factor, and the chosen scale was denoted by its index within that risk factor's range $K^{(j)}$. For example, for a certain risk factor *j* of project *i*, if the 1st scale (when $k = 1$) out of 5 scales $\left(K^{(j)} = 5\right)$ is selected according to $\boldsymbol{Po}_j$, then $S_i(j) = (1-1) \bullet 1 + (2-1) \bullet 0 + (3-1) \bullet 0 + (4-1) \bullet 0 + (5-1) \bullet 0 = 0$, which is the index of the 1st scale according to Python programming's indexing system.

$$S^I_{i=1}(j) = \sum_{k=1}^{K^{(j)}} (k-1) \times \text{Sample}\left(\boldsymbol{Po}_j\right) for\ all\ j = 1, \cdots, J \tag{1}$$

Where $S^I_{i=1}(j)$ is the array of simulated scales for each risk factor *j* across all *I* project cases; $K^{(j)}$ is the number of scales of the *j*-th risk factor; $(k-1)$ is the index of the *k*-th scale, indexing from 0; $\boldsymbol{Po}_j$ is the probability distribution of occurrence across the scales of the *j*-th risk factor; $\text{Sample}\left(Po_j\right)$ would return a 1 for the chosen scale and 0 for others, essentially picking *k*-th scale for the *j*-th risk factor based on $\boldsymbol{Po}_j$.

Determining $Po = \left\{\boldsymbol{Po_1}, \boldsymbol{Po_2}, \cdots \boldsymbol{Po_j} \cdots, \boldsymbol{Po_J}\right\}$ is essential for the Monte Carlo simulation. For a risk factor *j*, two approaches are adopted: (1) Data sources. The resource we relied on is a large, published text database (Yao and García de Soto, 2023) that includes six textual sources related to cybersecurity in construction. This approach resulted in the determination of 63% of the risk factors. (2) Delphi-based method. For the remaining risk factors for which required data is not available, we determined $\boldsymbol{Po}_j$ based on our internal expertise and an external expert review, employing a Delphi-inspired expert elicitation process (Galanis, 2018) as follows:

- **First Round:** One of the authors, with over five years of industry experience, assigned likelihoods to each scale for every risk factor, normalizing them to ensure the sum equaled 1. This process was repeated after a one-day interval, and the average likelihood for each risk factor was calculated and recorded.
- **Second Round:** A second internal expert, with over 20 years of industry experience, reviewed these initial probability distributions, providing feedback and suggesting adjustments as needed. The two experts discussed the recommendations and reached a consensus on any modifications. A questionnaire was then prepared to gather input from the external expert, offering two options for each risk factor: (1) The probability distribution is reasonable, or (2) The probability distribution is not reasonable, with required suggestions for improvement.
- **Third Round:** An external expert with over ten years of construction industry experience in the U.A.E. (as shown in Table 1) reviewed the distributions via the questionnaire without direct interaction with the internal experts to avoid bias. The internal team then discussed

Table 1
Information of the expert in reviewing.

Field	Background	Affiliation	Years of expertise	Expertise	Process of reviewing ***Po***
Construction	Executive MBA in operations and construction management	A leading construction company in the U.A.E.	10+	Expert in construction management, cybersecurity innovation in construction, business transformation, and new product development	- Gained an understanding of the scales of each risk factor - Reviewed the occurrence probability of each scale within each risk factor - Modified the probability distribution and discussed it with the authors for finalization through 3 in-person meetings, 5 ZOOM meetings, and 20+ emails - Ensured the sum of the scale probability distributions equals 1

the external feedback and made final adjustments to the probability distributions. This Delphi-based process ensures a comprehensive and unbiased assignment of scale probabilities, enhancing the reliability and validity of the results. The finalized *Po* is shown in the Appendix.

Risk Factors 2.2 and 2.3 relate to the layers of a construction project, which include eight predefined layers. This adds seven additional factors for each of these two risk factors, resulting in a total of 46 factors (32 initial factors plus 14 layer-specific ones) and 259 scales. Using Equation (1), we generated 1000 simulated project, producing the dataset shown in Table 2. This dataset consists of 1000 rows, each representing an individual project case, and columns indicating the indices of the selected scales for each risk factor, derived from probabilistic sampling.

3.2.2. Data labeling

For model training, labels indicating the risk degrees for the five cyber risks are needed for each generated project case. This study defines the risk degree within the range [0,1], interpreted as the probability of occurrence. This definition is based on the concept of the probability of the top event (the risk) occurring in the Fault Tree Analysis method (Lee et al., 1985) discussed later. The labeling process is represented by Equation (2).

$$p_i = f^l\left(r_{1,S_{i,1}}, r_{2,S_{i,2}}, \cdots, r_{j,S_{i,j}}, \cdots, r_{J,S_{i,J}}\right) \tag{2}$$

Where p_i is the labeled risk degree of a cyber risk for the i-th project, within [0,1]; f^l represents the labeling process; $S_{i,j}$ is the selected scale of the j-th risk factor of the i-th project; $r_{j,S_{i,j}}$ is the risk degree of the j-th risk factor that equals the risk degree of the selected scale, which, inspired by probability propagation in the Fault Tree Analysis, represents the probability of its progression into the worst-case scenario.

This study uses an ensemble approach that integrates multiple risk assessment methods to achieve two purposes: (1) mainly to complement individual assessment methods, combining them into a single overall risk degree label, and (2) to reduce the potential variance and bias often present in single-method labeling f^l (George et al., 2022). Equation (2) is revised to Equation (3) to reflect this integration. The ensemble approach combines Fault Tree Analysis (FTA) and a criteria-based method. FTA provides a systematic framework for modeling complex interdependencies among risk factors and analyzing their causal links to cyber risks, while the criteria-based method adds simplicity and complementarity. This approach ensures a more comprehensive and reliable risk labeling across projects. Obtaining $\bar{p}_i$ involves five steps, detailed in Sections 3.2.2.1 to 3.2.2.5.

$$\bar{p}_i = \sum_{e=1}^{E} w_e^l \bullet f_e^l\left(r_{1,S_{i,1}}, r_{2,S_{i,2}}, \cdots, r_{j,S_{i,j}}, \cdots, r_{J,S_{i,J}}\right) \tag{3}$$

Where $\bar{p}_i$ is the ensembled risk degree of the i-th project; w_e^l is the weight of the e-th labeling method, $\sum_{e=1}^{E} w_e^l = 1$; f_e^l is the e-th labeling method.

3.2.2.1. Determine the scale risk distribution (R). For a risk factor, the risk degrees of its associated scales make up the risk distribution for that risk factor, denoted as $\boldsymbol{R_j} = \left(r_{j,1}, r_{j,2}, \cdots, r_{j,k}, \cdots, r_{j,K^{(j)}}\right)$. The determination of this distribution can be approached in two ways, depending on whether the scales are ordinal or categorical.

(a) **The risk distribution of ordinal scales.** For risk factors with ordinal scales (38 risk factors), such as risk factor 1.4, the worst-case scenario is represented by a specific scale with a risk degree set to 1. Typically, the riskiest scale is either the first or the last. Risk degrees for the other scales are then considered to have a linear relationship with the worst-case scenario's risk degree, as shown in Equations (4) and (5).

If the first scale represents the worst scenario:

$$r_{j,k} = \frac{K^{(j)} - (k-1)}{K^{(j)}} \tag{4}$$

Table 2
Data structure.

Project No.	Risk Factor No.													
	1.1	1.2	1.3	1.4	1.5	1.6	1.7	…	5.2	5.3	5.4	5.5	5.6	5.7
1	2	2	3	2	2	3	1	…	2	0	0	1	0	2
2	3	3	3	3	3	4	1	…	3	1	1	1	1	3
3	2	2	3	2	2	3	1	…	2	0	1	1	1	2
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
998	5	4	5	4	4	4	2	…	4	1	1	3	2	4
999	0	0	1	1	1	1	0	…	1	0	0	0	0	1
1000	3	2	3	3	2	4	1	…	2	0	1	1	1	2

Notes.
Project No. — The numerical identifier for the project.
Risk Factor No. — The numerical identifier for risk factors.
Scale: Values in the table represent the index of the scale sampled for each corresponding risk factor, with indexing starting from 0.

If the last scale represents the worst scenario:

$$r_{j,k} = \frac{k}{K^{(j)}} \tag{5}$$

(b) **The risk distribution of categorical scales.** For risk factors with categorical scales (8 risk factors), such as risk factor 1.7, it is necessary to determine the risk degree for each scale. To reduce uncertainty, fuzzy set theory (Zadeh, 1965) can be employed. This method handles imprecise and ambiguous data in complex systems and is widely used in risk analysis across industries like nuclear power, chemicals, and oil and gas. Risk likelihoods are assigned using a seven-level natural language scale: {Very Low (VL), Low (L), Moderately Low (ML), Medium (M), Moderately High (MH), High (H), Very High (VH)}. These terms are then mapped to fuzzy numbers, which are defuzzified into a point value to obtain the final risk probability. The two most common fuzzy number representations are triangular $\widetilde{\boldsymbol{P}}_{\boldsymbol{A}} = (P_{a_1}, P_{a_2}, P_{a_3})$ and trapezoidal $\widetilde{\boldsymbol{P}}_{\boldsymbol{B}} = (P_{b_1}, P_{b_2}, P_{b_3}, P_{b_4})$. The relationship between these terms and fuzzy numbers, derived from their membership functions (Wang et al., 2022), can be shown as Fig. 2.

For defuzzification, the Center of Area (CoA) method (Zadeh, 1965), known for its straightforwardness and practical application, is commonly used. The defuzzification formula for the triangular and trapezoidal fuzzy number is given in Equations (6) and (7) respectively (Senol et al., 2015).

$$p_A^* = \frac{1}{3} \bullet (p_{a_1} + p_{a_2} + p_{a_3}) \tag{6}$$

$$p_B^* = \frac{1}{3} \bullet \frac{(p_{b_4} + p_{b_3})^2 - p_{b_4}p_{b_3} - (p_{b_1} + p_{b_2})^2 + p_{b_1}p_{b_2}}{p_{b_4} + p_{b_3} - p_{b_2} - p_{b_1}} \tag{7}$$

By implementing the method (a) for ordinal scales and (b) for categorical scales, we preliminarily determined the risk degrees for all 259 scales. These were presented to the same expert from the construction company and underwent minor modifications based on their professional feedback. The finalized scale risk distribution R is displayed in Appendix.

3.2.2.2. Convert scales into risk degrees. Now that the risk degree for each of the 259 pre-defined scales has been established, the scales generated in Table 2 can be converted into their corresponding risk degrees. This allows for the determination of the risk degree for each risk factor in every project, as shown in Table 3.

3.2.2.3. FTA labeling method. FTA (Lee et al., 1985) is a versatile analytical method used in risk assessment across various industries, such as the oil and gas industry (Wang et al., 2022), chemical industry (Senol et al., 2015), and construction industry (Aghaei et al., 2022), to investigate the causes and probabilities of an undesired event. In this study, we applied the FTA methodology as outlined in (Yao and García de Soto, 2023) and drew inspiration from the fault tree development processes in the above-mentioned literature to create fault trees specifically for each cyber risk.

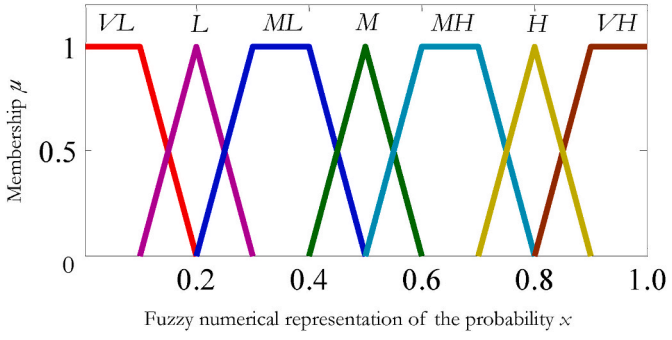


Fig. 2. Fuzzy numerical representation of natural language terms (Wang et al., 2022).

The FTA serves as a crucial framework for systematically modeling the complex interdependencies among risk factors and analyzing their causal relationships with cyber risks. This modeling approach is essential for capturing the non-linear interactions among risk factors and their collective impacts on overall cyber risks. The key significance of FTA in this study lies in its ability to provide a comprehensive understanding of how various risk factors interact and contribute to project vulnerabilities. In this study, the inputs to the fault tree are the risk degrees associated with each project's risk factors, as detailed in Table 3. These risk factors are sampled and simulated using the Monte Carlo simulation method implemented in Section 3.2.1, which helps in generating a range of potential construction project scenarios. The primary objective of employing FTA is to label each project by deriving a risk degree that indicates its associated cyber risk. These derived labels are crucial for subsequent machine learning training, as they inform the development of various models, such as linear regression and neural networks, as outlined in Section 3.3.2.

The process of constructing fault trees began by identifying the 'top event' (cyber risks) and positioning it at the top of the tree. We then broke down the top event into intermediate events representing broader categories of contributing factors, further decomposing these into basic events (individual risk factors) at the bottom. Relationships between events were mapped using logic gates like AND or OR to define how different events contribute to the top event. Each risk factor was carefully evaluated for its direct relevance to the intermediate and top events, ensuring that only factors with a clear relationship to cyber risks were included. Finally, the completed fault trees were reviewed and validated by an expert from the U.A.E. with extensive experience in construction management and cybersecurity innovation (shown in Table 1), ensuring that the trees accurately reflect specific cyber risks within the construction industry context.

The probability (risk degrees) of the top event can be derived from the risk degrees of the basic events using Boolean logic operators on gates. Events from which arrows originate are termed "parent events", while those the arrows point to are called "child events". For AND gate and for OR gate, the probability of the parent event is calculated as Equations (8) and (9) respectively.

$$P(Parent) = f_{AND}(A, B, \cdots, N) = P(A) \times P(B) \times \cdots \times P(N) \tag{8}$$

$$P(Parent) = f_{OR}(A, B, \cdots, N) = 1 - (1 - P(A)) \times (1 - P(B)) \times \cdots \times (1 - P(N)) \tag{9}$$

Where $P(Parent)$ is the probability of the parent event; $P(A), P(B), \cdots, P(N)$ are probabilities of the child events; $f_{AND}(\bullet)$ is the probability calculation function for AND gate; $f_{OR}(\bullet)$ is the probability calculation function for OR gate.

Taking the ransomware risk (fault tree in Fig. 3(a)) as an example, p_{T_i} for the i-th project can be derived using the pseudo-code presented in Algorithm 1. Similarly, p_{T_i} for the other four types of risks can be derived using the same logic.

Algorithm 1. Pseudocode for deriving p_{T_i} of Ransomware risk

1 Compute the probability (risk degree) of Risk Factor 2.3: $p_0 = max(r_{2.3.1,S_{i,2.3.1}}, r_{2.3.2,S_{i,2.3.2}}, r_{2.3.3,S_{i,2.3.3}}, r_{2.3.4,S_{i,2.3.4}}, r_{2.3.5,S_{i,2.3.5}}, r_{2.3.6,S_{i,2.3.6}}, r_{2.3.7,S_{i,2.3.7}}, r_{2.3.8,S_{i,2.3.8}})$

2 Compute the probability of Poor Security Training event: $p_1 = f_{AND}(r_{3.8,S_{i,3.8}}, r_{5.2,S_{i,5.2}})$

(continued on next page)

Table 3
Data structure indicating risk degrees.

Project No.	Risk Factor No.													
	1.1	1.2	1.3	1.4	1.5	1.6	1.7	…	5.2	5.3	5.4	5.5	5.6	5.7
1	0.90	0.6	0.50	0.6	0.50	0.8	1.0	…	0.67	1.0	0.2	0.33	0.2	0.6
2	0.65	0.8	0.50	0.8	0.67	1.0	1.0	…	0.50	0.2	1.0	0.33	0.4	0.4
3	0.90	0.6	0.50	0.6	0.50	0.8	1.0	…	0.67	1.0	1.0	0.33	0.4	0.6
…	…	…	…	…	…	…	…	…	…	…	…	…	…	…
998	0.60	1.0	0.17	1.0	0.83	1.0	0.7	…	0.33	0.2	1.0	0.67	0.6	0.2
999	0.80	0.2	0.83	0.4	0.33	0.4	0.4	…	0.83	1.0	0.2	0.17	0.2	0.8
1000	0.65	0.6	0.50	0.8	0.50	1.0	1.0	…	0.67	1.0	1.0	0.33	0.4	0.6

Notes.
Project No. — The numerical identifier for the project.
Risk Factor No. — The numerical identifier for risk factors.
Scale: Values populated in the table represent the risk degrees of risk factors for each project, each ranging from 0 to 1.

(*continued*)

3 Compute the probability of Poor OT Equipment Security event: $p_2 = f_{\text{AND}}\left(r_{4.1,S_{i,4.1}}, r_{4.2,S_{i,4.2}}, r_{4.4,S_{i,4.4}}\right)$
4 Compute the probability of Inadequate Cyber Security Measures event: $p_3 = f_{\text{AND}}\left(r_{3.6,S_{i,3.6}}, r_{5.3,S_{i,5.3}}, r_{5.4,S_{i,5.4}}\right)$
5 Compute the probability of Insufficient Network and System Management event: $p_4 = f_{\text{OR}}\left(p_0, r_{3.3,S_{i,3.3}}, r_{3.4,S_{i,3.4}}, r_{3.5,S_{i,3.5}}\right)$
6 Compute the probability of Vulnerability Network event: $p_5 = f_{\text{AND}}\left(r_{3.7,S_{i,3.7}}, r_{3.9,S_{i,3.9}}, r_{4.3,S_{i,4.3}}, r_{4.5,S_{i,4.5}}\right)$
7 Compute the probability of Improper Human Operation or Management event: $g_1 = f_{\text{OR}}\left(p_1, r_{5.5,S_{i,5.5}}\right)$
8 Compute the probability of Technological Malfunction or Insufficient Technological Protection event: $g_2 = f_{\text{OR}}(p_2, p_3, p_4, p_5)$
9 Compute the probability (risk degree) of the Ransomware risk: $p_{T_i} = f_{\text{AND}}(g_1, g_2)$

3.2.2.4. Criteria-based labeling method. For the factors not covered in fault trees, the criteria-based method can be adopted to derive the risk degree of cyber risks. This method is intuitive, defining a threshold T that a risk factor must satisfy to be considered significant in influencing the overall risk degree of the project, as shown in Equation (10).

$$p_{C_i} = f_C\left(r_{1,S_{i,1}}, r_{2,S_{i,2}}, \cdots, r_{j,S_{i,j}}, \cdots, r_{J,S_{i,J}}\right) = \frac{1}{J_c}\sum_{j=1}^{J_C} I\left(RF_j \in RF_C \cap r_{i,S_{ij}} \geq T\right) \tag{10}$$

Where J_c is the number of risk factors not in the fault tree; RF_j is the j-th risk factor not in the fault tree; I is a function that returns 1 if the condition inside the parentheses is true and 0 otherwise

3.2.2.5. Ensemble labeling. For each of the 1000 generated project case, both labeling methods are applied, resulting in two risk degrees (p_T and p_C) for each risk, the ensembled risk degree of this risk is shown as Equations (11) and (12). The weights of the labeling methods will be explored to train the ML models in Section 3.3.2.

$$\bar{p}_i = w_T \bullet p_{T_i} + w_C \bullet p_{C_i} \tag{11}$$

$$w_T + w_C = 1 \tag{12}$$

3.3. Module 1: model development for risk prediction

To capture non-linearity in the simulated dataset, various model candidates are tested. The final model selection for each risk involves a two-phase strategy: first, identifying the best model candidate, and then determining the most effective combination of labeling weights.

3.3.1. Model candidate exploration

The ensemble labeling method introduces non-linearity between risk factors and project risk degrees. To capture this non-linearity effectively, we explore different models. We used $Pr_{j,k}$ to denote the value of the k-th scale in the one-hot representation of the j-th risk factor, which is either 0 or 1, indicating whether this scale is selected or not. If the k-th scale is selected, $Pr_{j,k} = 1$, and $\boldsymbol{Pr}_j = \boldsymbol{Pr}_j^{e_{j,k}} = (0, \cdots, 1_{(k)}, \cdots, 0)_j$.

The selection of machine learning models was based on their ability to capture non-linear relationships among risk factors and their impacts on project risk degrees. We began with linear regression to establish a performance baseline, then employed non-linear polynomial regression to enhance the model's capacity to detect complex interactions. Finally, neural networks were chosen for their ability to automatically learn feature transformations, allowing them to effectively model more intricate non-linearities in the data. Our goal is to experiment with and determine which model is best suited for each cyber risk, ensuring that we do not use complex models for simple linear datasets, thereby saving resources.

(1) **Linear regression.** We began with linear regression models that have limited capacity to capture non-linearity: basic linear regression (Huang, 2022), Lasso regression by adding an L1 penalty (Tibshirani, 1996), and Ridge regression by adding an L2 penalty (Hoerl and Kennard, 1970). Both Lasso and Ridge regressions use regularization to prevent overfitting. The basic linear regression model is described in Equation (13).

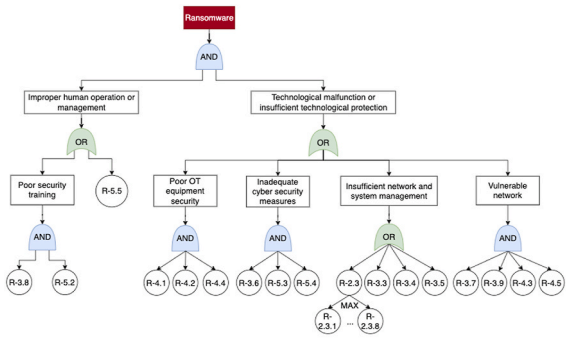
$$\widehat{p} = \beta_0 + \sum_{j=1}^{46}\sum_{k=1}^{K^{(j)}} \beta_{j,k} \bullet \boldsymbol{Pr}_{j,k} \tag{13}$$

Where $\widehat{p}$ is the predicted risk degree of a certain risk; β_0 is the intercept of the linear regression model; $\beta_{j,k}$ is the coefficient of the k-th scale of the j-th risk factor, totally 259; $\boldsymbol{Pr}_{j,k}$ is the value of the k-th scale in the one-hot representation of the j-th risk factor $\boldsymbol{Pr}_j$.

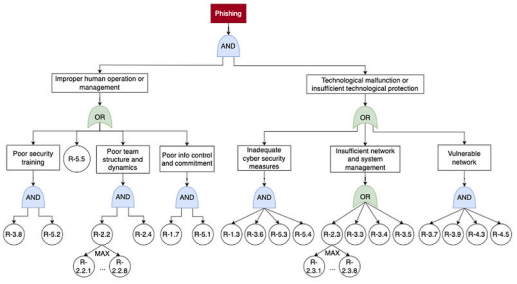
(2) **Non-linear regression.** Then, polynomial regression (Huang, 2022) was utilized to accommodate higher polynomial degrees of input features, enhancing the model's ability to detect non-linear relationships and capture the interactions between features. The polynomial regression of two features to the dependent variable y can be depicted as Equation (14).

$$y = \beta_0 + \sum_{i=1}^{d}\beta_{1i}x_1^i + \sum_{i=1}^{d}\beta_{2i}x_2^i + \sum_{i=1}^{d-1}\sum_{j=1}^{d-i}\beta_{ij}x_1^i x_2^j \tag{14}$$

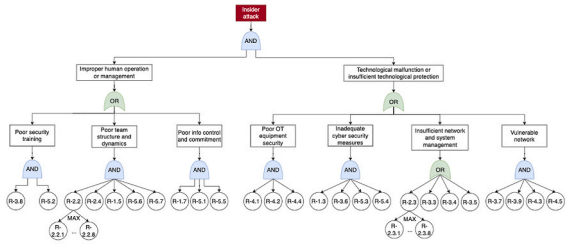
Where y is the dependent variable; x_1, x_2 are the features for exemplification; β_0 is the intercept; β_{1i} is the coefficients for the terms involving x_1 to the i-th power; β_{2i} is the coefficients for the terms involving x_2 to the i-th power; β_{ij} is the coefficients for the interaction terms, where x_1 is



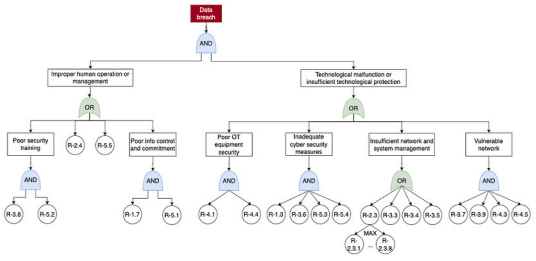
(a). Fault tree of ransomware



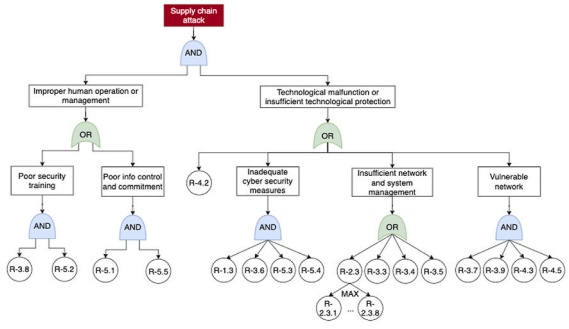
(b). Fault tree of phishing



(c). Fault tree of insider attack



(d). Fault tree of data breach



(e). Fault tree of supply chain attack

Fig. 3. Fault trees developed for the five cyber risks

Where OR represents the OR gate; AND represents the AND gate; ▭ represents the intermediate events; ◯ represents the basic events.

raised to the i-th power and x_2 is raised to the j-th power, with the constraint that $i+j$ does not exceed d; Given that the features in our study $Pr_{j,k}$ can only be 0 or 1 for one-hot encoding representation, all polynomial terms of a feature collapse to the original feature, so that Equation (14) can be simplified as Equation (15).

$$y = \beta_0 + \sum_{i=1}^{d} \beta_{1i} x_1 + \sum_{i=1}^{d} \beta_{2i} x_2 + \sum_{i=1}^{d-1} \sum_{j=1}^{d-i} \beta_{ij} x_1 x_2 \tag{15}$$

(3) **Neural networks.** Neural Networks (McCulloch and Pitts, 1943) can capture more non-linearity because they learn feature transformations automatically during training, using neurons in different layers to process and transfer information to the final output. In our study, the model includes an input layer with 259 neurons and is designed to predict the risk degree for single or multiple cyber risks. This multi-risk approach enables information sharing across the five cyber risk tasks, providing complementary insights and potentially enhancing generalization and performance. We tested nine neural network architectures, as detailed in Table 4, and experimented with three common activation functions: ReLU, LeakyReLU, and Tanh (Goodfellow et al., 2016).

3.3.2. *Training and selection*

To determine the optimal model for each risk, we need to test 13 model candidates and 11 labeling weights (ranging from 0 to 1 in 0.1 intervals), resulting in 143 combinations. To mitigate inefficiency, we adopted a two-phase selection inspired by the greedy algorithm (Edmonds, 1971), which prioritizes the best option at each step for local optimization. This approach reduces the number of training sessions to 24—13 for the initial model selection and 11 for weight testing—while still identifying the most effective model and weight combination.

(1) The model development strategy

Phase 1: Each of the 13 model candidates was trained for each risk using a weight of 0.5 for both labeling methods, treating them equally. The 1000 generated samples were split into training, validation, and testing sets in a 7:2:1 ratio. All model candidates were trained on the training set, with neural networks selected based on the epoch that achieved the lowest loss on the validation set. Performance was evaluated using four metrics: MSE, RMSE, MAE, and R^2. This multifaceted evaluation ensures comprehensive model assessment and offers diverse insights for stakeholders to understand performance.

Phase 2: A sensitivity analysis of the labeling weights was performed using the optimal model candidate. The best weights were selected based on model performance on two additional simulated projects. Initially, a specialist from a Chinese construction company reviewed and labeled the simulated projects, ensuring the risk factors were realistic and aligned with actual scenarios in their project database. The expert adjusted any unrealistic factors to better reflect reality. The revised

Table 4
Model structure and training configuration of neural networks.

Model Name	# of Hidden Layers	# of Neurons in Hidden Layers	Activation Function	# of Output Neurons
NN_1	1	100	ReLU	1
NN_2	2	150, 100	ReLU	1
NN_3	1	100	LeakyReLU	1
NN_4	3	200, 150, 100	ReLU	1
NN_5	1	100	Tanh	1
NN_6	2	200, 100	Tanh, ReLU	1
Combined_NN_1	1	100	ReLU	5
Combined_NN_2	2	150, 75	ReLU	5
Combined_NN_2	3	200, 100, 50	LeakyReLU	5

projects were then reviewed by two U.S.-based cybersecurity specialists, who labeled the projects to indicate the likelihood of specific risks occurring. This review and modification process ensured realistic simulations. Table 5 lists the experts, Table 6 presents the modified simulated projects, and Table 7 shows the labeling results. The labeling for each project risk is based on the majority vote from the three experts.

(2)Training details

(a)For linear regression, the commonly used loss function is MSE, as shown in Equation (16), while the loss functions for Lasso and Ridge regressions are presented in Equations (17) and (18), respectively.

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{p}_i - p_i)^2 \tag{16}$$

$$L_{Lasso} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{p}_i - p_i)^2 + \lambda\sum_{j=1}^{46}\sum_{k=1}^{K^{(j)}}|\beta_{j,k}| \tag{17}$$

$$L_{Ridge} = \frac{1}{N}\sum_{i=1}^{N}(\widehat{p}_i - p_i)^2 + \lambda\sum_{j=1}^{46}\sum_{k=1}^{K^{(j)}}\beta_{j,k}^2 \tag{18}$$

Where N is the number of project cases for training. λ is the regularization term. We optimized the hyperparameter λ for both Lasso and Ridge regressions using a grid search combined with 5-fold cross-validation. This method systematically evaluates different λ values, aiming to minimize the loss function while helping balance bias and variance, thereby enhancing the model's generalization capabilities while optimizing the trade-off between computational efficiency and accuracy.

(b) For non-linear regression, to train the model, we used the loss function from Equation (16), applying the same cross-validation as above on the same dataset to find the optimal polynomial degree. We limited polynomial degrees to 2 and 3 to avoid overfitting, considering the dataset's complexity does not warrant higher degrees.

(c) For neural networks, two loss functions were used respectively for the single output and multioutput, respectively using the loss functions in Equations (16) and (19). All models used the same training setup, utilizing the Adam optimizer with a learning rate of 0.004, 200 epochs for training, and a batch size of 16. The best-performing model on the validation set was saved.

$$L_{multi-output} = \frac{1}{5N}\sum_{i=1}^{N}\sum_{i=1}^{5}\left(\widehat{p}_{i,r} - p_{i,r}\right)^2 \tag{19}$$

Where $\widehat{p}_{i,r}$ is the predicted risk degree for the r-th risk of project case i; $p_{i,r}$ is the actual risk degree for the r-th risk of project case i

(3) Model and labeling weights selection results

By implementing the Phase-1 strategy, Table 8 displays the performance results with the best performances in bold. For each risk, the model with the best performance for each metric has been identified and summarized in Table 9. The optimal model for each risk, selected based on its consistently superior performance across metrics, is also detailed in Table 9.

Implementing the Phase-2 strategy of sensitivity analysis, Table 10 illustrates the predicted risk degrees for Risks 1 to 5 under different weight combinations, with a value of 0.5 or higher indicating the risk occurrence. The weight combination of 0.6 for FTA and 0.4 for Criteria-based labeling consistently matched actual labels (shown in Table 7) for all risks and both projects, leading to its selection as the final optimal weight combination for each risk in our study. We then retrained each selected base model with the optimal weight combination.

3.3.3. Analysis and evaluations

To demonstrate that our ensemble labeling methods, using the specified weight combinations in Phase 1 and the selected weight combination in Phase 2, produce a convincing dataset for training, a statistical analysis known as Kernel Density Estimation (KDE) (Parzen, 1962) was conducted. This analysis focuses on the distributions of the integrated weighted risk degrees of the generated samples, enabling an examination of their alignment with reality. Figs. 4 and 5 display the KDE plots of risk degrees for each risk category across 1000 generated project cases. Three observations are noteworthy: (1) The density distribution of the averaged risk degrees for all risks does not lean towards extremely high or low-risk degrees, confirming the unbiased nature and effectiveness of our ensemble labeling approach; (2) The density distribution of phishing and data breach tends slightly towards higher risk degrees, reflecting the prevalent occurrence of these cyber risks in the current construction cybersecurity landscape; (3) The density distribution for supply chain attack is skewed towards lower risk degrees, in line with their less frequent occurrence compared to phishing and data breach. These observations underscore the efficacy of our ensemble labeling method as it appears unbiased and mirrors the industry realities (Deloitte), (InfoCenter, 2024), (Salami Pargoo and Ilbeigi, 2023b), proving that our final selected models are reliable and have effectively encoded real-world information.

3.4. Module 2: risk factor analysis

The methods of Feature Contribution Analysis (FCA) (Rajbahadur et al., 2022) in ML were adapted to conduct analyses of risk factor importance and contribution. FCA was adapted to analyze the specific contribution of each risk factor to the project's risk at a particular time point in the progression of construction projects. Since this study involves various models, ranging from linear regression to neural networks, choosing an ML feature analysis method that is model-agnostic is advantageous, as it can be applied to a variety of different models.

Shapley Additive exPlanations (SHAP) (Lundberg and Lee, 2017) is recognized as a commonly used method in studies across various fields (Molnar, 2024). It boasts a model-agnostic nature, making it suitable for nearly all ML models. It adeptly manages data with high-dimensional

Table 5
Information of the experts in Phase 2.

Expert No.	Field	Background	Affiliation	Years of expertise	Expertise
1	Construction	Ph.D. in Civil Engineering from Tsinghua University	A Chinese construction company in Beijing	20+	Construction management, sustainable building practices, and urban development
2	Cybersecurity	Master's degree in Cybersecurity from Stanford University.	A cybersecurity scoring company in New York	20+	Expert in cybersecurity, governance, and risk management; develops security programs and advises on compliance, frameworks, and cloud-native security
3	Cybersecurity	Ph.D. in Computer Science from University of Tennessee, Knoxville	A cybersecurity scoring company in New York	10+	Expert in cybersecurity engineering and R&D strategy, threat research, cross-functional leadership with a focus on enhancing cybersecurity solutions

Table 6
The two simulated projects.

Project	Risk Factors																					
	1.1	1.2	1.3	1.4	1.5	1.6	1.7	2.1	2.2.1	2.2.2	2.2.3	2.2.4	2.2.5	2.2.6	2.2.7	2.2.8	2.3.1	2.3.2	2.3.3	2.3.4	2.3.5	2.3.6	2.3.7
S_1	2	2	3	2	2	3	1	4	2	1	3	0	0	0	0	5	3	3	5	0	1	1	1
S_2	0	0	0	0	0	0	0	0	0	1	0	1	0	5	0	0	0	1	0	3	1	7	0
Project	Risk Factors																						
	2.3.8	2.4	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.1	4.2	4.3	4.4	4.5	5.1	5.2	5.3	5.4	5.5	5.6	5.7
S_1	6	2	1	2	1	2	2	3	1	1	2	2	2	3	1	2	3	2	0	0	1	1	2
S_2	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Table 7
Labeling of the two simulated projects.

Project	Ransomware	Phishing	Insider Attack	Data Breach	Supply Chain Attack
S_1	Yes	Yes	Yes	Yes	No
S_2	No	Yes	No	No	No

features and can factor in interaction effects among features during analysis. For a specific sample, the contribution of each feature in the model can be quantified as the SHAP value demonstrated in Equation (20) (Lundberg and Lee, 2017), computed by iterating through all possible combinations of the features and determining how the introduction of the feature $Pr_{j,k}$ in question alters the prediction. By summing these variations over all combinations and averaging them, the SHAP value for a certain feature is obtained.

$$\mathrm{SHAP}_{j,k} = \sum_{Pr_s \subseteq Pr \setminus \{j,k\}} \frac{|Pr_s|!(|Pr| - |Pr_s| - 1)!}{|Pr|!} \left[f\left(Pr_s \cup \{Pr_{j,k}\}\right) - f(Pr_s) \right] \tag{20}$$

Where Pr is the set of all features; Pr_s is a subset of features not including $Pr_{j,k}$; $f\left(Pr_s \cup \{Pr_{j,k}\}\right)$ is the prediction with both the Pr_s and $Pr_{j,k}$; $f(Pr_s)$ is the prediction with just the features in Pr_s; $\frac{|Pr_s|!(|Pr|-|Pr_s|-1)!}{|Pr|!}$ is a combinatorial coefficient ensuring that each possible combination of features is weighted appropriately

For a specific project i, the contribution of a risk factor was determined by the sum of SHAP value across all its scales, as shown in Equation (21) where $\mathrm{SHAP}_{i,j,k}$ is the SHAP value of the k-th scale of the j-th risk factor for the i-th project case.

$$C_{ij} = \sum_{k=1}^{K^{(j)}} SHAP_{i,j,k} \tag{21}$$

3.5. Module 3: risk reduction strategy

The FCA analysis can identify the risk factors that significantly contribute to the project's predicted risk degree. This insight informs the development of targeted risk reduction strategies. To be efficient, the risk reduction strategies should be systematically developed, beginning with addressing the most critical ones. Equation (22) illustrates the brute force method of finding the global minimum risk degree by exploring all possible combinations of risk factors.

$$\boldsymbol{Pr}_1^{e_{1,k^*}}, \cdots, \boldsymbol{Pr}_j^{e_{j,k^*}}, \cdots, \boldsymbol{Pr}_J^{e_{J,k^*}} = \underset{1 \le e_{1,k} \le K^{(1)}, \cdots, 1 \le e_{J,k} \le K^{(J)}}{arg\ min} f\left(\boldsymbol{Pr}_1^{e_{1,k}}, \cdots \boldsymbol{Pr}_j^{e_{j,k}}, \cdots, \boldsymbol{Pr}_J^{e_{J,k}}\right) \tag{22}$$

Where $e_{j,k}$ denotes that the k-th scale of the j-th risk factor is selected; $\boldsymbol{Pr}_j^{e_{j,k}}$ is the one-hot representation of risk factor j, of which the k-th scale is selected so that $\boldsymbol{Pr}_{j,k} = 1$, denoted as denoted as $(0, \cdots, 1_{(k)}, \cdots, 0)_j$; $\boldsymbol{Pr}_j^{e_{j,k^*}}$ is the one-hot representation of risk factor j, of which the k^*-th scale is the optimal selection so that $\boldsymbol{Pr}_{j,k^*} = 1$, denoted as $(0, \cdots, 1_{(k^*)}, \cdots, 0)_j$.

The size of the search space is given by Equation (23). This rapidly expanding space is computationally demanding and impractical for quickly formulating risk reduction strategies. Therefore, a greedy optimization approach based on greedy principles (Edmonds, 1971) can be used. The pseudo-code for this process is shown in Algorithm 2. The core idea is to make the best choice at each step: selecting the highest positively contributing risk factor for optimization and choosing the scale that minimizes the predicted risk degree. This continues until the predicted risk drops below the threshold T. The final search space, if m risk factors are optimized, would be reduced to Equation (24) at most.

Table 8
Performance of the 13 models using the test set.

Risk	Metrics	Model												
		NN_1	NN_2	NN_3	NN_4	NN_5	NN_6	Linear	Ridge	Lasso	Polynomial	Combined_NN_1	Combined_NN_2	Combined_NN_3
Ransomware	MSE	2.06e-05	**1.27e-05**	1.48e-05	1.30e-05	1.91e-05	1.98e-05	2.02e-05	1.85e-05	4.22e-05	3.97e-05	2.01e-05	1.41e-05	1.99e-05
	RMSE	4.54e-03	**3.57e-03**	3.85e-03	3.60e-03	4.37e-03	4.45e-03	4.50e-03	4.31e-03	6.50e-03	6.30e-03	4.48e-03	3.75e-03	4.46e-03
	MAE	3.32e-03	2.63e-03	2.76e-03	**2.30e-03**	3.27e-03	3.28e-03	3.43e-03	3.22e-03	4.73e-03	4.78e-03	3.24e-03	2.66e-03	3.08e-03
	R^2	0.998	**0.999**	**0.999**	**0.999**	0.998	0.998	0.998	0.998	0.996	0.997	0.998	**0.999**	0.998
Phishing	MSE	6.12e-05	**4.97e-05**	7.82e-05	6.36e-05	7.58e-05	6.84e-05	1.27e-04	1.19e-04	1.23e-04	1.19e-04	6.42e-05	6.26e-05	5.99e-05
	RMSE	7.82e-03	**7.05e-03**	8.84e-03	7.97e-03	8.71e-03	8.27e-03	1.13e-02	1.09e-02	1.11e-02	1.09e-02	8.01e-03	7.91e-03	7.74e-03
	MAE	5.95e-03	**5.24e-03**	6.53e-03	5.40e-03	6.42e-03	5.87e-03	9.12e-03	8.76e-03	8.70e-03	8.58e-03	5.31e-03	5.89e-03	5.40e-03
	R^2	0.996	**0.997**	0.995	0.996	0.995	0.996	0.992	0.993	0.992	0.993	0.996	0.996	0.996
Insider attack	MSE	6.12e-06	2.76e-06	3.10e-06	3.36e-06	3.84e-06	2.96e-06	2.14e-06	**1.99e-06**	5.01e-06	4.73e-06	8.53e-06	9.44e-06	1.02e-05
	RMSE	2.47e-03	1.66e-03	1.76e-03	1.83e-03	1.96e-03	1.72e-03	1.46e-03	**1.41e-03**	2.24e-03	2.18e-03	2.92e-03	3.07e-03	3.20e-03
	MAE	1.83e-03	1.25e-03	1.35e-03	1.34e-03	1.54e-03	1.28e-03	1.09e-03	**1.05e-03**	1.83e-03	1.64e-03	2.32e-03	2.44e-03	2.43e-03
	R^2	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	**1**	0.999	0.999	0.999
Data breach	MSE	5.15e-05	5.62e-05	4.85e-05	5.24e-05	7.36e-05	5.15e-05	7.96e-05	7.38e-05	9.75e-05	1.07e-04	5.74e-05	4.36e-05	**3.66e-05**
	RMSE	7.18e-03	7.50e-03	6.96e-03	7.24e-03	8.58e-03	7.18e-03	8.92e-03	8.59e-03	9.87e-03	1.03e-02	7.58e-03	6.60e-03	**6.05e-03**
	MAE	4.33e-03	4.45e-03	4.44e-03	4.75e-03	6.18e-03	4.86e-03	6.89e-03	6.52e-03	7.24e-03	7.95e-03	4.46e-03	**3.83e-03**	4.11e-03
	R^2	0.995	0.994	0.995	0.995	0.992	0.995	0.992	0.992	0.99	0.989	0.994	0.995	**0.996**
Supply chain attack	MSE	1.24e-05	1.48e-05	1.15e-05	9.66e-06	7.14e-06	1.39e-05	9.92e-07	**8.93e-07**	1.98e-05	2.69e-05	9.59e-06	1.80e-05	1.42e-05
	RMSE	3.53e-03	3.84e-03	3.39e-03	3.11e-03	2.67e-03	3.72e-03	9.96e-04	**9.45e-04**	4.45e-03	5.19e-03	3.10e-03	4.24e-03	3.77e-03
	MAE	2.75e-03	2.90e-03	2.41e-03	2.24e-03	2.11e-03	2.55e-03	7.59e-04	**7.09e-04**	3.27e-03	3.76e-03	2.20e-03	2.61e-03	2.68e-03
	R^2	0.998	0.998	0.999	0.999	0.999	0.998	**1**	**1**	0.997	0.997	0.999	0.998	0.998

Note.
MSE (Mean Squared Error): Measures the average squared difference between the estimated values and what is estimated.
RMSE (Root Mean Squared Error): The square root of the mean of the square of all of the error. Used to measure how spread out the residuals are.
MAE (Mean Absolute Error): The average over the absolute differences between predicted values and actual values.
R^2 (Coefficient of Determination): Indicates the proportion of the variance in the dependent variable that is predictable from the independent variables. The interpretation of the R^2 testing results will be conducted in Section 5.1 to analyze the complexity of the cybersecurity landscape.

Table 9
Optimal base model selection results.

Risk	Metric	Best Performance	Best Models	Final Model
Ransomware (R_1)	MSE	1.27e-05	NN_2	NN_2
	RMSE	3.57e-03	NN_2	
	MAE	2.30e-03	NN_4	
	R^2	0.999	NN_2, NN_3, NN_4, Combined_Model_2	
Phishing (R_2)	MSE	4.97e-05	NN_2	NN_2
	RMSE	7.05e-03	NN_2	
	MAE	5.24e-03	NN_2	
	R^2	0.997	NN_2	
Insider attack (R_3)	MSE	1.99e-06	Ridge Regression	Ridge Regression
	RMSE	1.41e-03	Ridge Regression	
	MAE	1.05e-03	Ridge Regression	
	R^2	1	All except combined models	
Data breach (R_4)	MSE	3.66e-05	Combined_NN_3	Combined_NN_3
	RMSE	6.05e-03	Combined_NN_3	
	MAE	3.83e-03	Combined_NN_2	
	R^2	0.996	Combined_NN_3	
Supply chain attack (R_5)	MSE	8.93e-07	Ridge Regression	Ridge Regression
	RMSE	9.45e-04	Ridge Regression	
	MAE	7.09e-04	Ridge Regression	
	R^2	1	Linear Regression, Ridge Regression	

$$N_{\text{search space}} = \prod_{j=1}^{J} K^{(j)} \tag{23}$$

$$N_{\text{greedy search space}} = \sum_{j}^{m} K^{(j)} \tag{24}$$

Algorithm 2. Risk reduction strategy based on Greedy optimization

1: Initialize $j=1$, the risk factor with the most risk contribution
2: While $f\left(\boldsymbol{Pr}_j^{e_{j,k}}, \cdots, \boldsymbol{Pr}_J^{e_{J,k}} \middle| \boldsymbol{Pr}_1^{e_{1,k^*}}, \cdots, \boldsymbol{Pr}_{j-1}^{e_{j-1,k^*}}\right) > T$
3: $k^* = \underset{1 \le e_{j,k} \le K^{(j)}}{\arg\min} \left(f\left(\boldsymbol{Pr}_j^{e_{j,k}} = (0, \cdots, 1_{(k)}, \cdots, 0)_j, \boldsymbol{Pr}_{j+1}^{e_{j+1,k}}, \cdots, \boldsymbol{Pr}_J^{e_{J,k}} \middle| \boldsymbol{Pr}_1^{e_{1,k^*}}, \cdots, \boldsymbol{Pr}_{j-1}^{e_{j-1,k^*}}\right)\right)$
4: $\boldsymbol{Pr}_j^{e_{j,k^*}} = (0, \cdots, 1_{(k^*)} \cdots, 0)_j$
5: $j \leftarrow j+1$
6: m risk factors have been optimized; the risk reduction strategy is: $\left\{\boldsymbol{Pr}_1^{e_{1,k^*}}, \cdots, \boldsymbol{Pr}_j^{e_{j,k^*}}, \cdots, \boldsymbol{Pr}_m^{e_{m,k^*}}, \boldsymbol{Pr}_{m+1}^{e_{m+1,k}}, \cdots, \boldsymbol{Pr}_J^{e_{J,k}}\right\}$
7: End

4. Case study

To demonstrate the developed approach with three modules, we collaborated with Company A, a leading engineering and contracting firm in the U.A.E. (a subsidiary of a major investment corporation with over 12,000 employees). The company, known for its complex projects, provided data from a commercial building project in the UAE valued at over $5 million and lasting 24 months. During this period, the firm managed the construction phase. We explained each risk factor in online meetings and requested information on the 46 risk factors from the IT and construction teams, with a two-week deadline to return needed information. The real project data is shown in Table 11, with the "Scale" referring to the chosen scale's index, detailed in the Appendix. The teams reported encountering cyber risks, specifically phishing and data breaches. Phishing attacks occurred approximately every two months and were mostly intercepted by spam filters, though some occasionally bypassed these filters. These incidents were detected and communicated to staff to prevent data compromise, causing primarily internal disruptions. Effective filtering and alert systems prevented significant breaches. The data breach was due to unauthorized sharing of an internal memo by a staff member, not an external intrusion. This failure in data classification and awareness led to the premature public release of non-operational project branding details, affecting client relationships but resulting in no financial loss. The breach was detected manually shortly after the memo's release.

4.1. Module 1: risk prediction

The data from Table 11 were applied to the optimal model for each risk. Table 12 displays these risk degree predictions, highlighting phishing and data breaches as the most probable risks. Risks with predicted degrees higher than the 0.5 (a threshold used throughout this study) are classified into "occurrence" group. These predictions align with the actual events observed in the project, demonstrating a close match between predicted and actual occurrences. This consistency underscores the validity of our models.

4.2. Module 2: risk factor analysis

Based on the methods in Section 3.4, the feature contribution analysis for phishing and data breaches revealed the top 10 risk factors for each, as shown in Figs. 6 and 7. Notably, risk factor 2.2.2 significantly contributes to both risks, indicating that over 40 sub-teams at the project's second layer increase vulnerability to cyberattacks due to inadequate cybersecurity measures. This area requires urgent attention. More than 50% of the risk factors overlap between the two lists, suggesting that addressing these critical factors could mitigate multiple risks simultaneously. Category 1 factors, such as project budget (1.2) and duration (1.4), frequently rank high in contributing to phishing and data

Table 10
Sensitivity analysis results of ensemble labeling.

Weights		S_1 (Extra simulated project 1)					S_2 (Extra simulated project 2)				
FTA	Criteria	R_1	R_2	R_3	R_4	R_5	R_1	R_2	R_3	R_4	R_5
0.00	1.00	0.55	0.69	1	0.58	0.55	0.32	0.46	0.57	0.32	0.23
0.10	0.90	0.55	0.71	0.94	0.61	0.53	0.32	0.47	0.54	0.34	0.24
0.20	0.80	0.54	0.72	0.88	0.64	0.51	0.31	0.48	0.51	0.37	0.25
0.30	0.70	0.54	0.73	0.82	0.67	0.49	0.31	0.48	0.47	0.4	0.26
0.40	0.60	0.54	0.75	0.76	0.7	0.47	0.31	0.49	0.44	0.42	0.27
0.50	0.50	0.53	0.77	0.7	0.73	0.46	0.3	0.49	0.41	0.46	0.28
0.60	**0.40**	**0.53**	**0.78**	**0.64**	**0.76**	**0.44**	**0.3**	**0.51**	**0.38**	**0.48**	**0.29**
0.70	0.30	0.52	0.8	0.58	0.79	0.42	0.3	0.49	0.35	0.5	0.3
0.80	0.20	0.52	0.81	0.52	0.82	0.4	0.3	0.5	0.31	0.53	0.31
0.90	0.10	0.51	0.83	0.46	0.85	0.38	0.3	0.51	0.28	0.54	0.32
1.00	0.00	0.51	0.84	0.4	0.89	0.37	0.29	0.5	0.25	0.59	0.33

Notes: R_1 — Ransomware; R_2 — Phishing; R_3 — Insider attack; R_4 — Data breach; R_5 — Supply chain attack.

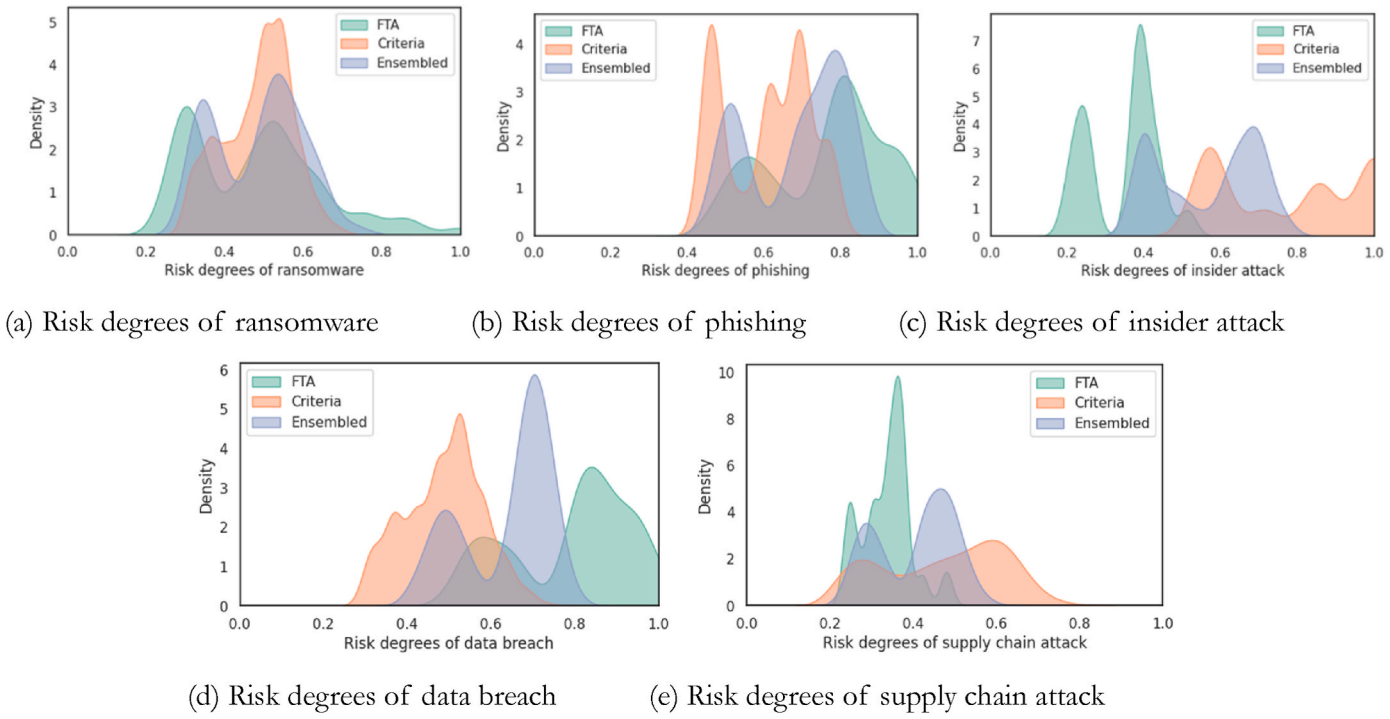


Fig. 4. KDE of the labeled risk degrees (FTA Weight: 0.5, Criteria Weight: 0.5).

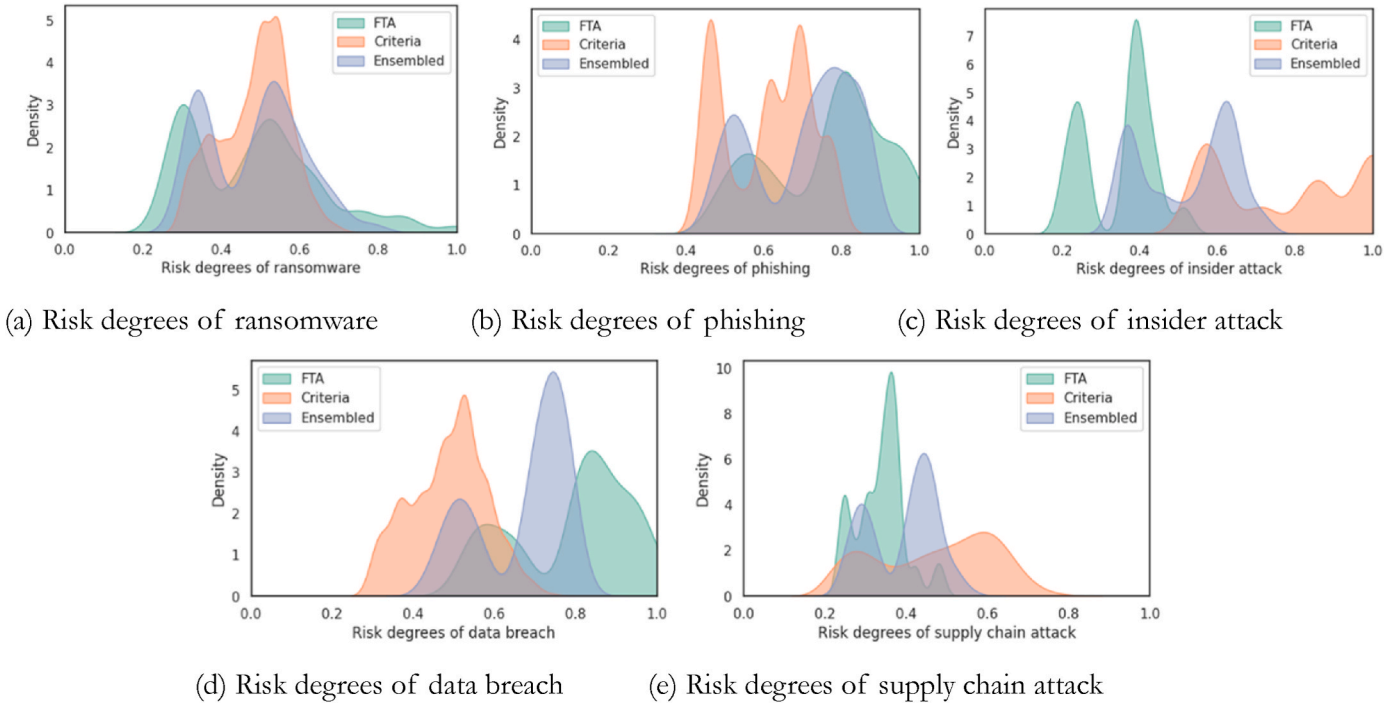


Fig. 5. KDE of the labeled risk degrees (FTA Weight: 0.6, Criteria Weight: 0.4).

breach risks, highlighting the need for integrating cybersecurity into early project planning. IT-related risks in Category 3 consistently appear in both lists, underscoring the importance of robust IT strategies for project cybersecurity. Categories 4 and 5 have minimal representation in the top 10, indicating that Construction Company A effectively manages risks associated with operational technology and human management.

4.3. Module 3: risk reduction strategy

Using the greedy optimization algorithm from Section 3.5, we mitigated key cyber risks: phishing and data breaches. This process involved prioritizing significant risk factors contributing to these risks. Fig. 8 displays the risk reduction results, with the x-axis representing the number of addressed risk factors and the y-axis showing the updated risk degree prediction. Each graph highlights the top three contributing

Table 11
Project data from Construction Company A.

Risk factor	1.1	1.2	1.3	1.4	1.5	1.6	1.7	2.1	2.2.1	2.2.2	2.2.3	2.2.4	2.2.5	2.2.6	2.2.7	2.2.8	2.3.1	2.3.2	2.3.3	2.3.4	2.3.5	2.3.6	2.3.7
Scale	0	4	2	4	5	3	1	1	2	4	5	5	5	5	5	5	1	3	7	7	7	7	7
Risk factor	2.3.8	2.4	3.1	3.2	3.3	3.4	3.5	3.6	3.7	3.8	3.9	4.1	4.2	4.3	4.4	4.5	5.1	5.2	5.3	5.4	5.5	5.6	5.7
Scale	7	0	4	2	0	4	4	1	2	2	1	5	1	0	2	1	1	0	0	1	1	0	1

Table 12
Comparison of predicted and actual risk occurrences.

Risk	Optimal Model	Predicted Risk Degree	Predicted Occurrence	Actual Occurrence
Ransomware	NN_2	0.47	No	No
Phishing	NN_2	0.70	Yes	Yes
Insider attack	Ridge Regression	0.48	No	No
Data breach	Combined_NN_3	0.64	Yes	Yes
Supply chain	Ridge Regression	0.39	No	No

factors for each risk. Detailed descriptions of these factors and how they have been addressed are provided in Table 13. For instance, for phishing, Risk Factor 1.5, which considers the total number of non-labor project participants, recommends selecting the risk factor scale with index 1, corresponding to a group size of 51–100 people, as detailed in the Appendix. By addressing the top 14 risk factors for phishing, the risk degree can be reduced to 0.481, and for data breaches, addressing the top 9 factors lowers the risk degree to 0.497. It is evident from the results that, for both phishing and data breaches, the lowest achievable risk degree exceeds 0.4 (0.418 for phishing and 0.404 for data breach as shown in the figure). This indicates that each project has inherent cybersecurity risks that cannot be eliminated. Thus, relevant parties should focus on controlling the risk factors to maintain project cyber risk at a tolerable level. If a more conservative approach is preferred, the risk threshold can be adjusted based on the project manager's criteria, the project's risk tolerance, and the feasibility and cost of mitigating these risks. For risks predicted not to occur, continuous monitoring is advised throughout the project. This approach involves using the trained models for ongoing risk predictions as project data is updated in subsequent project phases.

5. Discussions

5.1. The complexity of cybersecurity landscape

To explore the model behavior, R^2 values were extracted form Table 8 and plotted in Fig. 9. R^2 assesses the variance in labels explained by the model, indicating the linearity between training data and labels. High R^2 values suggest effective variance capture by the model. For simpler models like regression, high R^2 indicates a strong linear relationship between the dataset and labels. Conversely, low R^2 values in simpler models, but high in complex models like neural networks, imply significant non-linearity in the dataset.

Fig. 9 shows that the insider attack risk consistently achieves near-perfect R^2 values across all models, including simpler ones like Linear, Ridge, and Lasso regression, indicating a predominantly linear relationship and making basic models suitable for prediction. Similarly, supply chain attack risk also demonstrates high R^2 values with simpler models, confirming linearity. In cases of obvious linearity, complex models tend to overfit and underperform on test data, affirming our choice of simpler models for these risks. For ransomware attack, all models achieve high R^2 values, but complex models like neural networks outperform simpler ones, suggesting less linearity and more non-linearity. This pattern extends to phishing and data breach risks, with data breach showing the lowest R^2 values across all models, indicating significant non-linearity and the limited effectiveness of even sophisticated models. Thus, complex models like neural networks are more suitable for these risks, justifying our selection for predicting their degrees.

From the analysis above, insights into the cybersecurity landscape can be derived; the linearity analysis highlights the complexity of mapping risk factors to cyber risks. Nearly all risks exhibit some non-linearities, signifying a complex relationship between risk factors and cyber risks. This necessitates that project managers in the construction

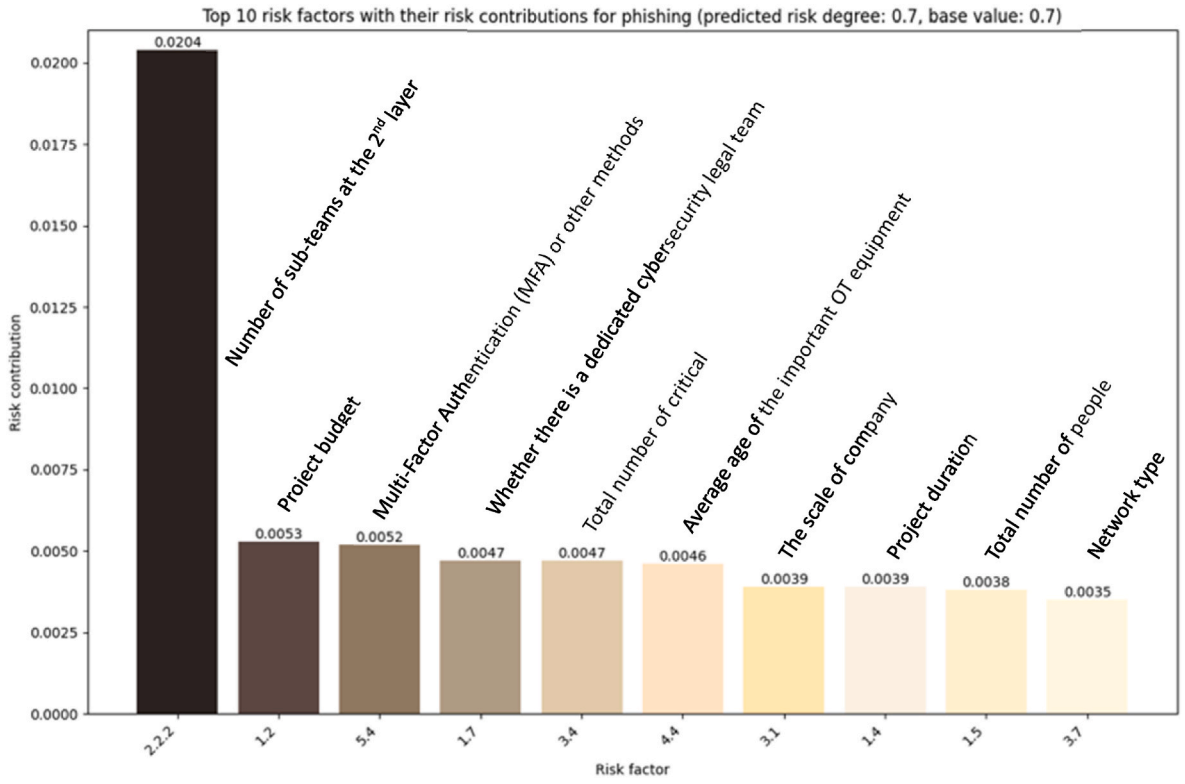


Fig. 6. Top 10 risk factors with their risk contributions for phishing.

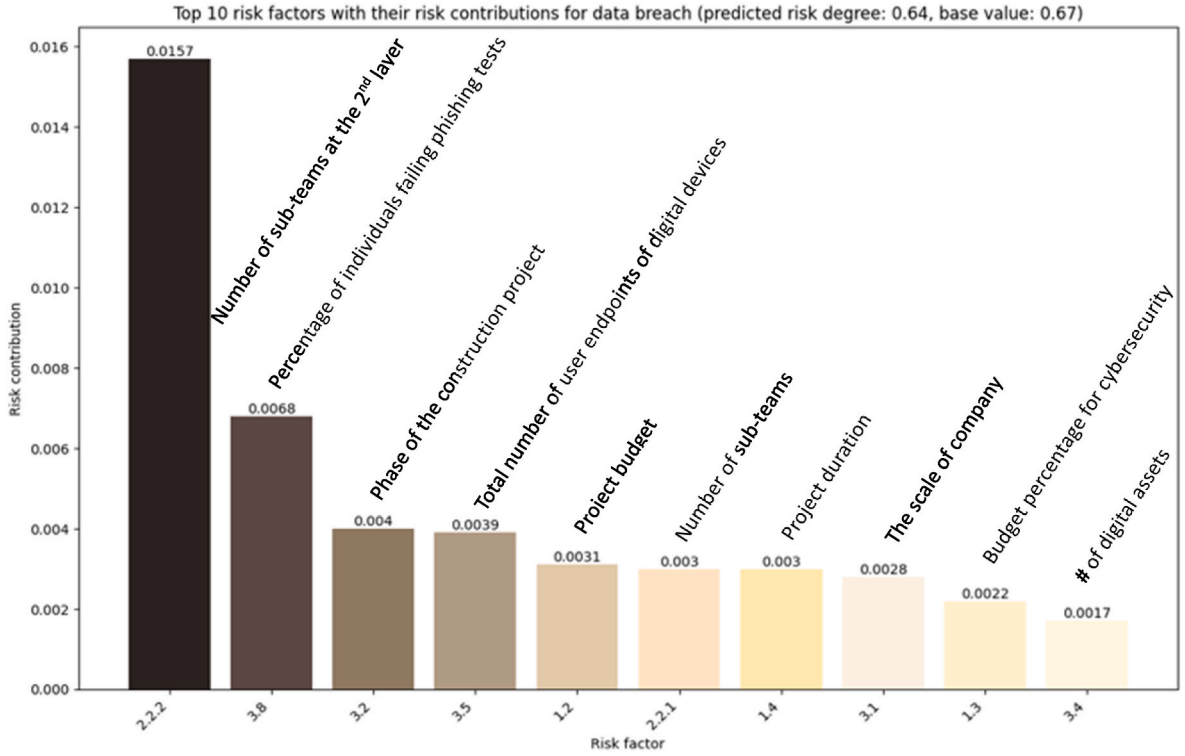
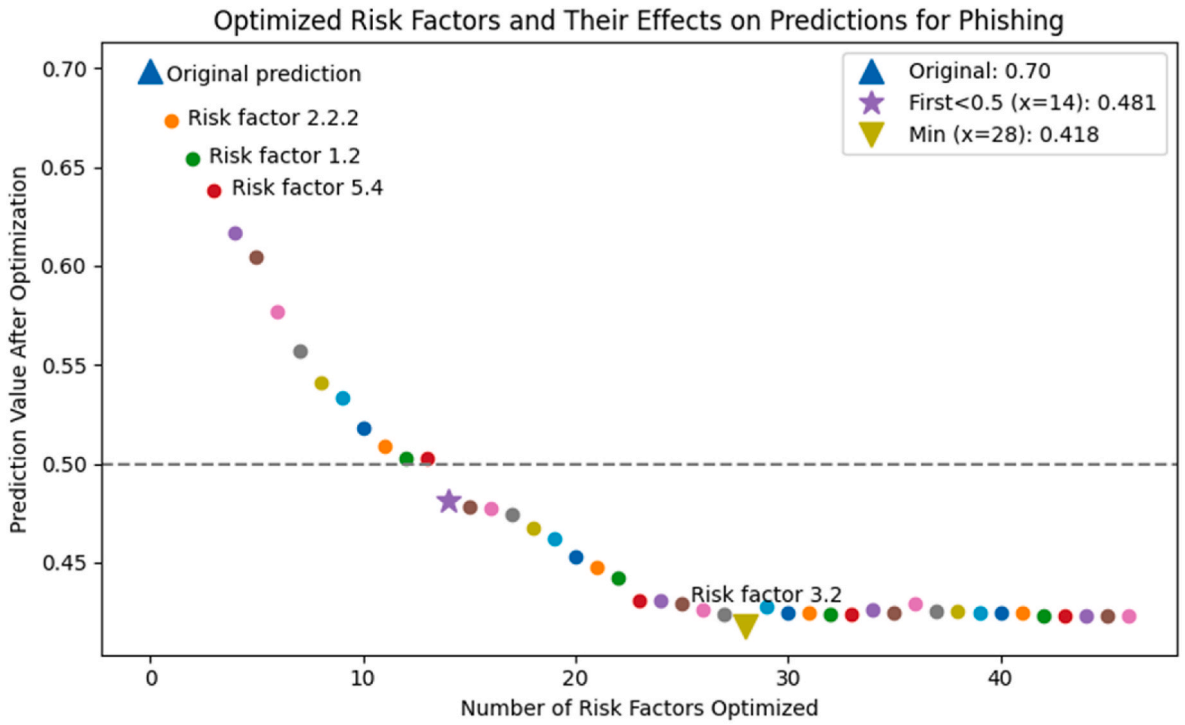
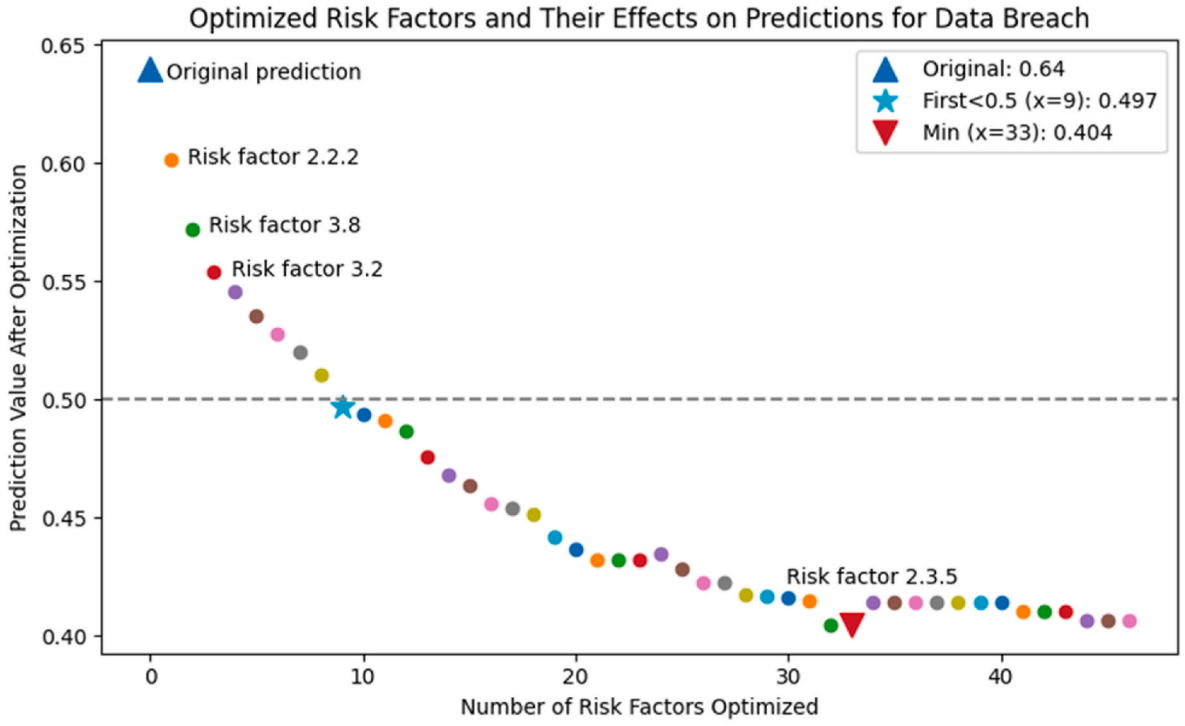


Fig. 7. Top 10 risk factors with their risk contributions for data breach.

industry have a comprehensive understanding of these risk factors to gain deeper insights into cyber risks, which in turn requires a commitment to cybersecurity and a holistic approach. Additionally, different cyber risks demonstrate distinct non-linear relationships with their respective risk factors, suggesting that effectively addressing each type of cyber risk may require strategies specifically tailored to these unique relationships. Due to these non-linearities, the effective reduction of cyber risk necessitates a coordinated approach that encompasses



(a) Phishing



(b) Data Breach

Fig. 8. Predicted risk degrees against the number of optimized risk factors.

multiple risk factors to achieve significant risk reduction, as addressing individual factors in isolation may not result in substantial risk mitigation.

5.2. *Risk factor of general importance*

To identify the risk factors that are generally important to a wide range of construction projects, their importance can be analyzed. The importance of risk factor j was determined as the mean absolute SHAP

Table 13
The risk predictions against the number of optimized risk factors.

Phishing	No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
	Risk factor	2.2.2	1.2	5.4	1.7	3.4	4.4	1.4	3.1	1.5	3.7	3.8	4.1	3.5	2.2.5	1.6	2.2.1	1.3	4.3	2.2.6	2.2.7	2.3.2	2.2.8	2.2.4
	Scale*	0	0	0	0	0	0	1	0	1	0	0	0	4	0	1	1	1	1	0	1	2	0	0
	Predicted risk	0.673	0.654	0.638	0.617	0.604	0.577	0.557	0.541	0.534	0.518	0.509	0.503	0.503	0.481	0.478	0.478	0.475	0.468	0.462	0.453	0.448	0.442	0.431
	No.	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
	Risk factor	2.3.8	3.9	5.5	2.3.3	3.2	2.3.6	2.2.3	5.3	2.3.4	4.5	2.3.7	5.6	2.3.5	5.1	5.2	2.1	3.3	3.6	1.1	2.4	4.2	5.7	2.3.1
	Scale*	1	0	2	2	1	1	0	0	1	1	1	4	0	2	0	7	0	1	4	0	1	0	1
	Predicted risk	0.431	0.43	0.426	0.424	0.418	0.428	0.425	0.425	0.424	0.424	0.426	0.424	0.429	0.425	0.425	0.424	0.424	0.424	0.423	0.423	0.423	0.423	0.423
Data breach	No.	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21	22	23
	Risk factor	2.2.2	3.8	3.2	3.5	1.2	1.4	2.2.1	3.1	1.3	3.4	4.1	1.5	3.7	4.4	1.7	1.6	5.4	5.5	2.2.7	2.2.8	4.3	3.9	5.3
	Scale*	0	0	1	0	0	1	0	0	1	0	0	0	0	0	0	1	0	0	0	0	1	1	0
	Predicted risk	0.601	0.572	0.554	0.546	0.535	0.528	0.52	0.51	0.497	0.494	0.491	0.487	0.476	0.468	0.464	0.456	0.454	0.452	0.442	0.436	0.432	0.432	0.432
	No.	24	25	26	27	28	29	30	31	32	33	34	35	36	37	38	39	40	41	42	43	44	45	46
	Risk factor	2.3.6	2.3.2	2.2.5	4.5	2.2.6	2.3.8	2.3.3	2.3.4	2.2.4	2.3.5	2.3.7	5.6	1.1	3.6	2.3.1	5.2	5.7	5.1	2.1	4.2	2.2.3	3.3	2.4
	Scale*	2	2	1	1	0	1	1	1	0	0	1	0	0	1	1	0	1	2	1	1	0	2	0
	Predicted risk	0.435	0.428	0.422	0.422	0.417	0.417	0.416	0.415	0.404	0.404	0.415	0.415	0.415	0.415	0.415	0.415	0.415	0.41	0.41	0.41	0.407	0.407	0.407

Note: "Scale*" means the index of the optimal scale of the optimized risk factor.

value for all its scales, averaged across all project cases in the testing dataset, described in Equation (25). To compare risk factor importance uniformly across models, we normalized the importance values over the 46 risk factors, as shown in Equation (26). As a reminder, the SHAP value computation equation is presented in Equation (20).

$$I_j = \frac{1}{N \bullet K^{(j)}} \sum_{i=1}^{N} \sum_{k=1}^{K^{(j)}} \left| \mathrm{SHAP}_{i,j,k} \right| \tag{25}$$

$$\widetilde{I}_j = \frac{I_j}{\sum_{j=1}^{46} I_j} \tag{26}$$

After computing the importance, Table 14 presents the top 10 risk factors for each risk helping project managers understand the relative significance of these factors. This allows them to determine the necessary measures for addressing critical risks. Note that these values primarily serve as a comparison indicator to rank the importance of the risk factors and do not have specific meanings or significant implications. Managers focusing on specific risks can use the ranked lists to develop targeted prevention strategies, while those seeking a holistic approach can refer to Table 15, which highlights the ten risk factors that appear most frequently across the five lists, emphasizing their importance in the broader cybersecurity landscape of construction projects.

From Table 15, several key insights can be derived and discussed.

(1) IT-related risk factors are the most impactful in construction cybersecurity risk predictions. A dedicated IT team (risk factor 3.3) is crucial for continuous threat monitoring and swift response. The choice of private networks (risk factor 3.7) significantly reduces vulnerability to external threats, while anti-phishing training (risk factor 3.8) greatly enhances employee awareness and minimizes breaches from human error. Project managers should prioritize these factors for effective resource allocation in cybersecurity planning.

(2) Factors related to project structure, such as the number of stakeholders across layers (risk factor 2.2) and team overlap across projects (risk factor 2.4), significantly affect cybersecurity. The distribution of teams must be managed carefully, as varying stakeholder perspectives can lead to inconsistent defenses. Striking a balance between team distribution and potential vulnerabilities is crucial. Minimizing team overlap is also essential, as it can complicate information security management and create execution gaps. Allocating team members to multiple projects may dilute focus, leading to errors, data breaches, and non-compliance with cybersecurity standards.

(3) Factors related to management side such as the implementation of MFA (risk factor 5.4) is crucial cybersecurity measures. MFA, including biometrics or face recognition, significantly enhances security by adding layers of verification, reducing unauthorized access risks.

(4) A dedicated cybersecurity legal team (risk factor 1.7) is also important for cyber risk management because it ensures compliance with evolving legal requirements and standards. This team can proactively identify potential legal issues related to cybersecurity, advise on risk mitigation strategies, and help navigate the complex landscape of international cyber laws, thus protecting the organization from legal and financial penalties.

(5) The phase of a construction project (risk factor 3.2) is also a determinant of its cybersecurity status. This observation is consistent with the understanding that different project phases may have varying types of threats and vulnerabilities and also varying levels of them, as demonstrated in the work (Mantha et al., 2021). Therefore, as a project progresses through its phases, a tailored set of cybersecurity plans and measures should be

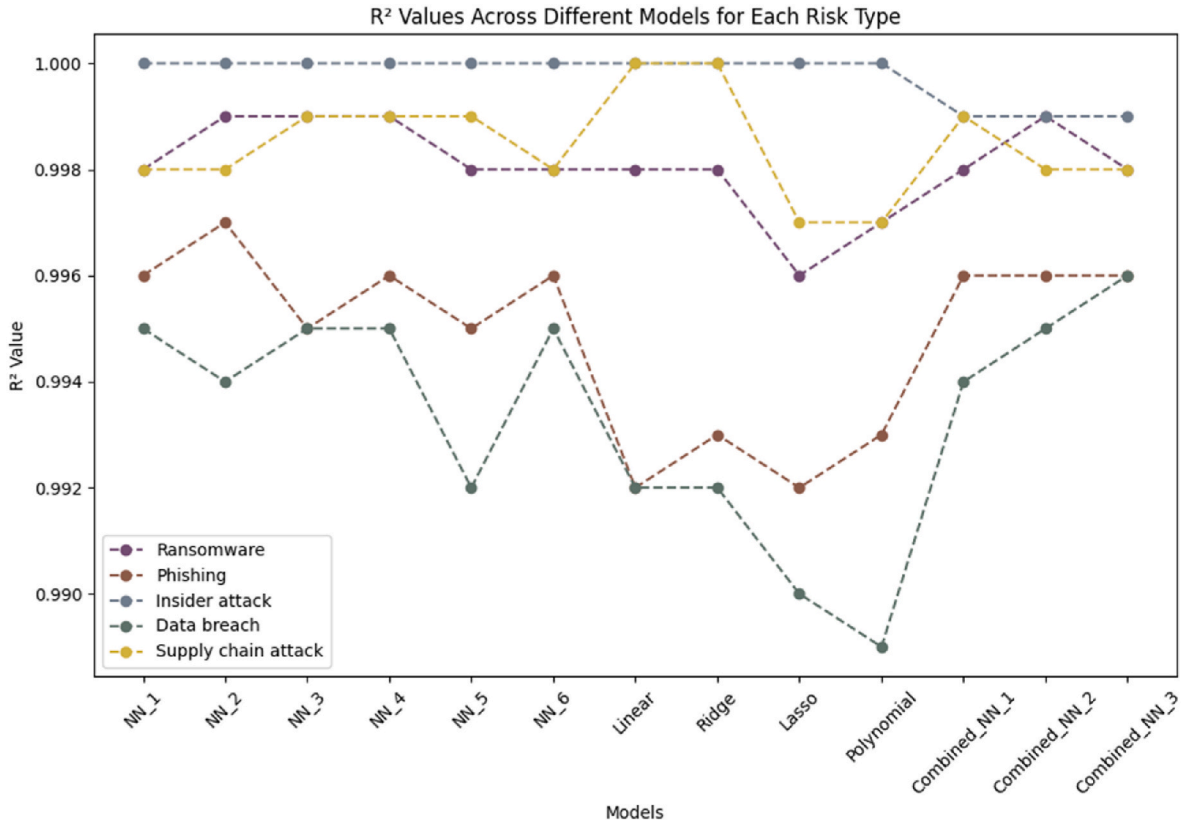


Fig. 9. R^2 values across different models for five cyber risks.

Table 14
Top 10 risk factors for each risk.

Ranking No. (from the most to the least important)		1	2	3	4	5	6	7	8	9	10
Ransomware	Risk factor	3.3	5.3	2.2.2	1.7	2.2.3	3.8	5.4	3.7	2.2.4	3.2
	Importance	0.0485	0.0402	0.0402	0.0379	0.0365	0.0338	0.0326	0.0304	0.0303	0.0281
Phishing	Risk factor	3.7	3.3	5.3	5.4	2.4	1.4	1.7	2.2.1	4.5	3.6
	Importance	0.0449	0.0440	0.0395	0.0388	0.0383	0.0365	0.0340	0.0323	0.0311	0.0303
Insider attack	Risk factor	3.3	3.7	1.7	5.1	3.8	3.2	1.6	1.4	3.1	5.4
	Importance	0.1193	0.0626	0.0587	0.0576	0.0555	0.0533	0.0412	0.0386	0.0346	0.0335
Data breach	Risk factor	2.2.2	2.2.3	3.7	3.3	2.2.4	2.4	3.1	3.6	1.7	2.2.5
	Importance	0.0565	0.0446	0.0434	0.0374	0.0356	0.0336	0.0313	0.0280	0.0268	0.0268
Supply chain attack	Risk factor	5.4	2.2.2	2.2.3	2.4	2.2.5	3.3	3.8	2.2.4	5.1	2.2.6
	Importance	0.1050	0.0411	0.0394	0.0368	0.0357	0.0349	0.0348	0.0332	0.0320	0.0304

Table 15
Top 10 risk factors based on appearance frequency.

Risk factor	3.3	5.4	1.7	3.7	2.2.4	2.4	3.8	2.2.3	2.2.2	3.2
Appearance frequency	5	4	4	4	3	3	3	3	3	2

prepared and implemented to address the evolving risk landscape.

5.3. The practicality and prospect of the models

In Section 3.3.3, we conducted KDE to demonstrate that our ensemble labeling methods produce a convincing dataset for training. The results underscore the efficacy of our ensemble labeling method, indicating that it is unbiased and accurately reflects industry realities. This proves that our final selected models are reliable and have effectively encoded real-world information. Additionally, the developed models successfully predicted the occurrence of cyber risks in two projects labeled by experts (Table 7) and in a real construction project (Section 4). These outcomes further demonstrate the validity and effectiveness of our models. In Section 4.2, the risk factor contribution analysis conducted through the models accurately identifies the factors contributing to the risks and the extent of their contributions. This analysis is invaluable for project managers, particularly during the project lifecycle. The models enable the prediction of cyber risk status at any stage of a construction project, allowing project managers to implement immediate risk reduction strategies. These strategies are informed by the model's greedy optimization algorithm, which aims to maximize resource allocation efficiency.

5.4. Limitations and future works

Our study's primary limitation is the absence of an existing dataset, leading us to use a simulated dataset based on defined probability distributions (*Po* in Section 3.2.1). Despite expert review and validation, these distributions might not fully mirror real-world scenarios, potentially introducing variance in our model results. Future plans include conducting sensitivity analyses to refine these distributions and expanding expert panel for the data simulation process, especially if simulations continue to be necessary. Additionally, this study uses a single real project to demonstrate the applicability of the approach; however, this is insufficient to establish the generalizability of the models. We are actively collaborating with local companies to collect additional authentic data for model validation and pursuing partnerships to access real-world data. Our ultimate goal is to replace the simulated dataset with authentic data to ensure the integrity of the models and enhance the validation process. Our long-term objective is to develop web and mobile applications for cyber risk assessment, specifically designed for construction practitioners and project managers. These applications will utilize our trained models, allowing users to input project details and receive tailored risk predictions and mitigation strategies directly.

6. Conclusions

This study developed an ML-centric approach for assessing the five most common cyber risks in construction projects: ransomware, phishing, insider attacks, data breaches, and supply chain attacks. The developed approach consists of three components: (1) dynamically predicting cyber risks throughout construction project progressions, (2) identifying the highest contributing risk factors, and (3) suggesting efficient strategies for risk reduction. A case study involving a real construction project was conducted to demonstrate the applicability of the approach. This study elucidates the complex relationship between risk factors and cyber risks in construction, emphasizing the need for project managers to deeply understand and strategically manage these risks. Effective cyber risk reduction depends on tailored strategies and a coordinated approach addressing multiple factors. It identifies critical risk factors, especially IT-related issues like the need for a dedicated IT team, the use of private networks, and comprehensive anti-phishing training. The study also stresses the importance of project-specific security measures, minimal personnel overlap between projects for improved security, and robust management practices such as MFA to establish a strong, multi-layered defense. Additionally, it highlights the role of a dedicated cybersecurity legal team in maintaining compliance with evolving legal standards and suggests tailoring cybersecurity plans to each project phase to adapt to changing risk landscapes throughout the project lifecycle. Future work will concentrate on replacing simulated data with real-world data to further improve the models' accuracy and applicability. Furthermore, we plan to develop web and mobile applications for cyber risk management tailored for practitioners and project managers. These applications will offer models that are regularly updated with new, real data, allowing users to input project details and instantly obtain risk predictions along with suggested strategies for reducing risks.

CRediT authorship contribution statement

Dongchi Yao: Writing – original draft, Visualization, Methodology, Formal analysis, Data curation, Conceptualization. **Borja García de Soto:** Writing – review & editing, Validation, Supervision, Project administration, Investigation.

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.

Acknowledgments

This work was supported by the Center for Cyber Security (CCS), funded by Tamkeen under the NYUAD Research Institute Award G1104. It was conducted in collaboration with the NYUAD Center for Interacting Urban Networks (CITIES), which is funded by Tamkeen under the NYUAD Research Institute Award CG001. We extend our gratitude to the experts at ALEC Engineering & Contracting LLC (ALEC) in Dubai, particularly Mr. Sabyasachi Jana, for providing the project data. Additionally, we appreciate the assistance of SecurityScorecard in New York, especially Prof. Mike Wilkes, for providing valuable feedback on the data simulation, and on the development and selection of our model.

Appendix. Table

Table A1
Summary of the risk factors

Category	No.	Risk factor	Scales	Scale Occurrence Probability Distribution (P_o)	Scale Risk Distribution (R)	Number of Actual Scales
1. Overall information of the project	1.1	What is the country of the project?	[Asia, Europe, Africa, North America, South America, Antarctica, and Oceania]. Although initially we planned to use countries as indexes, for simplicity we finally used continents.	[0.3, 0.14, 0.2, 0.14, 0.14, 0.04, 0.04]*	[0.8, 0.7, 0.9, 0.65, 0.7, 0.6, 0.7]**	7
	1.2	What is the project budget?	[≤ $100,000, $100,000 - $500,000, $500,000 - $1 million, $1 million - $5 million, > $5 million]	[0.25, 0.25, 0.2, 0.2, 0.1]*	[0.2, 0.4, 0.6, 0.8, 1.0]*	5
	1.3	What is the percentage of total project budget for cybersecurity management?	[≤1%, 1%–2%, 2%–3%, 3%–4%, 4%–5%, >5%]	[0.1, 0.15, 0.25, 0.25, 0.15, 0.1]*	[1.0, 0.83, 0.67, 0.50, 0.33, 0.17]*	6
	1.4	What is the project duration?	[≤3 months, 3–6 months, 6–12 months, 12–24 months, >24 months]	[0.1, 0.2, 0.35, 0.25, 0.1]*	[0.2, 0.4, 0.6, 0.8, 1.0]*	5

(*continued on next page*)

Table A1 (*continued*)

Category	No.	Risk factor	Scales	Scale Occurrence Probability Distribution (P_o)	Scale Risk Distribution (R)	Number of Actual Scales
	1.5	What is the total number of people involved in the project (labor excluded)?	[≤50, 51–100, 101–200, 201–300, 301–400, >400]	[0.2, 0.25, 0.25, 0.15, 0.1, 0.05]*	[0.17, 0.33, 0.50, 0.67, 0.83, 1.0]*	6
	1.6	What is the project type?	[Transportation Infrastructure Projects, Government Projects, Healthcare Projects, Large-Scale Commercial Projects, Residential Projects, Other types]	[0.15, 0.2, 0.1, 0.2, 0.3, 0.05]*	[0.17, 0.33, 0.50, 0.67, 0.83, 1.0]*	6
	1.7	Whether there is a dedicated cybersecurity legal team for the project?	[Yes, No, Unsure]	[0.3,0.6,0.1]*	[0.4, 1, 0.7] **	3
2. Project structure	2.1	What is the project delivery method?	[Design-Bid-Build (DBB), Design-Build (DB), Construction Manager at Risk (CMAR), Construction Management Multi-Prime (CMMP), Public-Private Partnership (PPP or P3), Integrated Project Delivery (IPD), Design/Build/Operate/Maintain (DBOM), Other types]	[0.2, 0.15, 0.1, 0.1, 0.1, 0.2, 0.1, 0.05]*	[0.7, 0.8, 0.7, 0.7, 0.7, 0.9, 0.7] **	8
	2.2 (2.2.1–2.2.8)	What is the number of sub-teams at different layers of the project? (The number of sub-teams at different layers of a project indicates the project's structural complexity and associated cyber risks. Outer-layer sub-teams, often smaller with fewer cybersecurity resources, are generally more vulnerable, while inner-layer teams are typically larger, better equipped, and more security-conscious. Understanding the distribution of sub-teams across layers is crucial for implementing effective, targeted risk management strategies.)	Eight layers, each layer's scales are: [≤10, 11–20, 21–30, 31–40, >40, N/A], "N/A" means this layer is not existent	[0.1, 0.2, 0.3, 0.25, 0.15, 0] [0.25, 0.3, 0.15, 0.15, 0.1, 0.05] [0.35, 0.25, 0.15, 0.1, 0.05, 0.1] [0.4, 0.2, 0.1, 0.1, 0.05, 0.15] [0.45, 0.15, 0.075, 0.05, 0.025, 0.25] [0.5, 0.15, 0.055, 0.02, 0.025, 0.25] [0.55, 0.1, 0.025, 0.015, 0.01, 0.3] [0.6, 0.025, 0.01, 0.01, 0.005, 0.35] **	[0.2, 0.4, 0.6, 0.8, 1.0, 0.0]* The same for other layers	47, excluding the one with occurrence probability of 0
	2.3 (2.3.1–2.3.8)	What is the number of communication channels at different layers in the model? (This factor assesses cybersecurity risks by examining how many pathways for data exchange exist among teams within each layer. A greater number can increase vulnerabilities due to more potential points of data interception. This factor necessitates implementing robust security protocols to safeguard sensitive information and ensure the integrity of systems across each layer.)	Eight layers, each layer's scales are: [≤50,≤100,≤150,≤200, <250,≤300, >300, N/A], "N/A" means this layer is not existent	[0.1, 0.15, 0.25, 0.2, 0.15, 0.1, 0.05, 0] [0.13, 0.16, 0.18, 0.17, 0.12, 0.09, 0.045, 0.03] [0.16, 0.17, 0.16, 0.15, 0.1, 0.08, 0.046, 0.06] [0.19, 0.18, 0.14, 0.13, 0.08, 0.07, 0.044, 0.09] [0.22, 0.19, 0.12, 0.11, 0.06, 0.06, 0.042, 0.12] [0.25, 0.2, 0.1, 0.09, 0.04, 0.05, 0.042, 0.05] [0.28, 0.21, 0.08, 0.07, 0.02, 0.04, 0.038, 0.18] [0.31, 0.22, 0.06, 0.05, 0.01, 0.03, 0.036, 0.21] **	[0.14, 0.29, 0.43, 0.57, 0.71, 0.86, 1.0, 0.0]* The same for other layers	63, excluding the one with occurrence probability of 0
	2.4	What is the percentage of teams overlapping in different projects? (The percentage of teams overlapping in different projects refers to the proportion of teams engaged simultaneously in multiple initiatives. This overlap can elevate cyber risks due to shared resources and personnel, which may introduce security gaps and dependencies. Increased team overlap heightens the likelihood of breaches and necessitates robust management strategies to address these vulnerabilities effectively.)	[≤20%, 21%–40%, 41%–60%, 61%–80%, 81%–100%]	[0.25, 0.3, 0.25, 0.15, 0.05] **	[0.2, 0.4, 0.6, 0.8, 1.0]*	5

(*continued on next page*)

Table A1 (*continued*)

Category	No.	Risk factor	Scales	Scale Occurrence Probability Distribution (P_o)	Scale Risk Distribution (R)	Number of Actual Scales
3. IT factors	3.1	What is the scale of your company?	[≤30, 31–60, 61–100, 101–150, >150]	[0.3, 0.35, 0.2, 0.1, 0.05]*	[1.0, 0.8, 0.6, 0.4, 0.2]*	5
	3.2	What is the phase of the construction project when your company is involved?	[Planning and Bidding phase, Design phase, Construction phase, Maintenance & Operation phase, Demolition phase]	[0.15, 0.20, 0.25, 0.3, 0.1] **	[0.6, 1, 0.8, 0.6, 0.2] **	5
	3.3	Is there a dedicated IT team for the project?	[Yes, No, Unsure]	[0.35, 0.55, 0.1]*	[0.2, 1, 0.6] **	3
	3.4	What is the total number of critical digital assets?	[≤50, 51–200, 201–400, 401–600, >600]	[0.20, 0.30, 0.30, 0.15, 0.05]*	[0.2, 0.4, 0.6, 0.8, 1.0]*	5
	3.5	What is the total number of user endpoints of digital devices for the project? (User endpoints refer to all digital devices, such as laptops, smartphones, and desktops, connected within a network. These devices serve as potential entry points for cyber threats, increasing the network's vulnerability to security breaches.)	[≤50, 51–200, 201–400, 401–600, >600]	[0.20, 0.30, 0.30, 0.15, 0.05]*	[0.2, 0.4, 0.6, 0.8, 1.0]*	5
	3.6	What is the percentage of digital devices with firewalls or intrusion detection systems involved in the project?	[≤20%, 21%–40%, 41%–60%, 61%–80%, 81%–100%]	[0.05, 0.15, 0.25, 0.30, 0.25]*	[1.0, 0.8, 0.6, 0.4, 0.2]*	5
	3.7	What is the network type used for the project: Public or Private? (Choosing between a public or private network impacts a project's cybersecurity. Public networks can be vulnerable to attacks due to easier access, while private networks offer enhanced security controls but may be costly and complex to manage. Each type requires specific security approaches.)	[Public network, Private network, Both public and private network]	[0.3, 0.4, 0.3] **	[1, 0.2, 0.6] **	3
	3.8	What is the percentage of individuals who fail phishing tests after completing mandatory training?	[≤20%, 21%–40%, 41%–60%, 61%–80%, 81%–100%]	[0.35, 0.30, 0.20, 0.10, 0.05] **	[0.2, 0.4, 0.6, 0.8, 1.0]*	5
	3.9	What is the estimated Mean Time to Respond (MTTR) in hours?	[Within 1 h, 1–4 h, 4–8 h, 8–24 h, Above 24 h]	[0.10, 0.30, 0.30, 0.20, 0.10] **	[0.2, 0.4, 0.6, 0.8, 1.0]*	5
4. OT factors	4.1	What is the total number of important OT equipment involved?	[≤30, 31–60, 61–90, 91–120, 121–150, >150]	[0.30, 0.25, 0.20, 0.15, 0.07, 0.03]*	[0.17, 0.33, 0.50, 0.67, 0.83, 1.0] *	6
	4.2	What is the level of physical access control mechanism to OT equipment? (Evaluating the level of physical access control mechanisms to OT equipment involves assessing the measures in place to prevent unauthorized physical access to sensitive systems. This includes secure entry points, surveillance systems, ID badges, biometric verification, and visitor audits. Higher levels of control suggest better security, crucial for protecting sensitive data and maintaining overall cybersecurity.)	[Level 1, Level 2, Level 3, Level 4, Level 5]	[0.10, 0.20, 0.30, 0.25, 0.15] **	[1.0, 0.8, 0.6, 0.4, 0.2] *	5
	4.3	What is the percentage of OT equipment isolated from project's general network?	[≤20%, 21%–40%, 41%–60%, 61%–80%, 81%–100%]	[0.05, 0.15, 0.30, 0.30, 0.20]*	[1.0, 0.8, 0.6, 0.4, 0.2] *	5
	4.4	What is the average age of the important OT equipment, in years?	[≤1, 1–3, 4–7, 8–10, >10]	[0.25, 0.35, 0.25, 0.10, 0.05]*	[0.2, 0.4, 0.6, 0.8, 1.0] *	5
	4.5	What is the level of authentication mechanism to access the HMI (Human Machine Interface)?	[Level 1, Level 2, Level 3, Level 4, Level 5]	[0.10, 0.40, 0.25, 0.15, 0.10] **	[1.0, 0.8, 0.6, 0.4, 0.2] *	5

(*continued on next page*)

Table A1 (*continued*)

Category	No.	Risk factor	Scales	Scale Occurrence Probability Distribution (P_o)	Scale Risk Distribution (R)	Number of Actual Scales
5. Management and human factors	5.1	What is the average level of commitment to corporate governance, ethical practices and cybersecurity policy?	[Level 1, Level 2, Level 3, Level 4, Level 5]	[0.05, 0.10, 0.20, 0.35, 0.30] **	[1.0, 0.8, 0.6, 0.4, 0.2] *	5
	5.2	What is the average frequency of security training per year?	[≤10, 11–20, 21–30, 31–40, 41–50, >50]	[0.20, 0.30, 0.20, 0.15, 0.10, 0.05]*	[1.0, 0.83, 0.67, 0.50, 0.33, 0.17] *	6
	5.3	Do you allow password reuse for any project-related software, systems, or accounts (e.g., project management tools, email, internal networks, file storage, etc.)?	[Yes, No]	[0.7, 0.3]*	[1, 0.2] **	2
	5.4	Does internet access within your construction project require Multi-Factor Authentication (MFA) or utilize other methods such as biometrics or face recognition?	[Yes, No]	[0.6, 0.4]*	[0.2, 1] **	2
	5.5	What is the percentage of people who have access to sensitive information in the project?	[≤10%, 11%–30%, 31%–50%, 51%–70%, 71%–90%, 91%–100%]	[0.45, 0.35, 0.10, 0.05, 0.03, 0.02]*	[0.17, 0.33, 0.50, 0.67, 0.83, 1.0] *	6
	5.6	What is the average team member variability over a 3-month period?	[≤20%, 20%–40%, 40%–60%, 60%–80%, 80%–100%]	[0.55, 0.30, 0.10, 0.03, 0.02] **	[0.2, 0.4, 0.6, 0.8, 1.0] *	5
	5.7	What is the average socioeconomic level of the people involved in the project? (This refers to the collective economic and social standing of the project's personnel, measured by income, education, and occupation. It is essential for assessing potential disparities in cybersecurity awareness and practices. Individuals with higher socioeconomic statuses often exhibit better cybersecurity attitudes and behaviors, influencing the overall project's cyber risk.)	[Level 1, Level 2, Level 3, Level 4, Level 5]	[0.10, 0.20, 0.40, 0.20, 0.10] **	[1.0, 0.8, 0.6, 0.4, 0.2] *	5

Notes.
* in the P_o column indicates the probability distribution is informed by the published database.
** in the P_o column indicates the probability distribution is based on estimation and expert validation.
* in the R column indicates the risk degree is derived based on linear assumption.
** in the R column indicates the risk degree is derived based on fuzzy set theory.
For detailed information about each risk factor, please refer to (Yao et al., 2023).

Data availability

Data will be made available on request.

References

Aghaei, P., Asadollahfardi, G., Katabi, A., 2022. Safety risk assessment in shopping center construction projects using fuzzy Fault Tree analysis method. Qual. Quantity 56 (1). https://doi.org/10.1007/s11135-021-01115-9.

Anysz, H., Apollo, M., Grzyl, B., 2021. Quantitative risk assessment in construction disputes based on machine learning tools. Symmetry 13 (5), 744. https://doi.org/10.3390/sym13050744.

Barbaschow, A., 2023. Bouygues Construction Falls Victim to Ransomware. ZDNET [Online]. Available: https://www.zdnet.com/article/bouygues-construction-falls-victim-to-ransomware/. (Accessed 30 September 2023).

Bello, A., Maurushat, A., 2020. Technical and behavioural training and awareness solutions for mitigating ransomware attacks. In: Applied Informatics and Cybernetics in Intelligent Systems. Springer International Publishing, Cham, pp. 164–176 [Online]. Available: https://link.springer.com/10.1007/978-3-030-51974-2_14. (Accessed 27 May 2024).

Catharine, Tunney, 2023. Ransomware attack on construction company raises questions about federal contracts. CBC News [Online]. Available: https://www.cbc.ca/news/politics/ransomware-bird-construction-military-1.5434308. (Accessed 30 September 2023).

Chris, Pash, 2023. How hackers and spies tried to steal the secrets of Australia's one-armed robot bricklayer. Yahoo Finance [Online]. Available: https://au.finance.yahoo.com/news/hackers-spies-tried-steal-secrets-103645052.html?guccounter=1. (Accessed 30 September 2023).

Chung, K., Kamhoua, C.A., Kwiat, K.A., Kalbarczyk, Z.T., Iyer, R.K., 2016. Game theory with learning for cyber security monitoring. In: 2016 IEEE 17th International Symposium on High Assurance Systems Engineering (HASE). IEEE, Orlando, pp. 1–8. https://doi.org/10.1109/HASE.

Deloitte. Building cybersecurity in the construction industry [Online]. Available: https://www2.deloitte.com/ce/en/pages/real-estate/articles/ce-building-cybersecurity-in-the-construction-industry.html. (Accessed 30 September 2023).

Edmonds, J., 1971. Matroids and the greedy algorithm. Math. Program. 1 (1). https://doi.org/10.1007/BF01584082.

El-Sayegh, S., Romdhane, L., Manjikian, S., 2020. A critical review of 3D printing in construction: benefits, challenges, and risks. Arch. Civ. Mech. Eng. 20 (2). https://doi.org/10.1007/s43452-020-00038-w.

Fan, C.-L., 2020. Defect risk assessment using a hybrid machine learning method. J. Construct. Eng. Manag. 146 (9), 04020102. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001897.

Fitzsimmons, J.P., Lu, R., Hong, Y., Brilakis, I., 2022. Construction schedule risk analysis – a hybrid machine learning approach. ITcon 27, 70–93. https://doi.org/10.36680/j.itcon.2022.004.

Galanis, P., 2018. The Delphi method. Arch. Hellenic Med. 35 (4). https://doi.org/10.4324/9781315728513-10.

García de Soto, B., Turk, Ž., Maciel, A., Mantha, B., Georgescu, A., Sonkor, M.S., 2022. Understanding the significance of cybersecurity in the construction industry: survey findings. J. Construct. Eng. Manag. 148 (9). https://doi.org/10.1061/(ASCE)CO.1943-7862.0002344.

George, M.R., Nalluri, M.R., Anand, K.B., 2022. Application of ensemble machine learning for construction safety risk assessment. J. Inst. Eng. India Ser. A 103 (4), 989–1003. https://doi.org/10.1007/s40030-022-00690-w.

Goh, G.D., Sing, S.L., Yeong, W.Y., 2021. A review on machine learning in 3D printing: applications, potential, and challenges. Artif. Intell. Rev. 54 (1), 63–94. https://doi.org/10.1007/s10462-020-09876-9.

Gondia, A., Siam, A., El-Dakhakhni, W., Nassar, A.H., 2020. Machine learning algorithms for construction projects delay risk prediction. J. Construct. Eng. Manag. 146 (1), 04019085. https://doi.org/10.1061/(ASCE)CO.1943-7862.0001736.

Gondia, A., Ezzeldin, M., El-Dakhakhni, W., 2022. Machine learning–based decision support framework for construction injury severity prediction and risk mitigation. ASCE-ASME J. Risk Uncertainty Eng. Syst., Part A: Civ. Eng. 8 (3), 04022024. https://doi.org/10.1061/AJRUA6.0001239.

Goodfellow, I., Bengio, Y., Courville, A., 2016. Deep Learning. MIT press.

Hoerl, A.E., Kennard, R.W., 1970. Ridge regression: biased estimation for nonorthogonal problems. Technometrics 12 (1). https://doi.org/10.1080/00401706.1970.10488634.

Huang, S., 2022. Linear regression analysis. In: International Encyclopedia of Education, fourth ed. https://doi.org/10.1016/B978-0-12-818630-5.10067-3

InfoCenter, E.N.R., 2024. Why YOUR construction company needs a good cybersecurity strategy. Eng. News Rec. [Online]. Available: https://www.enr.com/articles/56888-why-your-construction-company-needs-a-good-cybersecurity-strategy.

Jacobsen, B.N., 2023. Machine learning and the politics of synthetic data. Big Data & Society 10 (1), 205395172211453. https://doi.org/10.1177/20539517221145372.

Jiao, J., 2018. Discussion on the neural network model of comprehensive evaluation of computer network security. Inf. Commun. (10), 14–15.

Lee, W.S., Grosh, D.L., Tillman, F.A., Lie, C.H., 1985. Fault tree analysis, methods, and applications ☆ A review. IEEE Trans. Reliab. R-34 (3), 194–203. https://doi.org/10.1109/TR.1985.5222114.

Li, Q., et al., 2019. Safety risk monitoring of cyber-physical power systems based on ensemble learning algorithm. IEEE Access 7, 24788–24805. https://doi.org/10.1109/ACCESS.2019.2896129.

Liu, H., Tian, G., 2019. Building engineering safety risk assessment and early warning mechanism construction based on distributed machine learning algorithm. Saf. Sci. 120, 764–771. https://doi.org/10.1016/j.ssci.2019.08.022.

Lundberg, S., Lee, S.-I., 2017. A unified approach to interpreting model predictions. arXiv: arXiv:1705.07874. http://arxiv.org/abs/1705.07874. (Accessed 12 April 2024).

Mantha, B.R.K., García de Soto, B., 2019. Cyber security challenges and vulnerability assessment in the construction industry. In: Proceedings of the Creative Construction Conference 2019. Budapest University of Technology and Economics, pp. 29–37. https://doi.org/10.3311/CCC2019-005.

Mantha, B.R.K., García de Soto, B., 2021. Assessment of the cybersecurity vulnerability of construction networks. Eng. Construct. Architect. Manag. 28 (10), 3078–3105. https://doi.org/10.1108/ECAM-06-2020-0400.

Mantha, B.R.K., Jung, Y., García de Soto, B., 2020. Implementation of the common vulnerability scoring system to assess the cyber vulnerability in construction projects. In: Creative Construction E-Conference 2020. Budapest University of Technology and Economics, Budapest, Hungary, pp. 117–124. https://doi.org/10.3311/ccc2020-030.

Mantha, B., García de Soto, B., Karri, R., 2021. Cyber security threat modeling in the AEC industry: an example for the commissioning of the built environment. Sustain. Cities Soc. 66, 102682. https://doi.org/10.1016/j.scs.2020.102682.

Mantha, B.R.K., Sonkor, M.S., García de Soto, B., 2024. Investigation of the cyber vulnerabilities of construction networks using an agent-based model. Develop. Built Environ. 18, 100452. https://doi.org/10.1016/j.dibe.2024.100452.

Matsika, E., Robinson, M., Neill, C.O., 2016. Risk assessment tool for analysing terrorist attack impact on metro and light rail systems. In: *European Transport Conference 2016Association for European Transport (AET)*, Barcelona: European Transport Conference 2016, pp. 1–22.

McCulloch, W.S., Pitts, W., 1943. A logical calculus of the ideas immanent in nervous activity. Bull. Math. Biophys. 5 (4), 115–133. https://doi.org/10.1007/BF02478259.

Mohamed Shibly, M.U.R., García de Soto, B., 2020. Threat modeling in construction: an example of a 3D concrete printing system. In: Presented at the 37th International Symposium on Automation and Robotics in Construction. Kitakyushu, Japan. https://doi.org/10.22260/ISARC2020/0087.

Molnar, C., 2024. Interpretable Machine Learning. GitHub Repository [Online]. Available: https://christophm.github.io/interpretable-ml-book/. (Accessed 22 January 2024).

P, R.K.V., Mallidi, S.K.R., J. K, S., L. D, P., 2021. Bat optimization algorithm for wrapper-based feature selection and performance improvement of android malware detection. IET Netw. 10 (3), 131–140. https://doi.org/10.1049/ntw2.12022.

Pan, Z., Hariri, S., Pacheco, J., 2019. Context aware intrusion detection for building automation systems. Comput. Secur. 85, 181–201. https://doi.org/10.1016/j.cose.2019.04.011.

Pang, Y., Xue, X., Namin, A.S., 2016. Early identification of vulnerable software components via ensemble learning. In: 2016 15th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, pp. 476–481. https://doi.org/10.1109/ICMLA.2016.83.

Parn, E.A., Edwards, D., 2019. Cyber threats confronting the digital built environment: common data environment vulnerabilities and block chain deterrence. Eng. Construct. Architect. Manag. 26 (2). https://doi.org/10.1108/ECAM-03-2018-0101.

Parzen, E., 1962. On estimation of a probability density function and mode. Ann. Math. Stat. 33 (3), 1065–1076. https://doi.org/10.1214/aoms/1177704472.

Poh, C.Q.X., Ubeynarayana, C.U., Goh, Y.M., 2018. Safety leading indicators for construction sites: a machine learning approach. Autom. ConStruct. 93, 375–386. https://doi.org/10.1016/j.autcon.2018.03.022.

Rajbahadur, G.K., Wang, S., Oliva, G.A., Kamei, Y., Hassan, A.E., 2022. The impact of feature importance methods on the interpretation of defect classifiers. IEEE Trans. Software Eng. 48 (7). https://doi.org/10.1109/TSE.2021.3056941.

Raliile, M.T., Haupt, T.C., 2020. Machine learning applications for monitoring construction health and safety legislation and compliance. Proc. Int. Struct. Eng. Construct. 7 (2). https://doi.org/10.14455/ISEC.2020. CON-23.

RiskAmp, 2012. What Is Monte Carlo Simulation? RiskAmp.

Russell, R., et al., 2018. Automated vulnerability detection in source code using deep representation learning. In: 2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA). IEEE, pp. 757–762. https://doi.org/10.1109/ICMLA.2018.00120.

Salami Pargoo, N., Ilbeigi, M., 2023a. A scoping review for cybersecurity in the construction industry. J. Manag. Eng. 39 (2). https://doi.org/10.1061/JMENEA.MEENG-5034.

Salami Pargoo, N., Ilbeigi, M., 2023b. A scoping review for cybersecurity in the construction industry. J. Manag. Eng. 39 (2), 03122003. https://doi.org/10.1061/JMENEA.MEENG-5034.

Sanni-Anibire, M.O., Zin, R.M., Olatunji, S.O., 2022. Machine learning model for delay risk assessment in tall building projects. Int. J. Construct. Manag. 22 (11), 2134–2143. https://doi.org/10.1080/15623599.2020.1768326.

Sawyer, T., Rubenstone, J., 2021. Construction cybercrime is on the rise. Eng. News Rec. [Online]. Available: https://www.enr.com/articles/46832-construction-cybercrime-is-on-the-rise. (Accessed 23 April 2021).

Senol, Y.E., Aydogdu, Y.V., Sahin, B., Kilic, I., 2015. Fault tree analysis of chemical cargo contamination by using fuzzy approach. Expert Syst. Appl. 42 (12), 5232–5244. https://doi.org/10.1016/j.eswa.2015.02.027.

Sheikh, A., Kamuni, V., Patil, A., Wagh, S., Singh, N., 2019. Cyber attack and fault identification of HVAC system in building management systems. In: 2019 9th International Conference on Power and Energy Systems (ICPES). IEEE, pp. 1–6. https://doi.org/10.1109/ICPES47639.2019.9105438.

Shemov, G., García de Soto, B., Alkhzaimi, H., 2020. Blockchain applied to the construction supply chain: a case study with threat model. Front. Eng. Manag. 7 (4), 564–577. https://doi.org/10.1007/s42524-020-0129-x.

Sonkor, M.S., García de Soto, B., 2021a. Is your construction site secure? A view from the cybersecurity perspective. In: Proceedings of the 38th International Symposium on Automation and Robotics in Construction (ISARC). ISARC, pp. 864–871. https://doi.org/10.22260/isarc2021/0117.

Sonkor, M.S., García de Soto, B., 2021b. Operational technology on construction sites: a review from the cybersecurity perspective. J. Construct. Eng. Manag. 147 (12), 04021172. https://doi.org/10.1061/(ASCE)CO.1943-7862.0002193.

Sonkor, M.S., García de Soto, B., 2021c. Towards secure construction networks: a data-sharing architecture utilizing blockchain technology and decentralized storage. In: *Blockchain and the Digital Twin*, Design Computation. https://doi.org/10.47330/CBC.2021.NOKH7555.

Sonkor, M.S., García de Soto, B., 2023. Lessons learned from the 'hack my robot' competition and considerations for construction applications. In: *In Proceedings of the 40th International Symposium on Automation and Robotics in Construction (ISARC 2023)*, Chennai, India: International Association for Automation and Robotics in Construction (IAARC), pp. 577–584.

Stiles, M., 2021. Turner Construction Data Breach Exposes Hundreds in Washington to Possible Fraud. The Business Journals [Online]. Available: https://www.bizjournals.com/seattle/blog/techflash/2016/04/turner-construction-data-breach-exposes-hundreds.html. (Accessed 15 July 2021).

Thambawita, V., et al., 2022. SinGAN-seg: synthetic training data generation for medical image segmentation. PLoS One 17 (5), e0267976. https://doi.org/10.1371/journal.pone.0267976.

Thibault, M., 2024. Skender Hit by Ransomware Attack. ConstructionDive [Online]. Available: https://www.constructiondive.com/news/skender-ransomware-attack-chicago-maine/712844/. (Accessed 12 May 2024).

Tibshirani, R., 1996. Regression shrinkage and selection via the Lasso. J. Roy. Stat. Soc. B 58 (1). https://doi.org/10.1111/j.2517-6161.1996.tb02080.x.

Turk, Ž., García de Soto, B., Mantha, B.R.K., Maciel, A., Georgescu, A., 2022. A systemic framework for addressing cybersecurity in construction. Autom. ConStruct. 133, 103988. https://doi.org/10.1016/j.autcon.2021.103988.

Wang, Y., Su, C., Xie, M., 2022. A weakest T-Norm based fuzzy Fault Tree approach for probability assessment of natural gas pipeline. In: 2022 6th International Conference on System Reliability and Safety, ICSRS 2022. https://doi.org/10.1109/ICSRS56243.2022.10067575.

Wolf, V., Lugmayr, A., Danelljan, M., Van Gool, L., Timofte, R., 2021. DeFlow: learning complex image degradations from unpaired data with conditional flows. In: 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). IEEE, pp. 94–103. https://doi.org/10.1109/CVPR46437.2021.00016.

Yao, D., García de Soto, B., 2022. A preliminary SWOT evaluation for the applications of ML to cyber risk analysis in the construction industry. IOP Conf. Ser. Mater. Sci. Eng. 1218 (1), 012017. https://doi.org/10.1088/1757-899X/1218/1/012017.

Yao, D., García de Soto, B., 2023. A corpus database for cybersecurity topic modeling in the construction industry. In: Presented at the 40th International Symposium on Automation and Robotics in Construction, Chennai, India. https://doi.org/10.22260/ISARC2023/0072.

Yao, D., García De Soto, B., 2024a. Cyber risk assessment framework for the construction industry using machine learning techniques. Buildings 14 (6), 1561. https://doi.org/10.3390/buildings14061561.

Yao, D., García De Soto, B., 2024b. Enhancing cyber risk identification in the construction industry using language models. Autom. ConStruct. 165, 105565. https://doi.org/10.1016/j.autcon.2024.105565.

Yao, D., García de Soto, B., Wilkes, M., 2023. Identifying cyber risk factors associated with construction projects. SSRN J. https://doi.org/10.2139/ssrn.4648243.

Zadeh, L.A., 1965. Fuzzy sets. Inf. Control 8 (3). https://doi.org/10.1016/S0019-9958(65)90241-X.