buildings

MDPI

Article

Cyber Risk Assessment Framework for the Construction Industry Using Machine Learning Techniques

Dongchi Yao [1,2,*] and Borja García de Soto [1,2]

[1] S.M.A.R.T. Construction Research Group, Division of Engineering, New York University Abu Dhabi (NYUAD), Experimental Research Building, Saadiyat Island, Abu Dhabi P.O. Box 129188, United Arab Emirates; garcia.de.soto@nyu.edu

[2] Department of Civil and Urban Engineering, Tandon School of Engineering, New York University (NYU), New York, NY 11201, USA

* Correspondence: dongchi.yao@nyu.edu

Abstract: Construction 4.0 integrates digital technologies that increase vulnerability to cyber threats. A dedicated cyber risk assessment framework is essential for proactive risk mitigation. However, existing studies on this subject within the construction sector are scarce, with most discussions still in the preliminary stages. This study introduces a cyber risk assessment framework that integrates machine learning techniques, pioneering a data-driven approach to quantitatively assess cyber risks while considering industry-specific vulnerabilities. The framework builds on over 20 literature reviews related to construction cybersecurity and semi-structured interviews with two industry experts, ensuring both rigor and alignment with practical industrial needs. This study also addresses the challenges of data collection and proposes potential solutions, such as a standardized data collection format with preset fields that computers can automatically populate using data from construction companies. Additionally, the framework proposes dynamic machine learning models that adjust based on new data, facilitating continuous risk monitoring tailored to industry needs. Furthermore, this study explores the potential of advanced language models in cybersecurity management, positioning them as intelligent cybersecurity consultants that provide answers to security inquiries. Overall, this study develops a conceptual machine learning framework aimed at creating a robust, off-the-shelf cyber risk management system for industry practitioners.

Keywords: cybersecurity; automation; cyber-physical systems; digital twins; machine learning

check for updates

Citation: Yao, D.; García de Soto, B. Cyber Risk Assessment Framework for the Construction Industry Using Machine Learning Techniques. *Buildings* **2024**, *14*, 1561. https://doi.org/10.3390/buildings14061561

Academic Editor: Jurgita Antucheviciene

Received: 31 March 2024
Revised: 20 May 2024
Accepted: 24 May 2024
Published: 28 May 2024

1. Introduction

The global construction industry is transitioning towards Construction 4.0, which integrates digital technology, automation, and cyber-physical systems into building processes. This enhances efficiency, sustainability, and safety, aiming to construct, operate, and maintain assets faster, cheaper, and with higher quality through digitalized technologies [1]. In Construction 4.0, the cyber-physical system is the core component, involving the interaction between cyberspace and physical space through data. The cyber-physical system consists of three layers: physical, connecting, and digital. The physical layer captures on-site data using sensors and drones. The connecting layer, which includes technologies like BIM (building information modeling) and common data environment (CDE), not only stores but also processes and manages this data. Finally, the digital layer analyzes the data using advanced technologies such as cloud computing, big data, digital twins, and virtual reality (VR) [1]. For instance, in a high-rise construction project, sensors (physical layer) monitor concrete curing and send this data to the BIM systems (connecting layer), which dynamically adjust project timelines. Simultaneously, digital twins (digital layer) simulate building stress scenarios, aiding decision-making for enhanced safety and efficiency.

The transition towards Construction 4.0 marks a revolutionary shift from traditional practices, integrating real-time data, automation, and interconnected digital platforms into the construction process. This digital transformation introduces unique cybersecurity challenges, underscoring the urgent need to protect these innovative systems against increasingly sophisticated threats [2]. Despite the vital functions these technologies perform in construction projects, the industry remains highly vulnerable to cyberattacks, which can lead to economic losses, physical property damage, and even injuries due to the malfunction of operational equipment. Surprisingly, cybersecurity is not prioritized as a high business priority within the industry, and the construction industry has been ranked second as the target of cyberattacks, as indicated by a recent survey by the Department for Digital, Culture, Media, and Sport [3]. Cybersecurity incidents in construction have repeatedly occurred globally in recent years. These incidents include Turner Construction falling victim to a spear-phishing scam [4]. The information provided to the fraudulent email account included full names, Social Security numbers, states of employment and residence, as well as tax withholding data for 2015. All employees who worked for the company in 2015 were affected by the data breach. Jewson discovered an abnormal piece of code compromising personal data [5]. During this period, up to 2000 customers using the Jewson Direct online store may have been affected. Marous Brothers Construction did not receive a payment of USD 1.7 million due to malicious modification of routing numbers [6]. Bird Construction was breached by ransomware. Hacker organization MAZE claims to have stolen 60 GB of data from the company, which signed 48 contracts worth CAD 406 million with the Canadian Department of Defense between 2006 and 2015 [7]. Hoffmann Construction reported unauthorized access to employee information. This information includes employee name, address, date of birth, Social Security number, and welfare information [8].

One of the primary reasons for recurring cyber incidents in the construction industry is the sector's general lack of awareness and a profound skills gap in handling cybersecurity. This leads to the slow adoption of existing cybersecurity measures and the absence of cyber risk assessment frameworks specifically designed for construction (as stated in Section 2). The lack of such frameworks further hinders effective cybersecurity management for construction projects. Therefore, these challenges and gaps highlight the critical need for research that can develop and implement a construction-specific cyber risk assessment framework to identify and assess potential cyber risks, thus facilitating risk decision-making and proactively preventing the losses that may be incurred in a construction project.

In the context of cybersecurity, risk assessment is a systematic process that involves identifying various cyber risks that could affect assets in construction projects. It also includes evaluating the likelihood of such events occurring and determining the potential consequences of such attacks on these assets [9]. By understanding the risks, construction companies can implement targeted strategies to mitigate them, ensuring the security and resilience of their operations in the face of increasing digital threats. This process not only helps in prioritizing risks or risk factors that contribute to these risks based on their potential impact, but also guides the allocation of resources to areas where they are most needed for effective cybersecurity management [10]. Traditional risk assessment in construction, typically conducted through qualitative methods like fault tree analysis and fuzzy comprehensive assessment to assess the threat likelihood and vulnerability impact, etc., often relies heavily on human expertise, resulting in subjective outcomes and prolonged discussions. Given the rapid evolution of cybersecurity threats, there is a pressing need for more objective and efficient risk assessment techniques tailored to the construction industry's unique needs. Machine learning (ML) offers substantial advantages by automating data analysis, thus speeding up the risk assessment process. ML offers substantial advantages by automating data analysis, thus speeding up the risk identification and assessment process. ML not only enhances prediction accuracy through advanced statistical techniques but also manages the industry's complex datasets more effectively, extracting insights from diverse data sources such as project timelines, contracts, and work logs—tasks that traditional methods may struggle with over short periods. This enables project managers to make more informed decisions [11].

This study proposes a cyber risk assessment framework that utilizes ML techniques, which can facilitate risk decision-making and help project managers proactively prevent losses. This framework is specifically tailored to address vulnerabilities unique to the construction industry. The objectives are twofold: (1) To propose and detail the six modules of the cyber risk assessment framework, explaining how ML is integrated and applied; (2) To discuss some critical aspects of creating a model that is aligned with industry needs. This study contributes both academically and practically to the field of construction cybersecurity. Academically, it introduces a framework tailored to the construction industry, addressing the gap of the absence of an objective and efficient cyber risk assessment framework in current research and enriching the intersection of construction, cybersecurity, and machine learning. Practically, it enhances industry awareness of specific cyber vulnerabilities and sets the stage for providing a flexible, user-friendly ML tool, improving the cybersecurity infrastructure of construction projects and facilitating proactive risk management.

2. Related Works

Many studies, tools, and standards regarding this topic have been proposed, but mostly in the information technology industry. For example, the frameworks proposed by the National Institute of Standards and Technology (NIST) [12] and by the International Organization for Standardization's Information Security Management Systems (series of 27000s code) [13] are the most widely employed. Many sectors have adopted and tailored these tools, practices, and standards to fit into their own domains, like the Center for Internet Security framework for the healthcare sector [14], regulations proposed by the New York Department of Financial Services (NYDFS) for the finance sector [15], etc., suggesting that industry-unique studies are necessary for efficient cyber risk management [16]. Nevertheless, given the dynamic nature of construction projects, the construction industry faces distinct cybersecurity vulnerabilities and challenges (detailed in Section 3.1.3). General frameworks and methodologies designed primarily for sectors like manufacturing might not sufficiently address these unique cybersecurity requirements. Instead, tailored and industry-specific approaches are necessary.

However, only a few studies have been developed for the construction industry within the last few years. The purpose of these studies can be classified into general discussions, review papers and specific solutions [17]. General discussions on cybersecurity topics in construction include works by Bello and Maurushat [18], EI-Sayegh et al. [19], Mantha and García de Soto [2], Yao and García de Soto [11], Turk et al. [20], among others. Review papers include works such as those by Pargoo and Ilbeigi [17], Pärn and Edwards [21], and Goh et al. [22]. Specific methods or solutions cover blockchain technology [21,23], machine learning or deep learning algorithms [24,25], threat modeling [26,27], a framework proposal [20], and the Common Vulnerability Scoring System (CVSS) [28]. Some of the most recent related works are as follows: Mantha and García de Soto [28] used the CVSS to quantify the cyber vulnerability of different participants in construction projects. However, it is still manually implemented, and the assignment of the CVSS score for each participant is subjective and based on assumptions instead of real data. Shibly and García de Soto [27] developed a threat modeling method that adopted the Quantitative TMM method based on the STRIDE framework and applied the proposed framework to a 3D concrete printing system. However, the implementation of this framework involves assumed CVSS scores that propagate through an attack tree, which still involves great subjectivity. Mantha et al. [26] proposed a cybersecurity threat model tailored to the construction industry and placed it in the commissioning phase for a case study. However, this threat model only involves qualitatively identifying the threats and vulnerabilities across the project phases instead of providing a framework for quantifying the risks.

In 2023, Pargoo and Ilbeigi [17] conducted a scoping review on cybersecurity in the construction industry, again highlighting the limited research in this field. They found only 19 studies that provided specific technical solutions, of which four emphasized the importance and prospect of predictive models like ML but without implementations. Other methods, such as threat modeling and fault tree analysis, demand significant human input and are time-consuming. These techniques struggle to keep pace with rapidly evolving anomalies that are hard to detect, highlighting their misalignment with the construction sector's digitalization objectives. Additionally, even among those mentioning ML, they only identified it as a potential solution without providing any detailed framework or focused solely on general IT issues, failing to explore the unique vulnerabilities of the industry. The limitations of current methods, including their dependency on extensive human intervention, their inability to scale, and their inefficacy in timely anomaly detection, prove there is a pressing need to develop a machine learning framework specifically for cyber risk assessment in the construction industry.

In summary, research on cyber risk assessment within the construction industry remains in its early stages. Many of the previously mentioned methods and techniques demand significant human involvement and are inflexible, leading to inefficiencies and an inability to accurately predict rapidly evolving cyber risks. Notably, there is a clear lack of the practical application of ML, a powerful tool adept at swiftly managing the vast data sources prevalent in the modern construction sector and capturing interactive vulnerabilities. Given the gap, there is a pressing need to incorporate ML approaches for a more efficient and tailored cyber risk assessment.

3. The Cyber Risk Assessment Framework

In this section, an ML-integrated framework specifically designed for cyber risk assessment in the construction industry is introduced, which is shown in Figure 1. This framework consists of six modules: identifying cyber risks that might be encountered by assets, defining risk assessment objectives, designing features for ML models, collecting data, developing ML models, and prioritizing risk factors that aid in effective risk mitigation. The ML technique, mainly the supervised ML technique in this study, provides a model for quantifying the risks by analyzing the collected data in construction projects or within construction companies. It also provides a pathway for analyzing which factors are more important so that these factors can be addressed in prioritization. During the formation of the framework, an established ML process to assess delay risks in construction projects [29] is referred to, which is comprehensive in its ML application and aligns well with the conventional risk assessment framework. Simultaneously, interviews were conducted with experienced industry experts for refinement. Two specialists from a major construction company in the UAE were consulted.

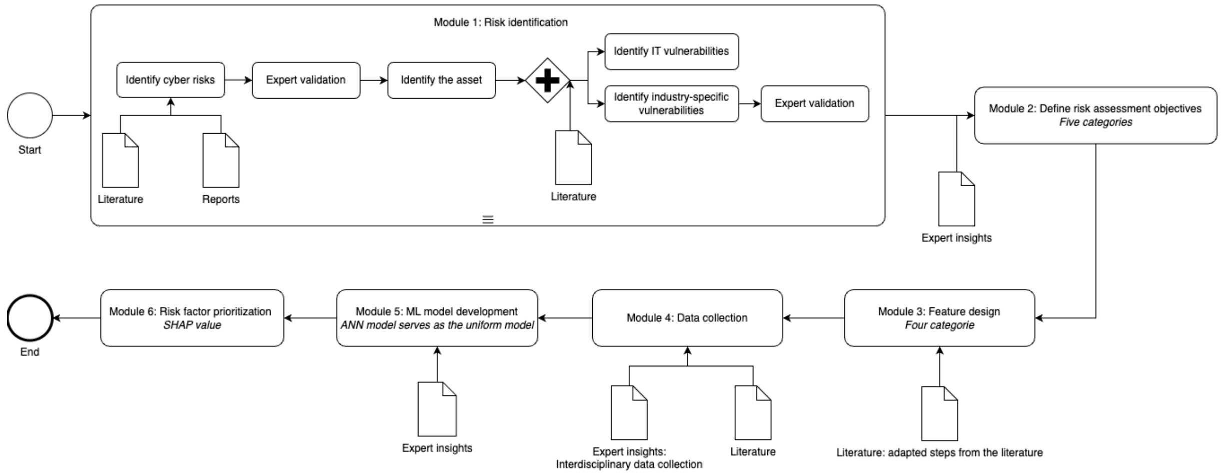
Module 1: Risk identification
Start
Identify cyber risks
Expert validation
Identify the asset
Identify IT vulnerabilities
Identify industry-specific vulnerabilities
Expert validation
Literature
Reports
Literature
Module 2: Define risk assessment objectives
Five categories
Expert insights
Module 3: Feature design
Four categorie
Module 4: Data collection
Module 5: ML model development
ANN model serves as the uniform model
Module 6: Risk factor prioritization
SHAP value
End
Expert insights
Expert insights:
Interdisciplinary data collection
Literature
Literature: adapted steps from the literature

Figure 1. The proposed framework.

The first expert is an IT specialist with more than 10 years of experience (Expert 1); The second is an expert in the construction industry with over 12 years of experience (Expert 2). This combination of expertise is crucial to ensure that our framework is not only relevant but also practical and comprehensive, integrating industry-specific knowledge with cybersecurity considerations. The interviews were semi-structured, lasting approximately 2.5 h in total, and involved both closed-ended questions and open-ended discussions. Each expert answered questions in turn, sharing their perspectives. Following each question, open discussions were initiated. The theoretical foundation of the previous work [29] and expert consultation collectively ensure that the framework developed is both rigorous and aligned with industrial practicality.

3.1. Risk Identification

Risk identification is the initial step in cyber risk assessment [29,30]. It is the process of determining which cyber risks potentially threaten an organization's assets by exploiting vulnerabilities. In this context, three key terms should be noted: "cyber risk", "asset", and "vulnerability".

3.1.1. Identifying the Cyber Risks

As noted in Section 1, the construction industry has experienced significant cyber incidents. By looking at the literature that focuses on the general discussions of cybersecurity in construction [2,10,16,28], coupled with official cyber incident analysis reports from two renowned organizations including Deloitte [31], which is a global professional services network known for its expertise in consulting, financial advisory, risk management and Engineering News-Record [32], a magazine and website that provides news, analysis, data, and opinion for the construction industry, five cyber risks can be identified as the risk types that are most interesting to stakeholders.

(1) Ransomware. These attacks encrypt critical project data and demand a ransom for its decryption. In construction, project managers may find crucial documents like blueprints and financial records encrypted, leading to project delays, financial losses from ransom payments, operational shutdowns, and damage to client relationships. Victims include Bouygues Construction [33] and Skender Construction [34].

(2) Phishing. Phishing schemes deceive individuals into disclosing sensitive information, often through deceptive emails. For construction professionals, phishing could involve fraudulent emails from seemingly legitimate sources, leading to unauthorized access to financial accounts or confidential data, resulting in financial loss and compromised project security. Victims include Marous Brothers Construction [6] and Turner Construction [35].

(3) Insider Attacks. These occur when individuals within an organization—typically trusted employees—act maliciously. In construction, this might involve the theft or sabotage of key materials or the leaking of proprietary information, leading to unforeseen expenses, project delays, and legal issues related to contract or intellectual property breaches. Victims include Target Stores and AECOM [36].

(4) Data Breaches. Construction projects store significant amounts of digital data, and unauthorized access to this information, whether by external hackers or insiders, can lead to immediate financial loss, legal consequences, and long-term reputational damage that may affect future business opportunities. Victims include the Ingérop firm [37] and Bird Construction [38], which suffered from data breaches as a result of attacks.

(5) Supply Chain Attacks. The construction supply chain is complex, and disruptions within it can have severe implications. Issues such as supplier insolvency and transportation problems can lead to project delays and increased costs as alternative sourcing solutions are sought. Additionally, challenges like substandard material quality may force the use of inferior materials, compromising the overall project quality. These consequences of supply chain vulnerabilities are repeatedly emphasized by MarshMcLennan [39].

To confirm the suitability and importance of the five identified cyber risks, experts were asked during the interview to rank them based on their perceived level of concern for the construction industry. The ranking range was from 1 to 5, with 1 representing the most concerning risk. The results have been analyzed and are organized in Table 1.

Table 1. Expert ranking and comment on the identified cyber risks.

Cyber Risk	Expert 1 Ranking	Expert 1 Comment	Expert 2 Ranking	Expert 2 Comment
Ransomware	4	"Important, but not as prevalent as other risks in our experience".	5	"Seen infrequently, but the impact can be significant when it occurs".
Phishing	1	"Most common and dangerous, especially due to employee vulnerability".	1	"A constant threat, often overlooked in our industry".
Insider Attacks	2	"It's a growing concern with the high turnover of staff".	2	"Hard to detect and can have devastating consequences".
Data Breaches	3	"Increasingly common with the digitization of our data".	4	"Significant but not as immediate a threat as phishing or insider attacks".
Supply Chain Attacks	5	"A risk, but more due to external factors than cyber threats per se".	3	"Particularly relevant given recent global events affecting supply chains".

There was a consensus that phishing is the most concerning risk, attributed primarily to its high likelihood of occurrence, particularly among untrained employees in small-to-mid-sized construction companies. Insider attacks were ranked as the second most significant threat, a concern heightened by the nature of the construction industry, where numerous individuals are constantly exchanging project data. However, the experts' perspectives diverged when considering other risks, a divergence that may be attributed to their differing areas of expertise. Expert 1 emphasized data breaches as more prevalent, especially from an IT perspective, thereby ranking them higher in the risk hierarchy. In contrast, Expert 2 identified supply chain attacks as the third most significant threat, noting the specific vulnerabilities within the construction industry. Ransomware was also acknowledged but was generally ranked lower compared to the aforementioned risks. When asked whether the five identified cyber risks covered most of the potential threats in their construction projects, both experts agreed that these risks were comprehensive and thorough. Therefore, any future risk assessment efforts should focus on these five cyber risks, with particular attention to phishing and insider attacks.

3.1.2. Identifying the Asset

In standard cyber risk assessment, the next step involves identifying assets that could be cyberattack targets. The International Telecommunication Union (ITU) [40] and the NIST [41] broadly define "assets" to include computing devices, personnel, infrastructure, applications, services, telecommunications systems, and all information within the cyber environment. This indicates that in the construction industry, both tangible and intangible elements are susceptible to cyber risks. Tangible assets refer to physical and material properties that have value and contribute to the operational capabilities of a business [42], which include machinery, vehicles, tools, construction equipment, real estate, and inventory. These physical items play crucial roles in building infrastructures. Conversely, intangible assets are non-physical assets that have value due to their attributes or rights [42], which include reputation, processes, company brand, intellectual property (patents, copyrights, and trademarks), software, skilled labor knowledge, business networks, organizational culture, licenses, permits, and client contracts. Such assets, while non-physical, significantly influence the industry's growth, credibility, and competitive advantage. There is an interplay between tangible and intangible assets; for example, cyberattacks on intangible assets, such as data breaches affecting business networks or software, can cause tangible losses by disrupting operations, delaying construction projects, and incurring significant financial costs to rectify the damage.

In the risk assessment process, prioritizing assets involves a detailed evaluation of their importance to project completion, susceptibility to cyber threats, and intrinsic value to the organization. Methods for prioritization typically include both quantitative and qualitative assessments. Quantitative evaluations often use scoring systems that calculate the potential impact of an asset being compromised, considering factors like cost of replacement, operational downtime, and impact on revenue. Qualitative assessments, on the other hand, delve into the severity and likelihood of disruptions or data breaches, examining aspects such as the strategic importance of the assets and the potential reputational damage from security failures. These combined approaches ensure a comprehensive understanding of asset criticality and guide effective cybersecurity measures.

3.1.3. Identifying Vulnerabilities of the Asset

The next step involves identifying the vulnerabilities of the asset, which can guide us in future feature design and data collection to accurately reflect these vulnerabilities. These vulnerabilities can be technical in nature, such as flaws in software or hardware, commonly associated with IT systems. They can also stem from aspects unique to the construction industry, related to operational processes and management, such as inadequate security training, poor process oversight, and the specific nature of construction projects [43]. This section delves into two aspects of vulnerabilities that can serve as a starting point when considering the vulnerabilities of a specific asset.

(1) The IT vulnerabilities (V_1)

The construction industry, like many other sectors, is vulnerable to IT threats due to its reliance on digital technologies and networked systems in project management, design software, and operational processes [43]. Different standards, such as NIST [11] and PAS 1192 [44], have been explored to identify IT vulnerabilities relevant to the construction industry. Our analysis categorized these vulnerabilities into four main categories, detailed in Table 2. These categories inform the feature design and data collection processes outlined in Sections 3.3 and 3.4. To ensure comprehensive identification of IT vulnerabilities for a specific asset, it is essential to collaborate with both project managers and IT personnel. This collective approach can utilize tools for vulnerability scanning or cybersecurity audits, such as Nessus v10.7.3, Qualys v10.27, OpenVAS v23.0.0, etc.

Table 2. Common IT vulnerabilities (V_1).

Aspect	Description	Examples of Vulnerabilities
Software Flaws ($V_{1\text{-}1}$)	Issues or bugs in software that could be exploited.	Buffer overflows, SQL injection, cross-site scripting (XSS), unvalidated inputs, directory traversal, insecure deserialization
Network Configurations ($V_{1\text{-}2}$)	Improper setup or outdated components in networks.	Open ports, misconfigured firewalls, default credentials in use, unpatched services, excessive permissions, insecure protocols
Communication Protocol Weaknesses ($V_{1\text{-}3}$)	Vulnerabilities in the way devices communicate.	Man-in-the-middle attacks, session hijacking, replay attacks, unencrypted communications, inadequate key management, insecure handshake procedures
Hardware Susceptibilities ($V_{1\text{-}4}$)	Physical vulnerabilities in devices or systems.	Firmware vulnerabilities, insecure interfaces, physical tampering risks, side-channel attacks, inadequate hardware encryption, component wear-out

(2) Industry-specific vulnerabilities (V_2)

In addition to general IT vulnerabilities, the construction industry has unique vulnerabilities due to the complex nature of its projects. Several studies [2,16,26,28,44] that discussed these vulnerabilities were reviewed, leading to the identification of five key aspects listed in Table 3. These aspects were presented to the industry experts during the interviews for feedback. In a structured interview, Expert 2, who specializes in construction, confirmed that these aspects are potential cyber risks, each with varying severity and importance.

Frequent changes in teams ($V_{2\text{-}1}$). In construction projects, teams often change to meet the demands of various project stages and specializations. Although crucial for project execution, this fluidity disrupts consistent workflows, particularly in communication and cybersecurity. Each transition can introduce vulnerabilities, mainly due to team members' unfamiliarity with established protocols and security practices. This inconsistency in team continuity raises the risk of security breaches, as team members may not be equally informed or vigilant. Expert 2 emphasized the significant impact of team changes, noting that they introduce vulnerabilities, especially when frequently integrating new members. This process involves recruiting, onboarding, training, process alignment, and access management, each of which carries potential risks that can lead to cybersecurity breaches. Expert 2 stressed that managing team changes is not merely administrative but critical to reducing the risk of security incidents in the dynamic construction environment.

Varied levels of cybersecurity knowledge among personnel ($V_{2\text{-}2}$). The construction sector employs a diverse range of roles, from field workers to IT experts, resulting in varying levels of cybersecurity knowledge. This diversity creates potential weak links as less informed personnel might inadvertently compromise project security through errors or phishing vulnerabilities. During the interview, Expert 2 expressed significant concern about this issue. He noted that the wide spectrum of cybersecurity understanding among workers could lead to weak points in security. Uninformed or poorly trained personnel could jeopardize project security, whether through simple mistakes or by falling victim to phishing attacks. To address this risk, he strongly advocated for enhanced training programs, emphasizing the need for comprehensive and regular cybersecurity training across all organizational levels.

Scattered and frequent communications ($V_{2\text{-}3}$). Communications in construction occur across various channels and locations, increasing the potential for security breaches. With diverse stakeholders involved, risks such as misinterpretation, data leaks, or unauthorized access are heightened. During the interview, Expert 2 acknowledged these issues within the industry and stressed the importance of managing them by monitoring communication channels. He recommended that communications should only occur through authorized and secure channels like approved emails or chat systems. The expert highlighted the operational importance of ensuring all personnel consistently use these approved channels to mitigate risks associated with varying levels of cybersecurity knowledge. He emphasized that while this issue requires attention, it is manageable with the right protocols in place.

Frequent exchange of digital information ($V_{2\text{-}4}$). The digital era has increased data exchanges in construction, particularly within supply chains. Every digital interaction, from sharing design blueprints to daily updates, presents a cybersecurity risk. Given the diversity of transferred data, specialized security measures are essential. During the interview, Expert 2 recognized the inevitability and importance of frequent digital exchanges, noting that they increase the vulnerability of construction projects to cyberattacks, especially within complex supply chain interactions. He emphasized that the solution is not to reduce these exchanges but to enhance their security. The expert pointed out that risks often stem from non-standardized communication protocols, inadequate software, and insufficient knowledge among personnel. He advocated for addressing these foundational issues by improving onboarding processes and equipping employees with the necessary knowledge to ensure safe and efficient digital information exchange.

Table 3. Vulnerabilities specific to the construction industry (V_2).

Aspect	Item	Description	Example
Frequent changes in teams ($V_{2\text{-}1}$)	Inconsistent security protocols	Varied security protocol application with changing team members.	For instance, while one team may use multifactor authentication (MFA) rigorously, another might only utilize basic password protocols, creating security inconsistencies.
	Lack of trust	Newly onboarded members may not have immediate trust, restricting access or data sharing.	In a new project phase, a subcontractor might hesitate to share real-time data feeds due to trust issues, possibly leading to delayed decision-making.
	Loss of knowledge	Exiting team members may take vital security knowledge with them.	An employee who departs midway through a project might have had unique access credentials or understanding of a specific cybersecurity protocol, leaving a security knowledge gap.
	Onboarding risks	New team members might introduce risks if not vetted properly.	A new contractor may unknowingly introduce malware through an infected USB drive or device during the initial setup phase.
	Limited accountability	Frequent changes can blur accountability, making fault tracking challenging.	If a breach occurs, identifying responsibility becomes challenging when team members have been regularly rotated out. For instance, a password leak from 2 months ago might involve tracking past team members.
Varied levels of cybersecurity knowledge among personnel ($V_{2\text{-}2}$)	Disparate security practices	Varied adherence to security best practices among employees.	While senior architects might use encrypted email services, newer interns might rely on personal emails, causing potential data breaches.
	Phishing susceptibility	Increased risk of less informed members succumbing to phishing or spear-phishing attempts.	Less tech-savvy team members, such as older craftsmen, might be more susceptible to clicking malicious links in scam emails.
	Improper data handling	Inadequate data storage, sharing, or processing due to ignorance.	An engineer might accidentally save sensitive project blueprints in a publicly accessible cloud folder.
	Usage of unapproved software	Employees might use software/tools not approved, risking security.	An architect might use a non-standard design software that has not been vetted for security, introducing potential risks.
	Misconfigured security settings	Incorrect security configurations due to lack of knowledge.	A team member might disable firewall settings to expedite a software installation, leaving systems vulnerable.
Scattered and frequent communications ($V_{2\text{-}3}$)	Unsecure communication channels	Risk of data interception across multiple communication points.	Using consumer-grade messaging apps for communicating about project specifics can risk data interception.
	Data integrity issues	Potential for inconsistent data due to frequent exchanges.	A subcontractor might receive an outdated design plan over email, leading to construction flaws.
	Version control issues	Stakeholders using outdated data versions can cause operational conflicts.	Without a central data repository, two teams might work on different versions of a project blueprint, leading to inconsistencies.
	Over-reliance on single channels	Relying heavily on one communication channel can create a single point of failure.	If a primary communication software faces an outage, it can halt the entire project's communication flow.
	Miscommunication	Risk of distorted or misunderstood data in fragmented communication environments.	Key safety instructions might be misunderstood or lost in long email threads, leading to on-site hazards.
Frequent exchange of digital information ($V_{2\text{-}4}$)	Data leak risk	Potential for data breaches when using insecure channels.	A contractor might unknowingly forward a confidential project blueprint to an external stakeholder, risking intellectual property.
	Data interception risk	Possible data theft during transmission.	A hacker might exploit an unencrypted data transfer, capturing sensitive financial details.
	Unauthorized data access	Data might be accessed without proper controls.	A shared project server might not have proper access restrictions, allowing unauthorized personnel to access confidential designs.
	Excessive data replication	Frequent data exchanges can lead to multiple, unnecessary data copies, increasing breach risk.	Each project subcontractor might maintain separate copies of project blueprints, increasing the data breach surface.
	Risk from third-party applications	Utilizing third-party apps for data sharing can introduce unknown vulnerabilities.	Utilizing a less-known third-party scheduling app can introduce vulnerabilities not present in industry-standard software.

Table 3. *Cont.*

Aspect	Item	Description	Example
Personnel overlapping across multiple projects ($V_{2\text{-}5}$)	Data confusion	Risks of data misplacement or incorrect stakeholder sharing.	With simultaneous projects, a blueprint for Project A might mistakenly be sent to Project B's team, leading to construction discrepancies.
	Resource clashes	Projects might compete for the same resources, leading to potential delays.	Two projects might unknowingly book the same crane on the same date, leading to logistical challenges.
	Financial mismanagement	Possible misallocation of funds across overlapping projects.	Funds allocated for one project might inadvertently be spent on another overlapping project due to accounting errors.
	Scheduling conflicts	Overlapping projects can lead to misaligned timelines, causing project delays.	Two projects' timelines might clash, leading to delays as resources are spread thin.
	Contractual conflicts	Potential for conflicting contractual obligations between projects.	Contractual obligations for one project might interfere with another, e.g., exclusivity clauses with suppliers causing supply chain disruptions.

Overlap of personnel across multiple projects ($V_{2\text{-}5}$). It is common for personnel to handle multiple construction projects simultaneously. While resource-efficient, this can blur project boundaries, risking unintentional data leaks or access. For instance, an architect working on different projects using a single device might inadvertently mix or share data. During the interview, Expert 2 acknowledged the possibility of personnel overlap across multiple construction projects being a cybersecurity vulnerability, but with less severity compared to other vulnerabilities. He noted that while this practice is common and might theoretically lead to issues like data confusion or resource clashes, he has yet to observe any direct incidents where such overlap has resulted in cybersecurity breaches like data leaks or phishing attempts. However, he did not entirely dismiss the potential risk, suggesting that while it could be considered a vulnerability, its impact on cybersecurity has not been as pronounced or evident in the industry.

When considering the vulnerabilities of a specific asset, to ensure our framework remains effective against the dynamic landscape of cyber threats, a periodic review process can be incorporated and threat intelligence feeds can be integrated. This allows us to continuously update our vulnerability databases and adjust our scanning protocols to identify and mitigate newly emerging threats, ensuring our security measures are always aligned with current risks.

3.2. Define Assessment Objectives

The next step involves defining the objective of the risk assessment, as outlined in [29], which is the output of the ML model. Traditional risk assessment typically focuses on predicting the likelihood of an incident and assessing its impact [30]. However, given the powerful predictive capabilities of ML models, the objectives of risk assessment can be expanded. To this end, experts were interviewed, who were asked to specify what they would like to see as the output of the ML model. This information would enable them to manage cybersecurity more effectively based on the results. Consequently, three additional objectives have been identified and added to the list, as detailed in Table 4. Table 5 provides a comprehensive breakdown of the five categories of objectives. When addressing a specific cyber risk assessment task, it is beneficial to start by consulting this table for defining objectives. These objectives can be customized to suit the unique context and intricacies of the task(s) at hand.

Table 4. Identified objectives and expert opinions.

Objective Number	Description	Alignment with Strategic Business Goals	Expert 1 Opinion	Expert 2 Opinion
O_1	To predict the probability of a potential cybersecurity incident occurring within a specified time frame if the vulnerabilities are exploited.	This objective allows organizations to identify and prioritize risks effectively, helping allocate resources efficiently to safeguard business continuity.	Not explicitly mentioned, but implied as a sub-goal alongside $O_{2.}$	Highlighted as a sub-goal, important for understanding the probability of incidents.
O_2	To predict the severity of consequences if a cyber incident occurs. The impact can be measured in terms of financial loss, operational disruption, or damage to reputation.	Understanding potential impacts aids in preparing effective contingency plans and minimizing financial and reputational damage, which is crucial for strategic risk management.	Not explicitly mentioned, but implied as a sub-goal alongside $O_{1.}$	Identified as a sub-goal, crucial for assessing the severity of consequences of cyber incidents.
O_3	To predict the overall risk that combines the likelihood and impact, often represented as a score or level. This provides a summarized view of how critical the cyber incident is.	Quantifying risk with scores helps decision-makers prioritize threats and allocate cybersecurity resources strategically, aligning with business priorities.	Emphasized as the primary focus; crucial for understanding overall cybersecurity status in line with ISO 27001 [45].	Agrees with the prioritization of O_3; focuses on this in current project alongside O_1 and $O_{2.}$
O_4	To predict potential metrics or numbers related to the incident, such as the downtime of a system due to potential threats.	Estimating system downtime enables organizations to develop robust disaster recovery strategies, maintaining operational efficiency and customer satisfaction.	Not part of current project scope; no explicit opinion provided.	Important for understanding potential incidents and preparing accordingly, but not part of the current project scope.
O_5	This objective focuses on making the model generate answers to questions about cybersecurity posture, preparedness, and resilience. The models can be large generative language models (LLMs), such as GPT-4 [46] and Ernie Bot [47].	Using models like GPT-4 and Ernie Bot for cybersecurity assessments aids strategic decisions by providing insights into organizational readiness and vulnerabilities, enhancing long-term resilience.	Not part of current project scope; recognizes potential future value.	Deemed to have great potential for future projects, especially in creating tailored language models for cybersecurity.

During the interview, Expert 1 emphasizes the importance of O_3—evaluating the risk through metrics—as a primary focus. This preference aligns with the ISO 27001 standard [45], which mandates an initial assessment of current security levels and a gap analysis. According to him, understanding an organization's current overall cybersecurity status is crucial before moving on to other objectives. This evaluation could broadly apply, for instance, in assessing the overall risk for IT services in a head office. His viewpoint underscores the significance of establishing a foundational understanding of existing risks as a precursor to addressing and mitigating them. Expert 2 agrees with the prioritization of O_3 but also highlights the importance of O_4, which is critical for understanding potential incidents and preparing accordingly. Additionally, Expert 2 notes that their current project focuses on O_3, along with O_1 and O_2 as sub-goals. However, O_4 and O_5 are not part of their current project scope, but O_5 is deemed to have great potential because creating a tailored language model to assist with cyber risk management tasks is of significant interest, potentially saving labor and time.

Table 5. Identified objectives for cyber risk assessment.

Objective	Focus Area	Description	Scale/Unit
Predicting the likelihood of an incident (O_1)	Ransomware	Malicious software designed to block access to a computer system until a ransom is paid.	Likelihood (%)
	Phishing	Fraudulent attempts, often via email, to steal sensitive information by disguising it as trustworthy.	Likelihood (%)
	Data breach	Unauthorized access and retrieval of sensitive data.	Likelihood (%)
	Malicious insider attack	Harmful actions taken against an organization from someone within (i.e., malicious insider).	Likelihood (%)
	Supply chain attack	Targeting less-secure elements in the supply chain to compromise a primary target.	Likelihood (%)
Estimating potential impact (O_2)	Operational downtime	Period when operations are halted, affecting productivity.	Time duration (e.g., hours, days)
	Financial losses	Direct and indirect monetary losses due to a cyber incident.	Currency (e.g., USD)
	Reputational damage	Negative impact on a company's reputation following a cyber incident.	Qualitative assessment (e.g., low, medium, high)
	Legal and regulatory consequences	Legal penalties and regulatory fines following non-compliance or breaches.	Qualitative assessment (e.g., low, medium, high)
	Loss of intellectual property	Unauthorized access and theft of proprietary designs, processes, or ideas.	Count (e.g., number of files/documents)
	Compromised safety	Threats to human safety due to a cyber incident.	Incident count
Evaluating the risk through metrics (O_3)	Risk score	Numeric value representing the severity of a risk.	Numeric score (e.g., 0–100)
	Risk level	Categorical evaluation (e.g., low, medium, high) of the severity of a risk.	Qualitative assessment (e.g., low, medium, high)
Projecting statistical figures related to the incident (O_4)	The number of events	Count of specific cyber incidents over a time frame.	Count
	Operational downtime	Total time systems are non-operational due to incidents.	Time duration (e.g., hours, days)
	Affected systems count	Number of IT systems impacted by a cyber incident.	System count
	Data volume compromised	Amount of data, often in GB or TB, accessed without authorization.	Data volume (e.g., Terabytes (TB))
	Incident response time	Time taken to identify, react, and address a cyber incident.	Time duration (e.g., hours, days)
	User accounts affected	Number of user accounts compromised in an incident.	Count
Answering qualitative questions about cybersecurity (O_5)	Document checking, decision making, solution suggestion	Provide initial answers to the questions the user input to the language model.	Text (e.g., How would you rate the employees' familiarity and compliance with our organization's cybersecurity guidelines and best practices?)

3.3. Feature Design

The next step in the risk assessment process, as detailed in [29], begins with the identification of risk factors. These factors are integral to the subsequent design of features for ML models, which is vital for the development of effective ML models. In cyber risk management, a risk factor is typically understood to be any characteristic, condition, or behavior that can increase the likelihood of encountering one of the five cyber risks identified in Section 3.1 [30]. In ML, a feature is defined as a distinct, measurable attribute or characteristic of an observed phenomenon [48] that can serve as an input for ML models.

The features designed should reflect the vulnerabilities of the construction industry. The methodology in [29] includes a three-tiered literature review and expert consultations, originally designed for predicting delays in building construction projects. This comprehensive and structured approach can be adapted and applied to our context for formulating cyber risk factors and features. During the interview with the two experts, this process was thoroughly discussed and adapted into six distinct steps, enhancing its applicability to our specific needs.

(1) Initial Broad Search: In various academic databases, conduct multiple searches on terms related to cybersecurity and the construction industry and their variants, with the timeframe covering the last 10 years to ensure contemporary relevance.
(2) Focused Review: Review abstracts, keywords, and titles to filter out publications that specifically address cyber risks in construction environments, considering the unique digital landscape and vulnerabilities of this sector.
(3) In-Depth Analysis: Scrutinize articles to identify comprehensive cyber risk factors that could affect construction project assets.
(4) Expert Consultation and Validation: Engage with cybersecurity and construction IT experts to validate and potentially revise the identified risk factors, ensuring they align with the expertise of both industries.
(5) Finalization of Cyber Risk Factors: Utilize the updated literature findings and expert feedback to finalize cyber risk factors and classify them into different categories.
(6) Feature Design: Develop features based on the cyber risk factors, making sure they are suitable for use as inputs for ML models.

In [29], nine features are designed and categorized qualitatively into five risk levels: very low, low, medium, high, or very high. These features are known as ordinal features in ML [48]. However, there are other potential features that are more quantitative in nature and could more accurately reflect the vulnerabilities of the identified asset. In machine learning, features can be broadly classified into four categories [49] shown in Table 6. Considering various categories of features ensures comprehensive feature extraction and design, thereby enhancing the effectiveness of the cyber risk assessment process. For instance, when considering phishing as a risk in a construction project, various types of features can be identified, as exemplified in Table 6. However, we acknowledge that not all identified features are equally useful. To optimize the model, selected features can be used to preliminarily test performance. This is followed by a feedback loop from real-world implementations of the risk assessment, which informs further adjustments. Techniques such as feature importance scoring and recursive feature elimination can be employed to assess the impact of each feature on the model's predictive accuracy. This allows us to fine-tune our feature selection. Such continuous validation and refinement ensure that our model remains robust and effective in identifying and mitigating cyber risks in dynamic environments.

Table 6. Feature categories and examples.

Feature Category	Explanation	Example Feature	Relevance to Vulnerabilities (Section 3.1.3)
Numerical	It represents data that can be measured and expressed numerically	Percentage of personnel with access to sensitive information	V_1, $V_{2\text{-}1}$, $V_{2\text{-}2}$
Ordinal	It classifies data into categories with a specific order or scale	Level of security mechanism of OT equipment (very low, low, medium, high, very high)	$V_{1\text{-}3}$, $V_{1\text{-}4}$, $V_{2\text{-}2}$
Categorical	It classifies data into distinct categories that lack a numerical or ordered relationship.	Project location (categorized by country or city)	$V_{2\text{-}3}$, $V_{2\text{-}4}$
Boolean	It is binary and denotes a condition as either true or false	Presence of a dedicated IT team for the project (yes/no)	V_1, $V_{2\text{-}1}$, $V_{2\text{-}2}$

3.4. Data Collection

After designing the feature, the next step is to collect data consisting of samples, each composed of a set of predetermined features, for training ML models. As a starting point, the literature on risk assessment was reviewed, focusing on data collection methodologies [29,50–55]. Our review revealed that data collection in this domain often employs human-centric approaches, such as meetings, discussions, questionnaires, and brainstorming. All these methods are applicable for our data collection. To further enhance these findings and tailor our approach to cyber risk assessment, expert opinions on key areas of data collection were integrated.

3.4.1. Interdisciplinary Data Collection

The designed ML features address various asset aspects. For example, IT features might include network security protocols, OT features could cover equipment performance metrics, and management features may relate to maintenance budget allocations. Experts in these areas are better equipped to understand specific data requirements, thus improving the quality and relevance of data collection. Previous studies like [29,51,55] lacked detailed descriptions of personnel roles in data collection. Addressing this gap, our interviews detailed the involvement of specific personnel types in data collection, ensuring the development of accurate and relevant ML features. Five key personnel types were identified as relevant.

(1) Project Managers: They are central to data collection due to their comprehensive understanding of construction projects. They are typically familiar with various project aspects, including the assets involved. If project managers cannot provide the necessary data, it is advisable to consult with department heads in logistics, IT, and operations for specialized insights.

(2) IT Personnel: They play a crucial role in ensuring the security and efficiency of computer systems and networks. Their responsibilities include managing network infrastructure, implementing security protocols, monitoring system performance, and addressing IT-related issues. Consequently, they are well-positioned to provide data for features related to metrics like system downtime and communication intensity, among others.

(3) OT Personnel: They are responsible for managing and maintaining operational technology systems. Their tasks often involve overseeing the operation of machinery, ensuring the efficiency and safety of production processes, and conducting routine maintenance and repairs. Consequently, they can provide data for features such as equipment performance metrics and maintenance records, among others.

(4) Administrative Staff: They oversee essential organizational and clerical tasks, crucial for maintaining smooth operations across various departments. Accordingly, they can provide data for features related to project documentation, financial records, personnel data, client databases, communication histories, compliance reports, network access records, detailed incident reports, and thorough inventories of both hardware and software.

(5) Logistics Managers: They are instrumental in handling the logistical aspects of projects. Their key responsibilities include managing resources, overseeing the supply chain, and coordinating various operational activities. They can provide data for features related to resource utilization records, supply chain efficiency metrics, transportation and delivery schedules, inventory management statistics, and operational coordination logs.

3.4.2. Data Sources

When approaching personnel for data collection, it is crucial to consider various sources of data relevant to cybersecurity to ensure thoroughness and comprehensiveness. In the modern construction industry, six distinct modes of data can be identified [56]: (1) structured logs; (2) time-series data; (3) spatial data; (4) image data; (5) textual data;

(6) audio recordings. As discussed in Section 3.1.3, vulnerabilities can be categorized into two primary types: common IT vulnerabilities and those unique to the construction industry. Through discussions with experts, these data sources have been mapped to their respective vulnerabilities. This mapping is detailed in two tables. Table 7 illustrates available data mapped to common IT vulnerabilities, while Table 8 focuses on data mapped to vulnerabilities inherent to the construction sector. They serve as a reference for identifying suitable data sources that align with specific requirements and circumstances in cyber risk assessment tasks.

Table 7. Data mapped to common IT vulnerabilities.

Vulnerability Aspect	Structured Logs (D_1)	Time-Series Data (D_2)	Spatial Data (D_3)	Image Data (D_4)	Text Data (D_5)	Audio Data Mode (D_6)
Software Flaws ($V_{1\text{-}1}$)	–Construction management software logs –Equipment firmware versions –Building information modeling (BIM) software records	–Timeline of software updates in project management tools –Frequency of detected software issues from AutoCAD v25.0 tools	–Locations of on-site devices running specific software –Geographic distribution of cloud-based tools' data centers	–Screenshots of errors in scheduling or modeling software –Drone footage capturing software-driven machinery malfunctions	–Error logs from construction-specific apps –Feedback forms from site managers on software	–Recordings of construction software training sessions –Feedback from workers on software usability
Network Configurations ($V_{1\text{-}2}$)	–Network layout of construction site trailers –Router and switch settings at temporary site offices –Access logs from on-site servers	–Traffic flow over time from construction site to headquarters –Unauthorized access attempts on site-specific networks	–Locations of network hardware across construction site –Geographic layout of Wi-Fi boosters for large sites	–Network diagram visualizations for site office –Images of site-specific network setups	–Network setup guidelines for construction sites –Site IT team's notes	–Audio logs of network setup briefings at construction sites –Recordings of IT consultations for site-specific needs
Communication Protocol Weaknesses ($V_{1\text{-}3}$)	–Configuration settings for construction communication tools –List of approved communication tools for site	–Timeline of changes or updates in site communication tools –Detected issues with on-site communication systems	–Locations of devices using construction-specific communication tools –Distribution of push-to-talk device users	–Screenshots of communication device configurations –Photos of communication hubs at construction sites	Communication tool guidelines –Feedback from site workers on communication issues	–Audio feedback sessions about communication tools –Recorded discussions about tool selections for sites
Hardware Susceptibilities ($V_{1\text{-}4}$)	–Inventory of IoT devices on site –Firmware logs for construction machinery –Maintenance logs for IT equipment in site trailers	–Timeline of machinery updates or replacements –Frequency of IoT device malfunctions on site	–Geographic distribution of smart equipment across construction sites –Locations of construction drones' landing zones	–Images of machinery control panels –Drone-captured photos of large machinery in action	–Maintenance and issue logs for construction equipment –User manuals for construction-specific IT equipment	–Recordings of machinery training sessions –Maintenance feedback or alerts captured via audio

Table 8. Data mapped to vulnerabilities specific to the construction industry.

Vulnerability Aspect	Structured Logs (D_1)	Time-Series Data (D_2)	Spatial Data (D_3)	Image Data (D_4)	Text Data (D_5)	Audio Data Mode (D_6)
Frequent changes in teams ($V_{2\text{-}1}$)	–Employees' start and end dates –Roles and access levels –Onboarding checklists	–Timeseries graph showing frequency of team changes –Audit trails of access rights changes	–Locations of team members –Site access logs by different teams	–Badges/ID cards –Access logs with time-stamped images	–Team meeting minutes –Personnel change notifications	–Recorded interviews/feedback on team transitions –Audio logs of onboarding training

Table 8. *Cont.*

Vulnerability Aspect	Structured Logs (D_1)	Time-Series Data (D_2)	Spatial Data (D_3)	Image Data (D_4)	Text Data (D_5)	Audio Data Mode (D_6)
Varied levels of cybersecurity knowledge among personnel ($V_{2\text{-}2}$)	–Results from cybersecurity training assessments –Logs of approved/unapproved software usage –Incident response logs	–Timeline of cybersecurity incidents or breaches –Frequency of cybersecurity training sessions	–Geographic locations of cybersecurity training held	–Screenshots of training modules or breach notifications –Images of on-site cybersecurity posters/guidelines	–Training manuals –Feedback forms post-training –Reports of cybersecurity breaches/incidents	–Recordings from training sessions –Audio alerts from cybersecurity systems
Scattered and frequent communications ($V_{2\text{-}3}$)	–Logs of communication platforms –Timestamps and participants –Frequency of channel switches	–Frequency and timing of communications –Peaks in communication before major milestones	–Geographic distribution of stakeholders –Locations of communication relay nodes or boosters	–Screenshots of communication tools/channels used –Infographics/charts shared in communication	–Email threads –Chat logs –Memos and official communications	–Recorded calls/messages –Voice notes on communication platforms
Frequent exchange of digital information ($V_{2\text{-}4}$)	–Logs of data transfer events –Sizes of transferred files –Encryption status logs	–Timestamps of data transfers and accesses –Graphs of data volume exchanged over time	–Geographic locations of major data transfers (if applicable) –Locations of servers storing key data	–Visual representations of data flows –Screenshots of file transfer progress	–Descriptions or notes on data transfers –Digital handover notes –Logs of file names and types transferred	–Voice logs or confirmations of successful data transfers –Audio alerts from data transfer systems
Personnel overlapping across multiple projects ($V_{2\text{-}5}$)	–Resource-allocation logs –Financial records per project –Timeline overlaps	–Timelines showing project milestones and deliveries –Gantt charts of overlapping projects	–Geographic overlap or proximity of project sites –Maps with resource locations for different projects	–Visual charts/graphs depicting overlapping timelines –Pictures of sites showing simultaneous work	–Written project briefs/descriptions –Logs of resource requests and allocations	–Audio updates or briefings about overlapping projects –Recorded meetings discussing project overlaps

To effectively collect cybersecurity data in the construction industry, a variety of the latest techniques can be adapted. Real-time data consolidation from multiple sources can be achieved through automated integrated systems, such as employing energy-efficient routing techniques within a cloud-based software defined network (SDN) system for intelligent-Internet of Things (I-IoT) networks [57]. Building information modeling (BIM) and geographic information systems (GISs) can be used to collect detailed 3D spatial data [58], while large-scale distributed devices and IoT sensors can be used to gather industrial processing data [57,59]. Additionally, machine vision and deep learning techniques can be utilized to collect and analyze on-site data points, images, audios, and videos, enhancing monitoring and security measures [60,61]. Techniques such as wireless sensor systems can help in gathering environmental parameters, and artificial speech recognition technology can capture voice commands in smart home environments, integrating these into broader security protocols [62]. Furthermore, the unsupervised clustering NLP techniques combined with Accimap modeling can extract and analyze factors from accident reports to identify potential security vulnerabilities [63]. The integration of activity theory with GIS aids in collecting and analyzing qualitative data on user interactions and system performance, offering insights into operational security [64]. Moreover, a cyber–physical fusion method can be employed to integrate and apply multistage equipment data, furthering comprehensive data usability and enhancing cybersecurity across various construction phases [65]. In situations where actual data collection is challenging, computational models and simulation techniques can also be used to generate data [66]. These techniques, originally developed for diverse applications, hold great promise for adaptation to enhance cybersecurity measures in the construction industry. Their implementation could significantly improve the detection, analysis, and management of cyber threats in this increasingly digitized sector. However, as highlighted in the interviews, there are challenges associated with interdisciplinary data collection in the construction industry. These challenges, along with potential solutions, will be discussed in detail in Section 4.1.

3.5. ML Model Development

With the dataset prepared, selecting an appropriate model for training the dataset and predicting the risk for new samples is crucial. Given the designed features, the objectives of risk prediction (O_1–O_4) are formalized, as shown in Equation (1). Objective O_5, related to text generation, can be addressed by fine-tuning modern language models such as PaLM [67], LaMDA [68], or GPT-4 [46] with an industry-specific text corpus, a topic to be discussed in Section 4.3.

3.5.1. Mathematical Expression

Risk assessment objectives O_1–O_4 can be formalized with Equation (1).

$$\hat{y} = f(rf_1, rf_2, \ldots, rf_j, \ldots, rf_J) \quad (1)$$

where $\hat{y}$ is the predicted value of the assessment objective; f is the ML model; rf_j is the j-th feature; J is the number of features. For numerical features, the features can be directly input to the model; for others, the features should be converted into one-hot encoding [69] and rf_j can be represented by $P_{r_j} = (Pr_{j,1}, \ldots, Pr_{j,k}, \ldots, Pr_{j,K^{(j)}})$, where $1 \le k \le K^{(j)}$ and $K^{(j)}$ is the number of scales of this feature. Then Equation (1) can be converted to Equation (2).

$$\hat{y} = f((Pr_{1,1}, \ldots, Pr_{1,K^{(1)}}), \ldots, (Pr_{j,1}, \ldots, Pr_{j,K^{(j)}}), \ldots, (Pr_{J,1}, \ldots, Pr_{J,K^{(J)}})) \quad (2)$$

For a numerical feature, $K^{(j)}$ equals to 1, and Pr_j can be simplified as the same as rf_j;

For an ordinal feature, $K^{(j)}$ is the number of levels; if k is the selected level for the feature, then $Pr_j = (0, \ldots, 1_{(k)}, \ldots, 0)$;

For a categorical feature, $K^{(j)}$ is the number of categories; if k is the selected category for the feature, then $Pr_j = (0, \ldots, 1_{(k)}, \ldots, 0)$;

For a Boolean feature, $K^{(j)}$ equals 2; then $Pr_j = (0, 1)$ or $Pr_j = (1, 0)$.

3.5.2. A Uniform Model

When selecting ML models, dataset size is crucial. Larger datasets often require more complex models to effectively learn from the abundant patterns without overfitting [69]. Conversely, simpler models are better suited for smaller datasets to prevent overfitting. However, with an increase in dataset size, a transition to more complex models may be necessary to capture data complexity, necessitating model re-training. In the construction industry, initial data collections are usually small, making simpler models like decision trees [70] and naïve Bayesian models [71] practical. However, it is important to prepare for data volume increases over time. During the interview, the experts were asked about their preference for using different models versus a uniform model for future scalability. The consensus favored a uniform model architecture to reduce the frequency of model updates and leverage prior knowledge, saving time and resources. This approach also simplifies model deployment and ongoing management, offering a more sustainable solution in the long term [72].

3.5.3. Artificial Neural Networks (ANNs)

In recent years, the use of ANNs in the construction industry for risk management has gained attention. In study [73], ANNs are utilized to analyze and predict the severity of occupational injuries within the industry. In study [74], neural networks are applied to examine the relationship between OSHMS elements and safety performance in Singapore, aiding in accident prediction and prevention. Additionally, study [75] showcases a model developed using ANNs to predict the impact of construction risks on cost flow forecasts, leveraging data from UK case studies and surveys. These examples of ANN applications illustrate the potential for utilizing ANNs for the cyber risk assessment field.

ANNs, inspired by the structure and cognitive processes of the human brain, provide a powerful computational model for pattern recognition and complex data processing. These networks mimic neuronal interactions in the brain, a concept that has been extensively explored in the literature, including seminal works by scholars like Ian Goodfellow [69]. A key strength of ANNs lies in their ability to learn complex, non-linear relationships between features. This contrasts with linear regression models, as ANNs can autonomously learn feature transformations, leading to more precise predictions. This is especially valuable in the construction industry, where data is often multifaceted and unpredictable [69]. Additionally, ANNs are known for their flexibility and ease of use, making them suitable for uniform application across diverse tasks.

In this study, ANN is selected as the uniform model used for cyber risk assessment in construction. A feedforward ANN, composed of neurons in multiple layers, utilizes hidden layers to perform this automatic transformation. The operation of a hidden layer with m neurons, following a layer with n neurons, is described by Equation (3).

$$h_j = \sigma(\sum_{i=1}^{n} w_{ij}x_i + b_j) \tag{3}$$

where

h_j—the value of the j-th neuron in the hidden layer,

x_i—the value of the i-th neuron in the previous layer,

w_{ij}—the weight of the i-th neuron in the previous layer that connects to the j-th neuron in the hidden layer,

b_j—the bias term for the j-th neuron in the hidden layer,

σ—the activation function. Various activation functions are available for use, for example: ReLU [76], LeakyReLU [77], and Tanh [69].

3.5.4. Loss Function

The features discussed in Equation (2) are fed into the first layer, while the outcome $\hat{y}$ of the prediction is obtained from the final layer. The training process focuses on reducing the difference between this predicted value and the actual value of the assessment objective. This difference is measured using a loss function. For regression tasks, where the goal is to predict a continuous numerical value, mean squared error (MSE) is often used as the loss function, as detailed in Equation (4) [69]. In contrast, for classification tasks that involve predicting discrete categories, the cross-entropy loss is employed, as shown in Equation (5) [69]. The choice of loss function is crucial in guiding the supervised learning process towards accurate predictions.

$$L_{MSE} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_i - y_i)^2 \tag{4}$$

$$L_{CE} = -\frac{1}{N}\sum_{i=1}^{N}\sum_{c=1}^{M} y_{i,c}\log(\hat{y}_{i,c}) \tag{5}$$

where N is the number of data samples for training; M is the number of categories of risk assessment objectives if discrete; $y_{i,c}$ is a binary indicator (0 or 1) if class label c is the correct classification for the data sample i; $\hat{y}_{i,c}$ is the predicted probability that data sample i belongs to class c.

3.5.5. Objectives and Loss Function

Risk assessment objectives O_1–O_4 can be categorized to a regression or classification task based on their output, and can then utilize the respective loss functions described in Section 3.5.4.

- O_1: This is formulated as a regression task when the goal is to predict the likelihood as a continuous numerical value ranging from 0 to 1. However, if the likelihood is divided into discrete levels (e.g., low, medium, high), it becomes a classification task.
- O_2: This is typically a regression task, as impacts (such as financial losses) are often quantified with continuous values. If, however, the impact is segmented into distinct levels, this objective then transitions to a classification task.
- O_3: This can be approached as a regression task when the aim is to predict a continuous risk score. In contrast, it becomes a classification task if the prediction involves discrete risk levels.
- O_4: This is generally treated as a regression task, particularly when the statistical figures of interest are continuous. Examples include system downtime or the number of incidents within a specific time frame.

To evaluate model performance, MAE, RMSE, R-squared, and MSE can be used for regression tasks, and accuracy, precision, recall, and F1 score can be used for classification [48]. Since there are no existing benchmarks for cyber risk assessment in construction, companies should establish a baseline model to assess improvements and track progress.

3.6. Risk Factor Prioritization

After training the ML model, feature importance analysis (FIA) can be used to evaluate the overall importance of different risk factors [78], providing insights into their impact on risk predictions. This analysis helps rank the risk factors, supplying asset managers with a prioritized list essential for effective risk management [30]. Such prioritization supports the development of proactive mitigation strategies. Additionally, feature contribution analysis (FCA) offers a more detailed view by focusing on how individual factors contribute to the risk of a specific asset. Applying FCA allows us to determine the most influential risk factors for a particular scenario [79], aiding asset managers in refining their focus on critical areas for risk factor prioritization.

Various methods have been developed to analyze feature importance and contribution in ML. Notably, SHAP [80] stands out as a unified approach applicable across different models and is frequently employed in risk analysis studies [81–84]. Its model-agnostic nature makes it suitable for a wide range of neural network architectures. SHAP effectively handles high-dimensional data and accounts for interaction effects among features, making it particularly useful for complex analyses [80]. The method's ease of implementation facilitated by Python's "SHAP" package adds to its practicality.

The contribution of each feature in ANNs is quantified as the SHAP value demonstrated in Equation (6) [80], which is 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 can be obtained.

$$SHAP_{j,k} = \sum_{Pr_S \subseteq Pr \setminus \{j,k\}} \frac{|Pr_S|!(|Pr| - |Pr_S| - 1)!}{|Pr|!} \left[f(Pr_S \cup \{Pr_{j,k}\}) - f(Pr_S) \right] \quad (6)$$

where

Pr—the set of all features,
Pr_S—a subset of features not including $Pr_{j,k}$,
$f(Pr_S \cup \{Pr_{j,k}\})$—the prediction with both the Pr_S and $Pr_{j,k}$,
$f(Pr_S)$—the prediction with just the features in Pr_S,

$\frac{|Pr_S|!(|Pr|-|Pr_S|-1)!}{|Pr|!}$—a combinatorial coefficient ensuring that each possible combination of features is weighted appropriately.

The importance of feature j is determined as the mean absolute SHAP value for all its scales, averaged across all samples in the training dataset, described in Equation (7).

$$I_j = \frac{1}{N \cdot K^{(j)}} \sum_{i=1}^{N} \sum_{k=1}^{K^{(j)}} \left| SHAP_{i,j,k} \right| \tag{7}$$

where

N—the number of samples in the training dataset,

$SHAP_{i,j,k}$—the SHAP value of the k-th scale of the j-th feature for the i-th data sample.

The contribution of a feature is determined by the sum of contributions across all its scales, shown in Equation (8).

$$C_j = \sum_{k=1}^{K^{(j)}} (SHAP_{j,k}) \tag{8}$$

Since this study focuses primarily on the conceptual intersection of cyber risk assessment and ML approaches, the intricacies of model training, testing, and validation will not be explored. The intention is to provide readers with a broader understanding of how ML techniques can be applied to cyber risk assessment tasks specific to the construction industry. For those interested in the details of model training, validation, and metric selection, the work by Goodfellow et al. [69] is highly recommended.

4. Discussions

4.1. Challenges of Data Collection

In the interview, data preservation was thoroughly discussed with two experts. Expert 2 pointed out that data preservation, especially for operational technology (OT) data crucial for monitoring cybersecurity [43], is still nascent within the construction sector. These data, often from heavy machinery monitoring, are mainly used for operational and productivity analyses and not cybersecurity. The collection of OT data is largely manual, handled by administrative rather than specialized technical staff, which hampers efficiency. In contrast, as Expert 1 noted, IT data collection is more advanced, offering systematic access to structured data such as logs and network configurations that reveal vulnerabilities. This highlights a significant gap in the construction industry's cyber risk assessment capabilities and underscores the need for a structured approach to data collection and preservation that includes both OT and IT, as well as management and project planning aspects. To improve ML model training, a deep understanding of data collection challenges in construction is crucial. In this section, four main challenges along with the proposed solutions were discussed.

4.1.1. Lack of Specific Data Structure

There is an absence of a structured approach for data collection, which makes it difficult to systematically collect and utilize data for model training or cybersecurity purposes. A structured format is needed to guide what data should be collected, especially for OT data. Once the feature design is completed, these features should be structured into a fixed format and presented to relevant companies for data collection. Implementing a standardized data collection protocol can guide the collection of relevant data. This protocol should outline the types of data needed, methods of collection, and the frequency of updates.

4.1.2. Data Collection from Closed Projects

Data are often collected retrospectively from projects that have already been completed, limiting real-time application or insights on live projects. Establishing a system for ongoing data collection during the lifecycle of a project could be a potential solution. This could involve implementing real-time data tracking tools and processes, which allow for immediate data analysis and application. Regular updates and reports during the project can provide timely insights that are more relevant and actionable.

4.1.3. Reactive Approach to Data Collection

The current data collection processes are reactive rather than proactive, with data often being reviewed after incidents have occurred. This approach limits the potential for predictive analysis and timely intervention. Shifting towards a proactive data collection strategy, which includes regular monitoring and real-time data analysis, can help in identifying potential risks early and mitigating them before they escalate.

4.1.4. Integrating Data from Diverse Sources

Integrating data from various departments like logistics, IT, and operations is challenging due to the diverse nature of the data sources. Developing a centralized data integration platform that can handle various data formats and sources is one potential solution. This platform should include robust data processing capabilities to harmonize and standardize data from different departments. Employing advanced data integration tools and middleware can facilitate seamless data merging and analysis, leading to a more comprehensive understanding of the project's dynamics. When collecting and using such data, stringent protocols must be enforced to ensure the confidentiality and integrity of the information. This involves encrypting data both in storage and in transmission, implementing access controls, and regularly auditing data usage and security policies. Model training, based on the data collected, should also maintain strict confidentiality and be protected during its deployment process, which should occur in close collaboration with IT and machine learning experts. Ethical considerations should guide the collection and utilization of data, ensuring that it does not infringe on privacy rights or expose sensitive information inadvertently. These practices safeguard valuable data and maintain trust among stakeholders.

4.2. Dynamicity of ML Models

As pointed out in [72], the dynamicity of the ML model is important for risk assessment. This concept refers to the continuous learning and adaptation of the ANN model in response to evolving data inputs, thereby enhancing its predictive capabilities. The importance of dynamicity is underlined by three key aspects discussed in the interviews, which are aligned with industry needs.

4.2.1. Continuous Training

The construction industry experiences daily variations in many factors like the number of devices, personnel changes, and operational adjustments. For a model to be dynamic, it is crucial not only to absorb new data reflecting these variations but also to continuously adapt and learn from this data. However, retraining models from scratch every time is computationally expensive and delays the time it takes to update the model with new data. To address this, a warm-starting system, as implemented by Cheng et al. [85], can be beneficial. This system initializes a new model with embeddings and weights from the previous model, enabling quicker updates and less computational demand.

4.2.2. Real-Time Monitoring

Experts have emphasized the necessity of real-time monitoring within a robust cybersecurity framework. By enabling the model to continuously predict risks, it facilitates immediate incident response and behavior analysis. This continuous prediction is achieved by feeding the model with real-time data, leading to ongoing updates in risk assessment. For instance, once the model is trained, it can be deployed in a live project, which is subject to daily changes. By making daily risk predictions, the model effectively monitors the cybersecurity posture of the construction project in real-time. Depending on the specific requirements for risk prediction, the interval between predictions can be adjusted, ranging from hours to minutes.

4.2.3. Changing Risk Factors

In the initial stages of model development, the features are usually predetermined and fixed. However, in the ever-changing construction industry, new risk factors may emerge, necessitating the incorporation of additional features into the model. The ability to integrate and process these evolving features is a key aspect of the model's dynamicity. This challenge of integrating new risk factors is referred to as the "emergence" problem in [72]. To effectively address this emergence challenge, progressive learning techniques can be applied according to Venkatesan and Er [86], which allow the model to learn new features as relevant information surfaces.

4.3. Advanced Language Models

During the interview, Objective 5, focusing on the use of generative models for cyber risk management in the construction industry, was discussed in-depth. The consensus among the experts is that language models, such as GPT-4 [46], Ernie Bot [47], LaMDA [68] hold significant potential for enhancing cybersecurity management. These models are particularly skilled at providing insights into current cybersecurity challenges, offering recommendations for security frameworks, and assisting in the strategic design of cybersecurity measures for specific projects. In scenarios requiring incident response, these models can offer prompt guidance and suggest potential solutions, thus aiding IT personnel in efficiently addressing cybersecurity concerns.

Compared to traditional cybersecurity risk assessment tools, which typically rely on manual data analysis and predefined algorithms, these language models represent a significant departure. They leverage natural language processing to understand and generate human-like responses, allowing them to analyze vast amounts of unstructured data rapidly. This capability provides a distinct advantage in identifying emerging threats and synthesizing complex information into actionable insights. However, language models have inherent limitations, notably their dependence on the training data, which can introduce biases and inaccuracies if the dataset is not comprehensive or current. Furthermore, in situations requiring strict regulatory compliance and precise technical accuracy, traditional methods may still outperform these models due to their consistent adherence to established protocols. Additionally, it is crucial to cross-verify the information provided by language models with other reliable sources. This is essential because relying solely on these models for critical decision-making could lead to significant oversights, particularly as these models may not always have access to the most recent or context-specific information.

To maximize their effectiveness within the construction industry, it is proposed that language models be fine-tuned through training on a corpus specifically related to construction and cybersecurity. The sentences and paragraphs in the corpus should be of really high quality in proper wording, correct syntax, and rich semantics. This tailored training, part of our future works, will ensure that the models are well-equipped with knowledge of industry-specific terminology, challenges, and best practices, thereby enhancing their utility and reliability in this domain.

4.4. Addressing the Practical Challenges of the Framework

While the proposed framework promises robustness, its practical implementation in the diverse environments of construction projects may expose several challenges, including variability in data quality, differing IT infrastructures across companies, and resistance to adopting new technologies among staff. Acknowledging the nascent stage of ML application in this field, our framework incorporates features for adaptability, enabling customization to the unique technological and operational contexts of various construction projects. For example, by allowing the customization of ML features, the framework can manage variations in data availability and granularity, addressing inherent ML limitations such as overfitting and training data biases. To further mitigate these biases, our approach includes using diverse datasets, robust validation techniques, and incorporating bias detection mechanisms within the ML models. Furthermore, the framework is designed to integrate feedback loops from initial deployments, enabling continuous refinement based on real operational feedback and evolving risk scenarios. Proactive strategies such as pilot testing and phased implementation are also part of our approach, facilitating gradual adaptation to real-world conditions and validating the framework's efficacy and reliability in practical settings. These steps are critical for bridging the gap between the theoretical aspects of our ML framework and its on-ground applications, thereby enhancing its practical utility and effectiveness, despite these adaptations not being tested within the scope of this study.

5. Conclusions and Future Works

This study develops an ML framework designed for cyber risk assessment in the construction industry. Insights were refined through semi-structured interviews with two industry experts. The framework consists of six modules: identifying risks, defining assessment objectives, designing features, collecting data, developing the ML model, and analyzing risk factors. This study discusses the development of effective ML models, focusing on key areas: (1) data collection challenges and solutions; (2) ML-model dynamicity, featuring continuous updates, real-time risk monitoring, and adaptive learning; (3) the use of advanced language models like GPT-4 in enhancing cybersecurity management; (4) practical challenges and potential solutions in implementing this framework. The future implementation of this framework promises enhanced cyber risk management, which can improve security measures by leveraging machine learning for real-time risk monitoring and industry-specific data, ultimately leading to more efficient and precise risk assessment and mitigation practices.

A limitation of this framework is that it has not yet been implemented or tested with real data. This study's focus has been primarily on the theoretical integration of ML and cyber risk assessment, specifically tailored to the construction sector. Instead, our study is based on an extensive literature review and in-depth semi-structured interviews with industry experts, which were instrumental in formulating, supporting, and validating our framework. This conceptual framework prepares for future implementation and testing. Our plans are: (1) to collaborate with a UAE construction company to develop a dedicated ML framework for cyber risk assessment, determining risk types, the assets, vulnerabilities, prediction objectives, and feature design; (2) to use synthetic data from Monte Carlo simulations for initial model training due to the unavailability of real data; (3) to create a standardized data collection protocol to help construction companies gather data, aiming to replace synthetic with real data for model training; and (4) to deploy the trained model to mobile or web applications, enabling real-time risk monitoring and decision-making. Success will be measured by the model's ability to prevent or mitigate cyber risks within a specified project period, a detail to be finalized with the construction companies. The goal is to deploy the real data-trained model on digital platforms within one year. Concurrently, for Objective 5, a large text database on cybersecurity in construction has been collected, which can be used for fine-tuning a language model for cybersecurity management.

In the long run, the vision is for the creation of an intelligent, off-the-shelf cyber risk management system that enables practitioners to seamlessly input relevant data and receive insightful cyber risk analysis results in return. This will be a joint effort by experts in cybersecurity, machine learning, and the construction industry. Experts in cybersecurity can provide valuable insights into the types of cyber risks and the cybersecurity-related features that need to be identified. Machine learning specialists will contribute to the model building, training, and deployment processes. Construction industry professionals will assist in identifying assets and pinpointing their vulnerabilities, which are also crucial for defining the features for the machine learning models. Their collective expertise will facilitate a more accurate and effective cyber risk assessment for the assets in construction projects.

Author Contributions: Conceptualization, D.Y.; Methodology, D.Y.; Software, D.Y.; Validation, D.Y.; Formal analysis, D.Y.; Investigation, B.G.d.S.; Resources, D.Y. and B.G.d.S.; Data curation, D.Y.; Writing—original draft, D.Y.; Writing—review & editing, B.G.d.S.; Visualization, D.Y.; Supervision, B.G.d.S.; Project administration, B.G.d.S. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Data Availability Statement: Dataset available on request from the authors.

Acknowledgments: This work was supported by the Center for Cyber Security (CCS), funded by Tamkeen under the NYUAD Research Institute Award G1104 and in collaboration with the NYUAD Center for Interacting Urban Networks (CITIES), funded by Tamkeen under the NYUAD Research Institute Award CG001. We thank ALEC Engineering & Contracting LLC (ALEC) for their involvement, in particular Sabyasachi Jana and Ali Raza, for participating in the interviews, supporting the validation of the framework and providing insightful feedback.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Klinc, R.; Turk, Ž. Construction 4.0—Digital Transformation of One of the Oldest Industries. *Econ. Bus. Rev.* **2019**, *21*, 393–496. [CrossRef]
2. Mantha, B.R.K.; García de Soto, B. Cyber Security Challenges and Vulnerability Assessment in the Construction Industry. In Proceedings of the Creative Construction Conference 2019, Budapest, Hungary, 29 June–2 July 2019; Budapest University of Technology and Economics: Budapest, Hungary, 2019; pp. 29–37.
3. Emma, J. *Cyber Security Breaches Survey 2020*; Department for Digital, Culture, Media & Sport: London, UK, 2020; Volume 2020, p. 4. [CrossRef]
4. Phishing Attacks in the Construction Industry. Infosec. Available online: https://resources.infosecinstitute.com/topic/phishing-attacks-construction-industry/ (accessed on 15 March 2021).
5. Kunert, P. Shut the Front Door: Jewson Fesses up to Data Breach. The Register. Available online: https://www.theregister.com/2017/11/14/jewson_suffers_data_breach/ (accessed on 15 March 2021).
6. Sawyer, T.; Rubenstone, J. Construction Cybercrime is on the Rise. Engineering News-Record. Available online: https://www.enr.com/articles/46832-construction-cybercrime-is-on-the-rise (accessed on 23 April 2021).
7. Tunney, C. Ransomware Attack on Construction Company Raises Questions About Federal Contracts. CBC News. Available online: https://www.cbc.ca/news/politics/ransomware-bird-construction-military-1.5434308 (accessed on 15 March 2021).
8. Korman, R. Hoffman Construction Reports Hack of Self-Insured Health Plan Data. Engineering News-Record. Available online: https://www.enr.com/articles/51232-hoffman-construction-reports-hack-of-self-insured-health-plan-data (accessed on 15 March 2021).
9. Christopher, H. *Cyber Risk Management: Prioritize Threats, Identify Vulnerabilities, and Apply Controls*; Jellyfish, Ed.; Kogan Page Limited: New York, NY, USA, 2019. Available online: https://books.google.com/books?hl=en&lr=&id=yuWYDwAAQBAJ&oi=fnd&pg=PR1&dq=ML+methods+can+make+full+use+of+the+abundant+past+cyber+risk+estimate+data+to+generate+accurate+results+with+higher+expediency.&ots=6_54ITiJsu&sig=wZwSvARLpPrgO12ALRdEmhNbEhU#v=onep (accessed on 10 April 2024).
10. Kalinin, M.; Krundyshev, V.; Zegzhda, P. Cybersecurity Risk Assessment in Smart City Infrastructures. *Machines* **2021**, *9*, 78. [CrossRef]
11. Yao, D.; García de Soto, B. A Preliminary SWOT Evaluation for the Applications of ML to Cyber Risk Analysis in the Construction Industry. *IOP Conf. Ser. Mater. Sci. Eng.* **2022**, *1218*, 012017. [CrossRef]
12. NIST (National Institute of Standards and Technology). *Framework for Improving Critical Infrastructure Cybersecurity*; Version 1.1; NIST: Gaithersburg, MD, USA, 2018.
13. *ISO/IEC 27000:2018*; Information Technology—Security Techniques—Information Security Management Systems—Overview and Vocabulary. ISO (International Organization for Standardization): Geneva, Switzerland. Available online: https://standards.iso.org/ittf/PubliclyAvailableStandards/c073906_ISO_IEC_27000_2018_E.zip (accessed on 11 October 2021).
14. CIS (Center for Internet Security). *Center for Internet Security Controls*; Version 7.1; CIS: New York, NY, USA, 2019. Available online: https://learn.cisecurity.org/20-controls-download?_gl=1*2ttlk*_ga*MjA0MDEzNDk4LjE2ODQyNTE4MDI.*_ga_N70Z2MKMD7*MTY4NDI1NDcwMS4yLjEuMTY4NDI1NDcxMy40OC4wLjA.*_ga_ZQVR7NM9HJ*MTY4NDI1NDcwMS4yLjEuMTY4NDI1NDcxMy4wLjAuMA (accessed on 11 October 2021).
15. Part 500 Cybersecurity Requirements for Financial Services Companies. 2017. Available online: https://govt.westlaw.com/nycrr/Browse/Home/NewYork/NewYorkCodesRulesandRegulations?guid=I5be30d2007f811e79d43a037eefd0011&originationContext=documenttoc&transitionType=Default&contextData=(sc.Default) (accessed on 11 December 2023).
16. Mantha, B.R.K.; García de Soto, B. Cybersecurity in Construction: Where Do We Stand and How Do We Get Better Prepared. *Front. Built Environ.* **2021**, *7*, 1–13. [CrossRef]

17. Salami Pargoo, N.; Ilbeigi, M. A Scoping Review for Cybersecurity in the Construction Industry. *J. Manag. Eng.* **2023**, *39*, 03122003. [CrossRef]
18. Bello, A.; Maurushat, A. Technical and Behavioural Training and Awareness Solutions for Mitigating Ransomware Attacks. In *Advances in Intelligent Systems and Computing*; Springer International Publishing: Cham, Switzerland, 2020; Volume 1226, pp. 164–176. [CrossRef]
19. El-Sayegh, S.; Romdhane, L.; Manjikian, S. A critical review of 3D printing in construction: Benefits, challenges, and risks. *Arch. Civ. Mech. Eng.* **2020**, *20*, 34. [CrossRef]
20. Turk, Ž.; García de Soto, B.; Mantha, B.R.K.; Maciel, A.; Georgescu, A. A Systemic Framework for Addressing Cybersecurity in Construction. *Autom. Constr.* **2022**, *133*, 103988. [CrossRef]
21. Parn, E.A.; Edwards, D. Cyber threats confronting the digital built environment: Common data environment vulnerabilities and block chain deterrence. *Eng. Constr. Archit. Manag.* **2019**, *26*, 245–266. [CrossRef]
22. Goh, G.D.; Sing, S.L.; Yeong, W.Y. A Review on Machine Learning in 3D Printing: Applications, Potential, and Challenges. *Artif. Intell. Rev.* **2021**, *54*, 63–94. [CrossRef]
23. Shemov, G.; García de Soto, B.; Alkhzaimi, H. Blockchain Applied to the Construction Supply Chain: A Case Study with Threat Model. *Front. Eng. Manag.* **2020**, *7*, 564–577. [CrossRef]
24. Pan, Z.; Hariri, S.; Pacheco, J. Context Aware Intrusion Detection for Building Automation Systems. *Comput. Secur.* **2019**, *85*, 181–201. [CrossRef]
25. Sheikh, A.; Kamuni, V.; Patil, A.; Wagh, S.; Singh, N. Cyber Attack and Fault Identification of HVAC System in Building Management Systems. In Proceedings of the 2019 9th International Conference on Power and Energy Systems (ICPES), Perth, WA, Australia, 10–12 December 2019; IEEE: Piscataway, NJ, USA, 2019; pp. 1–6. [CrossRef]
26. Mantha, B.; García de Soto, B.; Karri, R. Cyber security threat modeling in the AEC industry: An example for the commissioning of the built environment. *Sustain. Cities Soc.* **2020**, *66*, 102682. [CrossRef]
27. Mohamed Shibly, M.U.R.; García de Soto, B. Threat Modeling in Construction: An Example of a 3D Concrete Printing System. In Proceedings of the 37th International Symposium on Automation and Robotics in Construction, Kitakyushu, Japan, 27–28 October 2020. [CrossRef]
28. Mantha, B.R.K.; García de Soto, B. Assessment of The Cybersecurity Vulnerability of Construction Networks. *Eng. Constr. Archit. Manag.* **2021**, *28*, 3078–3105. [CrossRef]
29. Gondia, A.; Siam, A.; El-Dakhakhni, W.; Nassar, A.H. Machine Learning Algorithms for Construction Projects Delay Risk Prediction. *J. Constr. Eng. Manag.* **2020**, *146*, 04019085. [CrossRef]
30. Meyer, T.; Reniers, G. *Engineering Risk Management*; De Gruyter: Berlin, Germany, 2022. [CrossRef]
31. Deloitte Building Cybersecurity in the Construction Industry. Available online: https://www2.deloitte.com/ce/en/pages/real-estate/articles/ce-building-cybersecurity-in-the-construction-industry.html (accessed on 30 September 2023).
32. ENR InfoCenter. Why Your Construction Company Needs a Good Cybersecurity Strategy. Engineering News-Record. Available online: https://www.viewpoint.com/en-gb/blog/why-its-critical-to-have-a-sound-cybersecurity-strategy?redirected=y (accessed on 17 December 2023).
33. Barbaschow, A. Bouygues Construction Falls Victim to Ransomware. ZDNET. Available online: https://www.zdnet.com/article/bouygues-construction-falls-victim-to-ransomware/ (accessed on 30 September 2023).
34. Thibault, M. Skender Hit by Ransomware Attack. ConstructionDive. Available online: https://www.constructiondive.com/news/skender-ransomware-attack-chicago-maine/712844/ (accessed on 12 May 2024).
35. Stiles, M. Turner Construction Data Breach Exposes Hundreds in Washington to Possible Fraud. The Business Journals. Available online: https://www.bizjournals.com/seattle/blog/techflash/2016/04/turner-construction-data-breach-exposes-hundreds.html (accessed on 15 July 2021).
36. LaRosa, B. Cyber Security and Cyber Threats in the Construction Industry. WINDOVER. Available online: https://www.windover.com/blog/cyber-security-cyber-threats-construction-industry/ (accessed on 12 May 2024).
37. Cyware. Hackers Hit French Firm Ingerop Stealing 65 GB Data Relating to Nuclear Power Plants. CYWARE SOCIAL. Available online: https://cyware.com/news/hackers-hit-french-firm-ingerop-stealing-65-gb-data-relating-to-nuclear-power-plants-f193b9ba/ (accessed on 22 March 2023).
38. Coble, S. Major Canadian Military Contractor Compromised in Ransomware Attack. Infosecurity Magazine. Available online: https://www.infosecurity-magazine.com/news/bird-construction-compromised-in/ (accessed on 21 October 2022).
39. McCabe, M.; Tullett, J.; Bradshaw, A. Cyber Risk and the Construction Supply Chain. MarshMcLennan. Available online: https://www.marshmclennan.com/insights/publications/2021/april-/cyber-risk-and-the-construction-supply-chain.html (accessed on 12 May 2024).
40. Cybersecurity. International Telecommunication Union (ITU). Available online: https://www.itu.int/en/ITU-T/studygroups/com17/Pages/cybersecurity.aspx (accessed on 7 December 2023).
41. Wunder, J.; Halbardier, A.; Waltermire, D. Specification for Asset Identification. Available online: https://nvlpubs.nist.gov/nistpubs/Legacy/IR/nistir7693.pdf (accessed on 7 December 2023).
42. Greco, M.; Cricelli, L.; Grimaldi, M. A strategic management framework of tangible and intangible assets. *Eur. Manag. J.* **2013**, *31*, 55–66. [CrossRef]

43. Sonkor, M.S.; García de Soto, B. Operational Technology on Construction Sites: A Review from the Cybersecurity Perspective. *J. Constr. Eng. Manag.* **2021**, *147*, 04021172. [CrossRef]
44. Yao, D.; García de Soto, B. A corpus database for cybersecurity topic modeling in the construction industry. In Proceedings of the 40th International Symposium on Automation and Robotics in Construction, Chennai, India, 3–9 July 2023. [CrossRef]
45. *ISO/IEC ISO/IEC 27001:2022*; Information Security, Cybersecurity and Privacy Protection—Information Security Management Systems—Requirements. ISO/IEC: Geneva, Switzerland, 2022. Available online: https://www.iso.org/standard/27001 (accessed on 30 August 2023).
46. OpenAI GPT-4 Technical Report. 2023. Available online: http://arxiv.org/abs/2303.08774 (accessed on 17 March 2024).
47. Baidu Inc. Introducing ERNIE 3.5: Baidu's Knowledge-Enhanced Foundation Model Takes a Giant Leap Forward. Baidu Research. Available online: http://research.baidu.com/Blog/index-view?id=185 (accessed on 28 November 2023).
48. Ethem, A. *Introduction to Machine Learning—Ethem Alpaydin—Google Books*; MIT Press: Cambridge, MA, USA, 2020.
49. Feature Types—Designing Machine Learning Systems with Python. Baidu Research. Available online: https://subscription.packtpub.com/book/data/9781785882951/7/ch07lvl1sec42/feature-types#:~:text=There%20are%20three%20distinct%20types,a%20type%20of%20categorical%20feature. (accessed on 8 December 2023).
50. Sharma, S.; Goyal, P.K. Fuzzy Assessment of the Risk Factors Causing Cost Overrun in the Construction Industry. *Evol. Intell.* **2022**, *15*, 2269–2281. [CrossRef]
51. Baloi, D.; Price, A.D.F. Modelling Global Risk Factors Affecting Construction Cost Performance. *Int. J. Proj. Manag.* **2003**, *21*, 261–269. [CrossRef]
52. Abd El-Karim, M.S.B.A.; Mosa El Nawawy, O.A.; Abdel-Alim, A.M. Identification and Assessment of Risk Factors Affecting Construction Projects. *HBRC J.* **2017**, *13*, 202–216. [CrossRef]
53. Chileshe, N.; Boadua Yirenkyi-Fianko, A. An Evaluation of Risk Factors Impacting Construction Projects in Ghana. *J. Eng. Des. Technol.* **2012**, *10*, 306–329. [CrossRef]
54. Hwang, B.G.; Shan, M.; Phua, H.; Chi, S. An Exploratory Analysis of Risks in Green Residential Building Construction Projects: The Case of Singapore. *Sustainability* **2017**, *9*, 1116. [CrossRef]
55. Aghaei, P.; Asadollahfardi, G.; Katabi, A. Safety Risk Assessment in Shopping Center Construction Projects Using Fuzzy Fault Tree Analysis Method. *Qual. Quant.* **2022**, *56*, 43–59. [CrossRef]
56. Bilal, M.; Oyedele, L.O.; Qadir, J.; Munir, K.; Ajayi, S.O.; Akinade, O.O.; Owolabi, H.A.; Alaka, H.A.; Pasha, M. Big Data in the construction industry: A review of present status, opportunities, and future trends. *Adv. Eng. Inform.* **2016**, *30*, 500–521. [CrossRef]
57. Udayaprasad, P.K.; Shreyas, J.; Srinidhi, N.N.; Kumar, S.M.D.; Dayananda, P.; Askar, S.S.; Abouhawwash, M. Energy Efficient Optimized Routing Technique With Distributed SDN-AI to Large Scale I-IoT Networks. *IEEE Access* **2024**, *12*, 2742–2759. [CrossRef]
58. Syed Abdul Rahman, S.A.F.; Abdul Maulud, K.N.; Wan Mohd Jaafar, W.S. BIM-GIS in Catalyzing 3D Environmental Simulation. In *Advances in Geoinformatics Technologies*; Yadava, R.N., Ujang, M.U., Eds.; Earth and Environmental Sciences Library; Springer Nature Switzerland: Cham, Switzerland, 2024; pp. 183–200. [CrossRef]
59. Arulkumar, V.; Kavin, F.; Arul Kumar, D.; Nagu, B. IoT Sensor Data Retrieval and Analysis in Cloud Environments for Enhanced Power Management. *ARASET* **2024**, *38*, 77–88. [CrossRef]
60. Wong, P.K.; Luo, H.; Wang, M.; Cheng, J.C.P. Enriched and discriminative convolutional neural network features for pedestrian re-identification and trajectory modeling. *Comput. Aided Civ. Eng.* **2022**, *37*, 573–592. [CrossRef]
61. Baek, J.; Kim, D.; Choi, B. Deep learning-based automated productivity monitoring for on-site module installation in off-site construction. *Dev. Built Environ.* **2024**, *18*, 100382. [CrossRef]
62. Zhu, J.; Wang, D.; Zhao, Y. Design of smart home environment based on wireless sensor system and artificial speech recognition. *Meas. Sens.* **2024**, *33*, 101090. [CrossRef]
63. Ma, Z.; Chen, Z.-S. Mining construction accident reports via unsupervised NLP and Accimap for systemic risk analysis. *Autom. Constr.* **2024**, *161*, 105343. [CrossRef]
64. Bawa, D. Activity Theory Approach and Geographic Information Systems Affordance for Effective Land Management and Administration Actualization. *Sci. Afr.* **2024**, *23*, e01970. [CrossRef]
65. Zheng, Q.; Ding, G.; Xie, J.; Li, Z.; Qin, S.; Wang, S.; Zhang, H.; Zhang, K. Multi-stage cyber-physical fusion methods for supporting equipment's digital twin applications. *Int. J. Adv. Manuf. Technol.* **2024**, 1–20. [CrossRef]
66. Asgarkhani, N.; Kazemi, F.; Jakubczyk-Gałczyńska, A.; Mohebi, B.; Jankowski, R. Seismic response and performance prediction of steel buckling-restrained braced frames using machine-learning methods. *Eng. Appl. Artif. Intell.* **2024**, *128*, 107388. [CrossRef]
67. Chowdhery, A.; Narang, S.; Devlin, J.; Bosma, M.; Mishra, G.; Roberts, A.; Barham, P.; Chung, H.W.; Sutton, C.; Gehrmann, S.; et al. PaLM: Scaling Language Modeling with Pathways. *J. Mach. Learn. Res.* **2022**, *24*, 1–113.
68. Thoppilan, R.; De Freitas, D.; Hall, J.; Shazeer, N.; Kulshreshtha, A.; Cheng, H.-T.; Jin, A.; Bos, T.; Baker, L.; Du, Y.; et al. LaMDA: Language Models for Dialog Applications. *arXiv* **2022**, arXiv:2201.08239.
69. Goodfellow, I.; Bengio, Y.; Courville, A. *Deep Learning*; MIT Press: Cambridge, MA, USA, 2016.
70. Quinlan, J.R. Induction of decision trees. *Mach. Learn* **1986**, *1*, 81–106. [CrossRef]
71. Lowd, D.; Domingos, P. Naive Bayes Models for Probability Estimation. In Proceedings of the 22nd International Conference on Machine Learning—ICML '05, Bonn, Germany, 7–11 August 2005; ACM Press: Bonn, Germany, 2005; pp. 529–536. [CrossRef]

72. Paltrinieri, N.; Comfort, L.; Reniers, G. Learning about risk: Machine learning for risk assessment. *Saf. Sci.* **2019**, *118*, 475–486. [CrossRef]
73. Mohammadfam, I.; Soltanzadeh, A.; Moghimbeigi, A.; Alizadeh Savareh, B. Use of Artificial Neural Networks (ANNs) for the Analysis and Modeling of Factors That Affect Occupational Injuries in Large Construction Industries. *Electron Physician* **2015**, *7*, 1515–1522. [CrossRef]
74. Goh, Y.M.; Chua, D. Neural network analysis of construction safety management systems: A case study in Singapore. *Constr. Manag. Econ.* **2013**, *31*, 460–470. [CrossRef]
75. Odeyinka, H.A.; Lowe, J.; Kaka, A.P. Artificial neural network cost flow risk assessment model. *Constr. Manag. Econ.* **2013**, *31*, 423–439. [CrossRef]
76. Nair, V.; Hinton, G.E. Rectified Linear Units Improve Restricted Boltzmann Machines. In Proceedings of the ICML 2010—Proceedings, 27th International Conference on Machine Learning, Haifa, Israel, 21–24 June 2010.
77. Maas, A.L.; Hannun, A.Y.; Ng, A.Y. Rectifier Nonlinearities Improve Neural Network Acoustic Models. In Proceedings of the in ICML Workshop on Deep Learning for Audio, Speech and Language Processing, Atlanta, GA, USA, 16–21 June 2013.
78. Wojtas, M.; Chen, K. Feature Importance Ranking for Deep Learning. *arXiv* **2020**, arXiv:2010.08973.
79. Roy, D.; Murty, K.S.R.; Mohan, C.K. Feature selection using Deep Neural Networks. In Proceedings of the 2015 International Joint Conference on Neural Networks (IJCNN), Killarney, Ireland, 12–17 July 2015; IEEE: Killarney, Ireland, 2015; pp. 1–6. [CrossRef]
80. Lundberg, S.M.; Lee, S.I.; Lundberg, S.M.; Lee, S.I. A unified approach to interpreting model predictions. In *Advances in Neural Information Processing Systems 30 (NIPS 2017)*; Neural Information Processing Systems Foundation, Inc.: South Lake Tahoe, NV, USA, 2017; Volume 32, pp. 4765–4774.
81. Lin, K.; Gao, Y. Model interpretability of financial fraud detection by group SHAP. *Expert Syst. Appl.* **2022**, *210*, 118354. [CrossRef]
82. Wen, X.; Xie, Y.; Wu, L.; Jiang, L. Quantifying and comparing the effects of key risk factors on various types of roadway segment crashes with LightGBM and SHAP. *Accid. Anal. Prev.* **2021**, *159*, 106261. [CrossRef] [PubMed]
83. Bussmann, N.; Giudici, P.; Marinelli, D.; Papenbrock, J. Explainable Machine Learning in Credit Risk Management. *Comput. Econ.* **2021**, *57*, 203–216. [CrossRef]
84. Futagami, K.; Fukazawa, Y.; Kapoor, N.; Kito, T. Pairwise acquisition prediction with SHAP value interpretation. *J. Financ. Data Sci.* **2021**, *7*, 22–44. [CrossRef]
85. Cheng, H.-T.; Koc, L.; Harmsen, J.; Shaked, T.; Chandra, T.; Aradhye, H.; Anderson, G.; Corrado, G.; Chai, W.; Ispir, M.; et al. Wide & Deep Learning for Recommender Systems. In Proceedings of the 1st Workshop on Deep Learning for Recommender Systems, Boston, MA, USA, 15 September 2016. [CrossRef]
86. Venkatesan, R.; Er, M.J. A novel progressive learning technique for multi-class classification. *Neurocomputing* **2016**, *207*, 310–321. [CrossRef]

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.