Quantifying the Financial Impact of Cyber Incidents: A Machine Learning Approach to Inform Legal Standards and Risk Management

Reza Alamsyah[1,*], Sri Wahyuni[2]

[1]Information Systems, Faculty of Engineering and Computer Science, Dharmawangsa University, Indonesia

[2]Institute Informatics and Business Darmajaya, Indonesia, Bandar Lampung, Indonesia

ABSTRACT

The escalating frequency and sophistication of cyber incidents present a significant challenge for organizations, insurers, and legal systems, which often struggle to quantify financial risk and establish clear standards of liability. Traditional risk assessments are frequently subjective and lack the empirical rigor needed to connect specific incident characteristics to financial outcomes. This research addresses this gap by developing a machine learning model to predict the financial impact of security breaches and identify the key drivers of cost. Using a dataset of 5,000 incidents enriched with threat, asset, and organizational data, this study employed two ensemble models, a Random Forest Regressor and XGBoost, to perform a regression analysis. The results demonstrate that a predictive model can successfully account for a significant portion of the variance in breach costs. The Random Forest model emerged as the superior performer, explaining approximately 49.3% of the variance ($R^2 = 0.4932$) in financial impact on unseen test data, with a Mean Absolute Error of $174.89k. The feature importance analysis yielded a clear and powerful insight: the volume of data breached is the single most dominant predictor of financial loss, with an importance score (~0.83) that far exceeds all other variables, including threat type, asset vulnerability, and incident resolution time. This finding has profound implications, suggesting that legal and regulatory standards of 'due care' should prioritize controls aimed at data minimization and the prevention of large-scale data exfiltration. The study provides a quantitative framework to help courts assess damages more empirically, allows insurers to refine underwriting criteria based on data exposure risk, and guides organizations to focus cybersecurity investments on protecting their most valuable data assets at scale.

Keywords Cyber Risk, Cybersecurity, Data Breach, Financial Impact, Machine Learning

Submitted 30 July 2025
Accepted 16 August 2025
Published 1 September 2025

*Corresponding author
Reza Alamsyah,
89rezaalamsyah@gmail.com

Additional Information and Declarations can be found on page 279

DOI: 10.63913/jcl.v1i3.15

Introduction

The escalating challenge of quantifying cyber risk in a complex legal and technological environment has gained considerable prominence in both academic discourse and practical applications. Cyber incidents result in increasing financial consequences driven by operational disruptions, regulatory penalties, and litigation, contributing to the pressing necessity for effective risk management frameworks. A leap in understanding is vital, as a significant gap exists between the inherently technical nature of cyber threats and the qualitative frameworks often employed in legal and insurance assessments, potentially leading to misalignment in risk expectations and evaluations.

One of the pressing issues in quantifying cyber risk lies in the difficulties faced by organizations in evaluating the multifaceted dimensions of cyber incidents.

How to cite this article: R. Alamsyah and S. Wahyuni, "Quantifying the Financial Impact of Cyber Incidents: A Machine Learning Approach to Inform Legal Standards and Risk Management," *J. Cyber. Law.*, vol. 1, no. 3, pp. 264-281, 2025.

Lois et al discuss the intricate relationship between internal auditing and cybersecurity, emphasizing the need for auditors to incorporate technological safety measures aligned with evolving cyber threats in their evaluations [1]. This complexity necessitates an expansion of traditional audit procedures to address modern technological risks, reflecting that a real-time dynamic approach is essential for assessing the cyber landscape accurately. Similarly, Radanliev et al advocate for predictive analytics supported by artificial intelligence and machine learning to provide real-time assessments of cyber risk [2]. Without such frameworks, organizations may rely on outdated methodologies that fail to capture the fluid nature of cyber threats, thereby increasing vulnerability to incidents.

Furthermore, the integration of dynamic supply chain management is fundamental in assessing the pathways and consequences of cyberattacks. Cyberattacks can propagate through interconnected systems, necessitating the mapping of attack paths for effective risk management [3]. This perspective is echoed by Crosignani et al, who highlight the necessity for organizations to improve their risk management strategies to combat potential supply chain disruptions arising from cyber incidents [4]. A robust understanding of how cyber vulnerabilities can emerge from interconnected supply chains is crucial, as firms must develop contingency planning to ensure operational continuity amidst cyber threats, underscoring the importance of knowledge-sharing and collaboration across industries.

The evolving complexity of cyber incidents further suggests that the legal ramifications of such risks can result in severe financial consequences for organizations. Crosignani et al reinforce the need for improving resilience across supply chains, thereby aiding in the mitigation of operational risk as firms navigate the regulatory and litigation landscape that follows any cyber incident [4]. Similarly, research has established a potential correlation between cyber risk materialization and diminished organizational reputation, which can lead to substantial financial pitfalls [5]. Thus, the quantification of cyber risks becomes intertwined not only with direct operational costs but also with the broader implications for market perceptions and stakeholder trust.

Adding to this discourse is the advent of cyber insurance as a tool for managing the financial ramifications of cyber incidents. The methodologies surrounding cyber insurance are increasingly recognized as critical for enterprises seeking to align their risk management strategies with the realities of cyber threats [6]. The relevance of this approach is underscored by the fact that organizations must navigate not only the technical threats but also the legal interpretations that arise from their cybersecurity measures—or lack thereof. Therefore, the role of cyber insurance becomes crucial as it not only provides financial backing in the wake of incidents but also incentivizes the adoption of better security measures to reduce policy premiums [7].

Effective cyber risk quantification also involves understanding the psychosocial dimensions tied to technological resilience. For example, the empirical survey by Ogbeide et al suggests that small and medium-sized enterprises (SMEs) in Nigeria lack sufficient processes to counteract evolving cybersecurity threats adequately [8]. The consequences of inadequate cyber risk management often manifest as litigation costs and loss of customer trust and revenue, highlighting that the repercussions of cyber incidents extend beyond immediate financial ramifications. As a result, frameworks must be enhanced to encapsulate the

holistic view of cyber risks, combining both quantitative and qualitative insights across various dimensions of organizational impact.

Moreover, emerging frameworks such as those proposed by Dupont emphasize that cyber resilience is not merely about risk avoidance but also encompasses the capacities to withstand and recover from cyber-attacks [9]. By fostering an integrated approach that marries the technical aspects of cybersecurity with legal compliance and operational resilience, organizations can position themselves more effectively against the multifaceted challenges posed by cyber threats. This comprehensive understanding of operational, legal, and reputational risks highlights the necessity for insurers and auditors to adopt a unified framework that captures the breadth of cyber risk.

The necessity for empirical, data-driven models to define liability and due care within cybersecurity risk management has been accentuated by the growing financial implications of cyber incidents. Traditional risk assessments often suffer from subjectivity and a lack of rigorous quantitative methods, thereby undermining the establishment of robust, defensible standards. Empirical frameworks are essential not only for compliance with regulations but also for ensuring that organizations can clearly articulate their cybersecurity posture and readiness in the face of evolving threats. The imperative for objective methodologies that effectively identify the factors contributing to financial loss due to cyber incidents is crucial in today's complex digital landscape.

A foundational reference for understanding the implementation of quantitative risk analysis in cybersecurity is the work of Bentley et al, who propose a multivariate model that quantifies and mitigates cybersecurity risk. While they caution against absolute reliance on quantitative metrics, they highlight that utilizing these models can significantly enhance the transparency of risk communication within organizations [10]. By requiring stakeholders to engage with clear parameters regarding potential threats, damages, and the efficacy of different mitigations, organizations are better equipped to navigate the intricate dynamics of cyber risk. The model effectively depicts risk in numerical terms and translates these into more comprehensible language for broader stakeholder engagement.

Sheehan et al further contribute to the discourse with their development of a bow-tie risk classification framework, which integrates both qualitative and quantitative aspects of cyber risk. They highlight that barriers to effective cyber insurance market development include a lack of standardized measurements and insufficient claims data, complicating risk quantification [11]. This gap illustrates the necessity for a standardized, empirical approach to classify risks and inform due diligence. Their framework offers a structured mechanism for assessing vulnerabilities systematically, thus aligning risk management activities with operational realities.

In terms of specific methodologies, Dawodu et al underscore the importance of an integrated approach to cybersecurity risk assessment in banking, advocating for advanced technologies to enhance risk assessments [12]. By leveraging technology, institutions are better equipped to confront the dynamic threat landscape and maintain stakeholder trust. Their emphasis on technological integration highlights the need for organizations to continuously evolve their risk management strategies to incorporate the latest insights and innovations in cybersecurity.

Importantly, advances in risk assessment techniques must incorporate methods like Monte Carlo simulations, as discussed by Shete. This approach provides a quantifiable, financial perspective of cybersecurity risks, aiding organizations in making informed decisions regarding resource allocation and strategic planning [13]. Such models can simulate various scenarios to project potential breaches, thereby enabling decision-makers to visualize the economic implications of their risk management strategies.

The urgency for empirical models is mirrored by the escalating costs of cyber incidents and the ensuing need for effective resource allocation. For example, Fagade et al illustrate how utilizing Monte Carlo predictive modeling can improve resource allocation decisions, thus mitigating the risks associated with both over- and under-resourcing cybersecurity capabilities [14]. This underscores the critical relationship between effective risk assessment and financial implications, showing that poorly directed investments in cybersecurity can lead to substantial losses.

The primary objective of this paper is to develop and evaluate a machine learning model capable of predicting the financial impact of a security breach. To achieve this, the research integrates a diverse set of incident, threat, asset, and organizational data to build a comprehensive analytical framework. The goal is to move beyond traditional, often qualitative, risk assessments by creating a quantitative tool that can generate an empirical forecast of potential losses, providing a more objective basis for risk management and strategic decision-making.

Beyond simple prediction, this study aims to identify and rank the key factors that most significantly contribute to financial loss following a cyber incident. A specific focus is placed on determining the relative importance of the volume of data compromised versus other contributing factors, such as threat type or asset vulnerability. By deconstructing the model to understand its key drivers, the research seeks to provide actionable insights that can help organizations prioritize their security investments and inform the development of more effective legal and regulatory standards.

Literature Review

Prior Approaches to Modeling Data Breach Costs and Cyber Risk

The modeling of data breach costs and the assessment of cyber risk have garnered significant attention within both the academic and insurance communities, especially as the consequences of cyber incidents continue to escalate. Established statistical and actuarial models utilized in the cyber insurance industry serve dual purposes: they price policies and estimate potential losses associated with cyber incidents. A comprehensive review of the literature reveals a variety of approaches deployed to address the complex nature of cyber risk and its financial implications, as well as efforts to correlate specific incident characteristics with resulting financial outcomes.

One prominent area of research relates to the use of generalized linear models (GLMs) for data breach incidents, as illustrated by Sun and Lu [15]. Their study develops a Bayesian generalized linear mixed model that analyzes data breaches chronologically, beginning from 2001. This model effectively captures the interdependence between the frequency and severity of losses due to cyber attacks. The implications of such models are critical for cyber insurers, as they

provide insights into estimating potential losses based on historical incident data. By employing a Bayesian framework, the study allows for a more nuanced understanding of incident dynamics, thus facilitating refined policy pricing.

Building on this statistical foundation, Pal et al explore the limitations of traditional correlative risk models in their research on cyber-insured IT firms [16]. They extend the dimensional copula density approach to account for nonlinear correlations between risk variables affecting cyber-risk quantification. The findings indicate that inadequate modeling can lead to loose estimates of correlated IT risks, ultimately affecting the profitability of coverage policies. Their work underscores the necessity for robust statistical methodologies to avoid significant mispricing within the cyber insurance market.

The complexity of pricing cyber insurance is further exacerbated by the nature of cyber incidents themselves. Historical data has shown that breaches can vary significantly based on industry sector and breach type, as detailed in the works of Granato and Polácek [17]. They highlight that high-profile incidents, such as the 2017 Equifax data breach, have had profound financial repercussions, emphasizing the urgent need for comprehensive actuarial analyses that can categorize and quantify these risks effectively. Their review articulates the upward trajectory of attack frequency and cost, directly correlating with the increasing demand for cyber insurance.

Another critical aspect is integrating data breach characteristics with financial outcomes, which Romanosky and Sayers examine in their inquiry into enterprise risk management [18]. They scrutinize how firms incorporate cyber risk into their broader risk management strategies, reinforcing that the financial ramifications of significant data breaches are now monitored at the boardroom level. This heightened visibility reflects a corporate recognition that cyber risks are paramount, thus necessitating models that can articulate their financial impact comprehensively.

The increasing complexity of cyber risks also requires that insurers grapple with systemic issues. As pointed out by Awiszus et al [19]. classical actuarial methods work for idiosyncratic and systematic cyber risks, systemic risks necessitate advanced approaches that account for interdependencies among networks and technological platforms. Their findings advocate for more sophisticated modeling that captures both strategic interactions across organizations and the broader network of dependencies inherent in our digital ecosystem.

This complexity is echoed in the work of Mamanazarov [20], who identifies limitations stemming from historical data deficiencies and opaque controls that hamper effective risk modeling. He emphasizes that these challenges can lead to ambiguities in claims processing and insurance pricing, which impede insurers' ability to develop robust coverage and premium structures. The study calls for enhanced risk management practices aligned with data transparency to allow for a clearer understanding of potential losses.

In this milieu, empirical studies have also revealed significant insights into how cyber insurers can behave as regulatory partners for businesses, as analyzed by Talesh [21]. This emerging narrative posits that insurance companies often play a proactive role in shaping compliance behaviors among firms, indirectly influencing risk mitigation strategies. By offering cyber risk assessment services and compliance guidelines, insurers help organizations navigate not only their

risk appetites but also their regulatory obligations.

A noteworthy contribution to the integration of cyber risk into corporate risk frameworks is highlighted by Peters et al [22]. They discuss model risk, focusing on how discrepancies between model assumptions and real-world outcomes can lead to mispriced insurance premiums. By addressing these modeling uncertainties and their implications for pricing strategies, the study reinforces the necessity of developing a robust framework for quantifying cyber risks accurately.

The Legal Framework for Cybersecurity and Data Breach Liability

The legal framework governing cybersecurity and data breach liability has evolved significantly over recent years, driven by the rise of data protection regulations such as the General Data Protection Regulation (GDPR) and the California Consumer Privacy Act (CCPA). These regulations establish essential legal requirements for data protection while imposing strict penalties for non-compliance, thereby shaping the landscape of corporate responsibility regarding cybersecurity. Understanding these legal frameworks is crucial for organizations operating in increasingly complex digital environments, as these laws dictate how personal data must be managed and the ramifications of failing to protect such information.

The GDPR, which came into force in May 2018, is considered one of the most comprehensive legislative measures regarding data protection. It emphasizes that personal data must be processed lawfully, transparently, and for specified purposes Chiara (2022). This regulation not only mandates strict compliance requirements within the EU but also has extraterritorial implications—meaning that organizations outside Europe must also comply if they handle data of EU residents (Zhao & Chen, 2019). Zhao and Chen discuss the GDPR's strict personal data protection mechanism and the challenges regarding its applicability in non-EU jurisdictions, particularly in regions like China [23]. Thus, organizations must navigate these international complexities to avoid substantial penalties.

The GDPR outlines significant penalties for non-compliance. Organizations risk fines up to €20 million or 4% of their total global annual turnover, whichever is higher [24]. Such substantial penalties underscore the importance of understanding and implementing the regulatory measures dictated by the GDPR. Furthermore, the CCPA, which became effective on January 1, 2020, complements the GDPR by granting California residents robust rights concerning their personal information. The CCPA allows consumers to know about and control the personal data businesses collect and share, with enforceable penalties for violations, further establishing a serious landscape for corporate negligence in data protection [24].

Corporate liability in the realm of cybersecurity is elucidated through significant case law that defines "reasonable security." Notable instances include the Equifax breach, where a lack of adequate security measures led to a breach affecting over 147 million customers, resulting in both public outcry and substantial financial penalties for failing to uphold reasonable security standards [25]. The courts have consistently interpreted reasonable security as encompassing a range of security practices tailored to the context of the organization's operations and the types of data it handles. As noted by Elendu et al, evolving case law underscores that businesses must adopt robust

cybersecurity measures that comply with regulatory frameworks and are effectively communicated to stakeholders [25].

Analysis of significant rulings reveals trends in judicial interpretations of corporate negligence following data breaches. The courtroom environment has increasingly held organizations accountable for insufficient data protection measures. In cases such as Yahoo and Target, courts scrutinized whether these organizations took necessary precautions to safeguard customer data, ultimately deciding in favor of plaintiffs who argued that the companies' negligence directly resulted in harm [25]. Such legal precedents provide a basis for determining what constitutes an acceptable standard of care and associated liabilities in the event of a data breach.

The interplay between regulations and case law outlines a comprehensive framework whereby organizations must conform to established security practices. Lindén et al discuss how the evolution of privacy policy landscapes post-GDPR is reshaping expectations for corporate responsibility concerning customer data protection [26]. The GDPR embodies a "privacy by design" approach—obligating organizations to implement strong data protection measures throughout the data lifecycle. Such requirements necessitate organizational commitment at all management levels, as noncompliance could lead to costly legal repercussions and reputational damage.

To further assist organizations, appropriate response strategies include establishing incident response plans and conducting regular security assessments, ensuring that preparations align with both legal obligations and industry best practices. Furthermore, the increasing reliance on cybersecurity insurance highlights the evolving risk assessment associated with data breaches. As demonstrated by Saqib et al, compliance with various regulatory frameworks requires organizations to map their security requirements according to established data protection laws [27]. This mapping facilitates the identification of potential compliance difficulties, assisting them in developing tailored security frameworks to effectively minimize risk exposure.

The Role of Technical and Human Factors in Incident Severity

The severity of cybersecurity incidents is influenced by both technical and human factors, and numerous studies illuminate the complex relationship between the sophistication of threat actors, system vulnerabilities, and the efficacy of security controls. Analyzing these aspects provides valuable insights into how organizations can mitigate risks and enhance their cybersecurity posture to protect against an increasingly sophisticated threat landscape.

One key technical factor affecting incident severity is the sophistication of threat actors. As indicated by the IBM X-Force Threat Intelligence Index, a significant proportion of cybersecurity incidents stem from fundamental oversights, such as misconfigurations and the use of weak passwords [28]. This underscores the need for organizations to adopt comprehensive security frameworks capable of addressing various attack vectors and ensuring that security protocols are continually adapted to the evolving threat landscape. Additionally, the prevalence of reentrancy vulnerabilities in smart contracts serves as a demonstration of how technical weaknesses can be exploited by adept attackers, further underscoring that both the technical depth of a system and its vulnerabilities can dictate the potential severity of an incident [29].

System vulnerabilities, often inherited from legacy infrastructure, also play a

pivotal role in determining incident severity. Jevtić and Alhudaidi emphasize that information security policies must be deeply embedded within organizational culture to effectively address vulnerabilities and enhance safety [30]. When vulnerabilities remain unmitigated, they provide attackers with opportunities to orchestrate severe cyber incidents, making it imperative for organizations to conduct regular vulnerability assessments and proactively address any weaknesses.

The effectiveness of specific security controls is another significant consideration in the context of incident severity. Research indicates that implementing robust incident response protocols and tailored security measures can significantly reduce the likelihood of exploit success and minimize the damages incurred from incidents [30]. For instance, employing security information and event management (SIEM) systems can enhance an organization's situational awareness regarding potential threats, enabling timely responses that can reduce incident severity [31]. Furthermore, organizations employing comprehensive cybersecurity strategies can quantify the risk tied to various vulnerabilities quantitatively, thus enabling preemptive actions to minimize the potential impact of threats.

However, the human element cannot be understated in its contribution to incident severity. Employee behaviors and adherence to security policies significantly influence cybersecurity outcomes. Liu et al elucidate that employees' perceptions of risk severity and vulnerability play a critical role in shaping their adherence to security protocols [32]. When employees receive comprehensive training that addresses their understanding of security threats, they are more likely to engage in protective behaviors, thus reducing the potential for security incidents.

Training programs that enhance security awareness have proven effective in promoting compliance among employees. He et al conclude that tailored, evidence-based training on malware and cyber threats can empower employees to take cybersecurity more seriously and heighten their engagement in organizational security measures [33]. The cultural integration of cybersecurity awareness within organizations fosters a proactive attitude toward incident prevention, aligning employee mindsets with corporate security objectives.

Insider threats, comprising unintentional risks from negligent employees or deliberate malicious actions, represent another significant challenge within the cybersecurity paradigm. As noted by Chu and So, unethical employee behavior can severely undermine organizational information security, emphasizing the need for companies to foster an ethical climate and cultivate a culture of reporting security incidents [34]. The interplay between human behavior and technical controls highlights that cyber hygiene must be treated holistically, where the effectiveness of technical safeguards hinges on employee engagement and compliance.

Method

Research Design and Data Integration

To quantify the financial impact of cyber incidents, this study employed a quantitative, predictive modeling design, chosen for its ability to produce objective, replicable, and data-driven insights that contrast with traditional qualitative risk assessments. The core of the methodology involved developing

and evaluating supervised machine learning models to perform a regression analysis, as the target variable—financial impact—is a continuous value. The approach began with the integration of four distinct datasets, which collectively detailed 5,000 security incidents and their associated threats, targeted assets, and involved employees. The incidents dataset served as the foundational log, containing records of each breach, its detection time, and resolution duration. The threats, assets, and employees datasets provided rich contextual information, including threat actor types, asset vulnerability scores, and employee security training status.

These disparate sources were systematically merged into a single, comprehensive analytical file to ensure that each incident record contained a holistic set of features for modeling. This integration was executed using a series of left joins, a strategy that preserves every record from the primary incidents table, thereby ensuring that no incident was lost from the analysis, even in cases of incomplete contextual data from the other tables. The data was linked on their respective primary keys (threat_id, asset_id, employee_id), resulting in a wide-format dataset where each row represented a single incident enriched with a full spectrum of explanatory variables.

Feature Engineering and Preprocessing

Following data integration, a series of feature engineering and preprocessing steps were executed to prepare the data for analysis. The primary target variable was defined as financial_impact_k, representing the total financial loss in thousands of U.S. dollars. To capture temporal patterns that might influence incident cost—such as attacker behavior during off-hours or the impact of staff availability on response times—the detection_time timestamp was decomposed into several numerical features: detection_hour, detection_day_of_week, and detection_month. This transformation converts a cyclical, non-linear feature into a format that regression models can more easily interpret.

Categorical data was transformed into a numerical format suitable for machine learning algorithms. The binary status of employee security_training ('Completed' or 'Pending') was logically encoded into a 1 or 0 format, respectively. For nominal features with no inherent order—specifically threat_type, asset_type, and employee_department—one-hot encoding was applied via the pandas get_dummies function. This technique was deliberately chosen over label encoding to prevent the models from inferring a false and misleading ordinal relationship between categories (e.g., that one department is numerically "greater" than another). This process created new binary columns for each category, expanding the feature space but ensuring accurate representation. Finally, to reduce noise and prevent model overfitting, non-predictive columns such as unique identifiers (incident_id, employee_id, etc.) and original text-based fields that had already been encoded were removed, resulting in a clean, model-ready dataset optimized for predictive performance.

Model Selection, Training, and Evaluation

For the predictive modeling phase, two powerful ensemble learning algorithms were selected: Random Forest Regressor and XGBoost. These tree-based models were chosen for their high predictive accuracy, their ability to handle complex non-linear relationships, and their inherent functionality for generating feature importance scores, which is crucial for identifying the key drivers of financial loss. The Random Forest model, which operates by constructing a

multitude of decision trees and outputting their collective average, was configured with n_estimators=100. The XGBoost model, an efficient implementation of gradient boosting, was set with an equivalent number of estimators and a reg:squarederror objective function to guide its optimization toward minimizing squared prediction errors. For reproducibility, a random_state of 42 was used throughout the modeling process.

The integrated dataset was partitioned into an 80% training set and a 20% testing set. This critical step ensures that the models' performance is validated on unseen data, providing an unbiased estimate of their ability to generalize to new, real-world incidents. The models were trained on the training data, and their predictive performance was subsequently evaluated on the test set using three standard regression metrics. Mean Absolute Error (MAE) was used to provide a direct, interpretable measure of the average prediction error in thousands of dollars. Root Mean Squared Error (RMSE) was also calculated, as it penalizes larger errors more heavily, making it particularly sensitive to high-cost outlier incidents. Finally, R-squared (R^2) was used to measure the proportion of the variance in financial impact that is predictable from the features, offering a clear assessment of the models' overall explanatory power.

Result and Discussion

Exploratory Data Analysis Results

Exploratory data analysis revealed several key trends and relationships within the dataset that informed the subsequent modeling process. A correlation matrix of the numerical features (figure 1) provided the first indication of the primary cost drivers. It showed a strong positive correlation of 0.87 between the volume of data breached (data_breached_gb) and the financial impact (financial_impact_k), highlighting this as the most significant linear relationship in the data. Other variables, such as vulnerability_score and time_to_resolve_hours, showed much weaker positive correlations with the financial outcome, suggesting they were less influential.

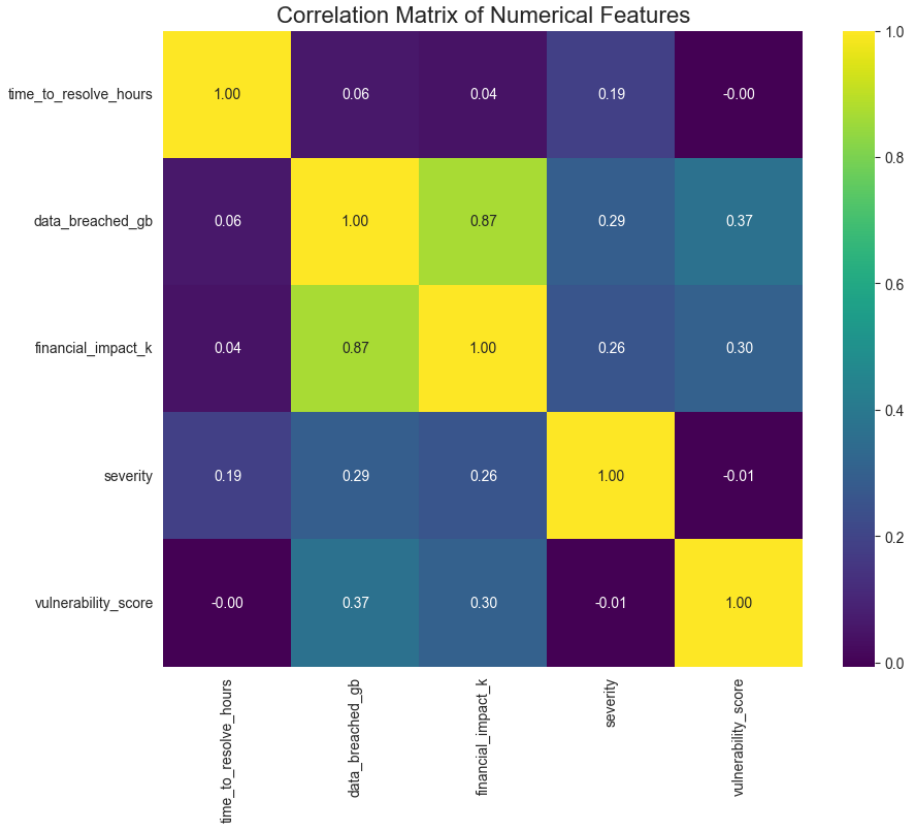

Figure 1 Correlation Matrix of Numerical Features

This strong relationship was visually confirmed by a scatter plot of financial impact versus data breached (figure 2). The plot clearly illustrates that as the amount of data compromised increases, the financial cost tends to rise accordingly. While the relationship is not perfectly linear, the upward trend is unmistakable and underscores the importance of data volume as a key variable.

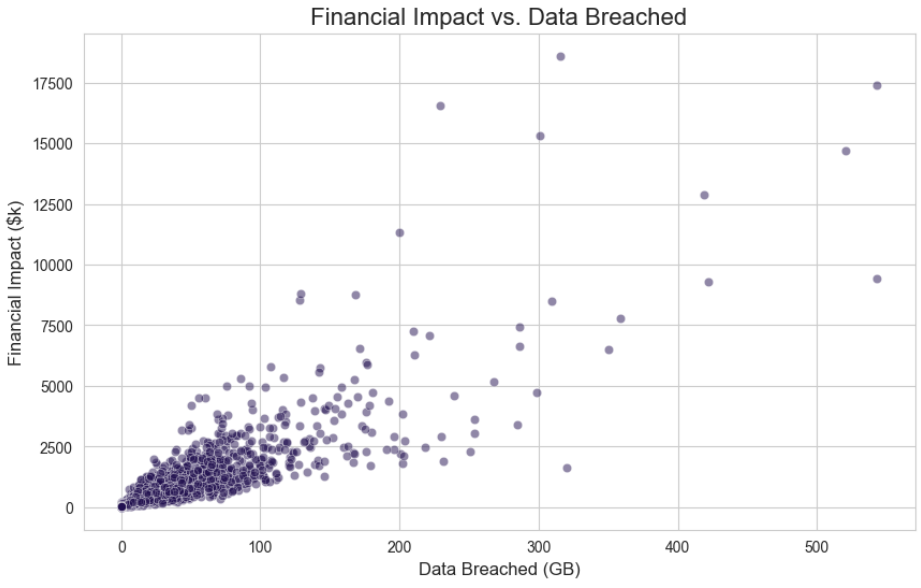

Figure 2 Financial Impact vs Data Breached Scatter Plot

Furthermore, figure 3 illustrating the total number of incidents by employee department showed that certain departments, such as Consulting Civil Engineer and Database Administrator, were associated with a higher frequency of incidents, pointing to potential areas of concentrated risk within the organization.

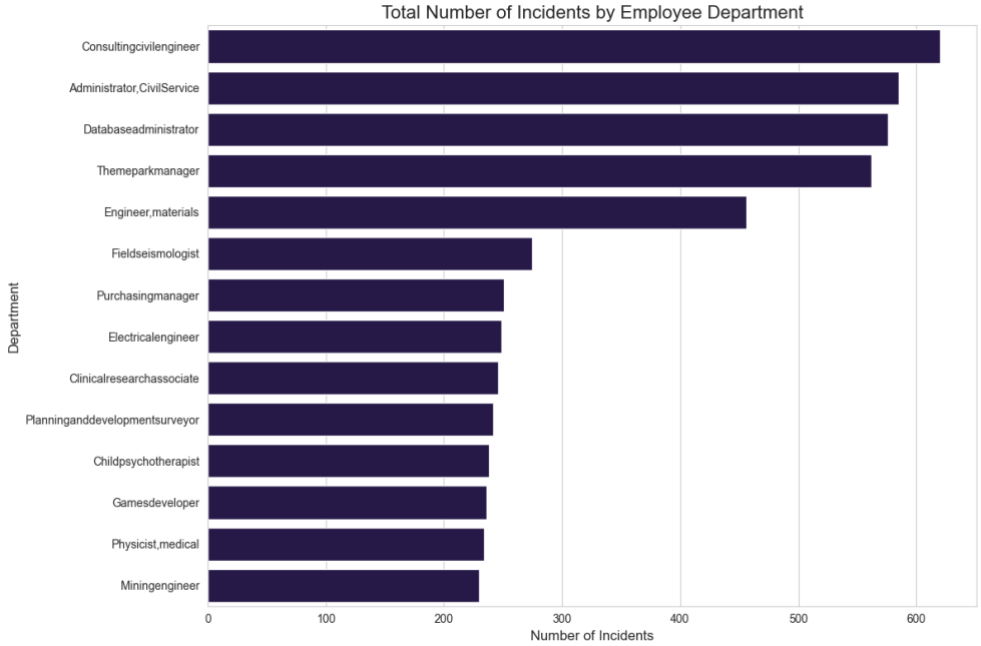

Figure 3 **Total Number of Incidents by Employee Department**

Finally, a time-series plot of incidents per month (figure 4) showed a consistent, albeit volatile, rate of occurrence over the observed period, without a clear upward or downward trend, indicating that the threat landscape was persistent and ongoing.

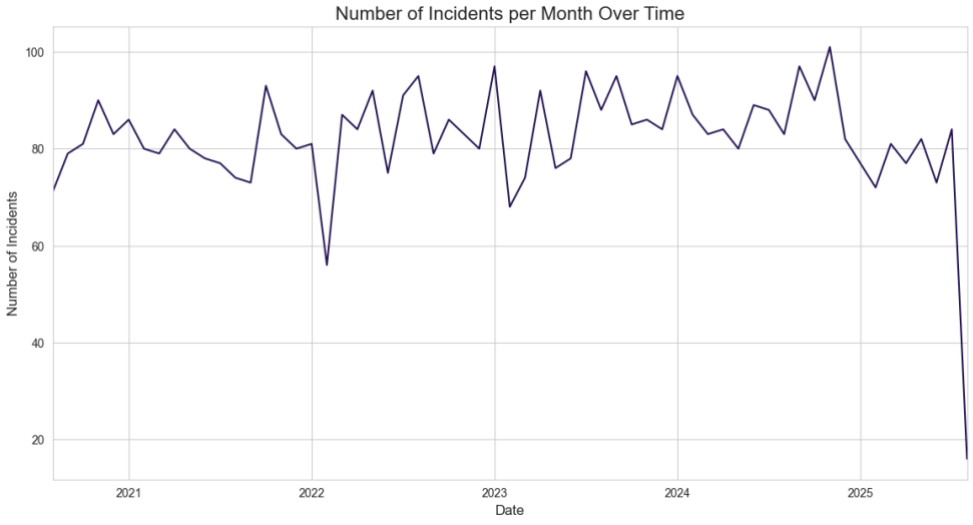

Figure 4 **Number of Incidents per Month Over Time**

Predictive Model Performance

The evaluation of the two machine learning models revealed that both were capable of predicting the financial impact of cyber incidents with a reasonable

degree of accuracy, though the Random Forest Regressor emerged as the superior performer. When evaluated on the unseen test set of 1,000 incident records, the Random Forest model achieved an R-squared (R^2) value of 0.4932. This indicates that the model was able to explain approximately 49.3% of the variance in financial impact based on the provided features. In practical terms, the model's Mean Absolute Error (MAE) was $174.89k, signifying that, on average, its predictions deviated from the actual financial cost by this amount. The XGBoost model, while also effective, demonstrated slightly lower performance, with an R^2 of 0.4625 and an MAE of $175.20k. Given its higher explanatory power and marginally lower prediction error, the Random Forest model was selected for the subsequent feature importance analysis.

Identification of Key Financial Impact Drivers

The feature importance analysis, conducted using the trained Random Forest model, as shown in figure 5, produced a clear and decisive primary finding: the volume of data breached (data_breached_gb) is the single most dominant predictor of financial loss. This variable registered an importance score of approximately 0.83, making it exponentially more influential than any other factor in the model. This empirical result provides strong quantitative evidence that the scale of data exfiltration is the principal determinant of the ultimate cost incurred from a security breach.

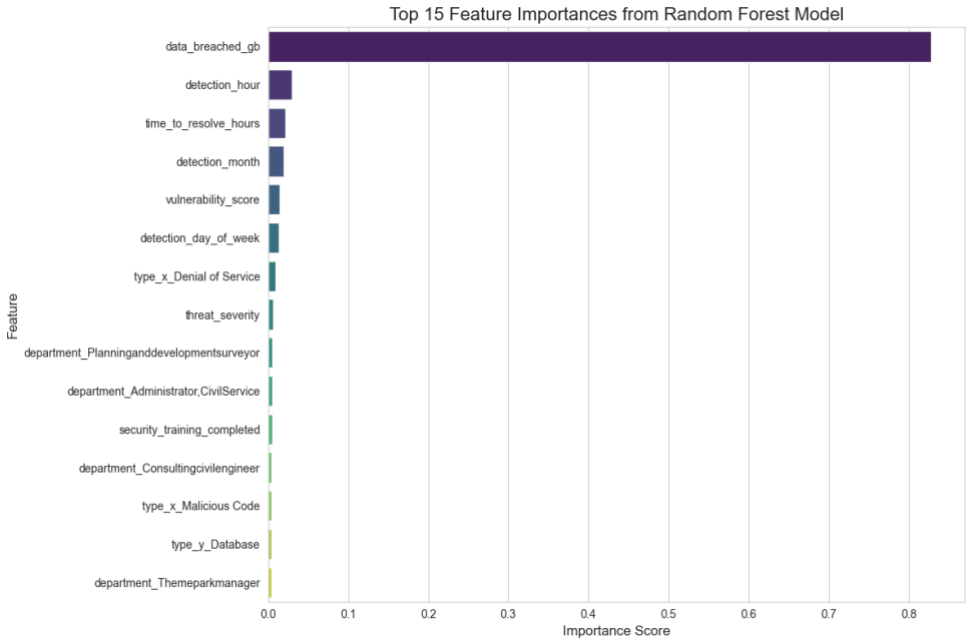

Figure 5 Top 15 Feature Importance

While data volume was the primary driver, the analysis also identified several secondary predictors of significance. Temporal factors, including the hour of incident detection (detection_hour) and the month of the incident (detection_month), emerged as the next most important variables. This suggests that the timing of a breach—perhaps correlating with periods of lower staff vigilance or slower response capabilities—has a measurable effect on its financial consequences. Other notable factors included the time required for resolution (time_to_resolve_hours) and the inherent vulnerability of the targeted asset (vulnerability_score), reinforcing the intuitive understanding that longer-

lasting incidents and attacks on weaker assets lead to higher costs. In contrast, factors such as the completion of security training or the specific type of threat, while still relevant, had a comparatively minor influence on the overall financial outcome.

Comparison with Previous Research

These findings both align with and diverge from previous research in the field of cyber risk quantification. Many prior studies, often relying on statistical analysis of survey data, have identified factors such as industry sector, company size, and threat type (e.g., insider vs. external) as significant cost drivers. While our model confirms the relevance of some of these variables (such as threat type), its primary contribution is the quantification of data volume's overwhelming importance. The finding that data_breached_gb outweighs all other factors so significantly provides a more granular and actionable insight than broader categorical analyses. It suggests that while what kind of attack occurs is relevant, the ultimate financial impact is far more dependent on how much data is compromised, a conclusion that has not been as empirically established in prior literature.

Discussion of Legal and Policy Implications

The empirical findings from this research carry significant implications for legal standards, corporate risk management, and cyber insurance underwriting. The unequivocal importance of data volume as the primary cost driver suggests that the legal standard of 'due care' in cybersecurity should be heavily weighted toward controls that specifically address data minimization, access control, and the prevention of large-scale data exfiltration. Rather than a checklist approach to security, this finding advocates for a risk-based framework where the quantity and sensitivity of data are central to defining what constitutes 'reasonable security'.

For legal proceedings, the model provides a quantitative framework that could be used by courts and expert witnesses to assess damages more empirically. Instead of relying on qualitative assessments, this data-driven approach can help establish a more objective baseline for financial liability based on the specific characteristics of a breach, particularly the amount of data compromised. Similarly, for the cyber insurance industry, these results can inform more precise underwriting criteria. Insurers could refine policy pricing based on an organization's data exposure risk, placing a greater emphasis on the volume of sensitive data it holds and the maturity of its data loss prevention (DLP) controls. Ultimately, this research provides a data-driven foundation for shifting the focus of cybersecurity strategy from a broad threat-based perspective to a more targeted, impact-oriented approach centered on protecting large volumes of data.

Limitations of the Study

Despite the robustness of the model, several limitations must be acknowledged. First, the model's R-squared value of 0.4932 indicates that approximately half of the variance in financial impact remains unexplained by the features included in this study. This suggests the influence of variables not captured in the available datasets. Such factors could include the specific security controls in place at the time of the incident, the effectiveness of the organization's public relations and crisis management response, pre-existing brand reputation, and

the specific regulatory environment. Second, the data is derived from a specific set of 5,000 incidents, and while comprehensive, it may not be fully representative of all industries or geographic regions, potentially limiting the generalizability of the findings. Finally, the financial impact data itself may not capture all long-term or intangible costs, such as sustained reputational damage or loss of customer trust.

Directions for Future Research

The limitations of this study present clear directions for future research. To improve the model's predictive power, subsequent studies should aim to incorporate additional variables. Integrating data on the specific security controls and technologies deployed by an organization could reveal which defenses are most effective at mitigating the financial impact of a breach. Furthermore, incorporating natural language processing (NLP) to analyze unstructured data from incident response reports could uncover nuanced operational details that are currently missed. Future research could also test the model's validity across different industry sectors (e.g., finance, healthcare, retail) to determine if the key cost drivers vary in different contexts. Finally, longitudinal studies that track the financial impact on organizations over several years post-breach would provide a more complete picture of the long-tail costs associated with cyber incidents.

Conclusion

This research successfully demonstrated that the financial impact of cybersecurity incidents can be quantitatively modeled using machine learning techniques, providing a data-driven alternative to traditional qualitative risk assessments. By integrating diverse datasets and applying ensemble regression models, this study empirically established that the amount of data compromised is the most critical factor determining the ultimate financial cost of a breach. The findings revealed that the volume of data exfiltrated supersedes other technical and organizational variables, such as threat type or employee training, in its predictive power. This highlights a crucial focal point for risk mitigation: the scale of a data breach is a more significant cost determinant than the specific tactics used by an attacker. The primary contribution of this work is the provision of an objective, empirical foundation for legal and corporate stakeholders to better understand and manage cyber risk. By quantifying the direct relationship between data volume and financial loss, this study offers a clear directive for shaping legal standards of 'due care', refining cyber insurance underwriting, and prioritizing corporate security investments. The focus is shifted from a broad, perimeter-based defense posture towards a more targeted, data-centric strategy. This research provides a quantitative tool to help bridge the gap between technical security measures and their tangible financial consequences, enabling organizations to align their cybersecurity strategies more closely with their financial risk exposure.

Declarations

Author Contributions

Conceptualization: R.A.; Methodology: S.W.; Software: S.W.; Validation: S.W.; Formal Analysis: R.A.; Investigation: S.W.; Resources: R.A.; Data Curation: S.W.; Writing Original Draft Preparation: R.A.; Writing Review and Editing: R.A.;

Visualization: S.W.; All authors have read and agreed to the published version of the manuscript.

Data Availability Statement

The data presented in this study are available on request from the corresponding author.

Funding

The authors received no financial support for the research, authorship, and/or publication of this article.

Institutional Review Board Statement

Not applicable.

Informed Consent Statement

Not applicable.

Declaration of Competing Interest

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

References

[1] P. Lois, G. Drogalas, A. Karagiorgos, A. Thrassou, and D. Vrontis, “Internal Auditing and Cyber Security: Audit Role and Procedural Contribution,” *Int. J. Manag. Financ. Account.*, 2021, doi: 10.1504/ijmfa.2021.116207.

[2] P. Radanliev *et al.*, “Design of a Dynamic and Self-Adapting System, Supported With Artificial Intelligence, Machine Learning and Real-Time Intelligence for Predictive Cyber Risk Analytics in Extreme Environments – Cyber Risk in the Colonisation of Mars,” *SSRN Electron. J.*, 2021, doi: 10.2139/ssrn.3562925.

[3] N. Polatidis, M. Pavlidis, and H. Mouratidis, “Cyber-Attack Path Discovery in a Dynamic Supply Chain Maritime Risk Management System,” *Comput. Stand. Interfaces*, 2018, doi: 10.1016/j.csi.2017.09.006.

[4] M. Crosignani, M. Macchiavelli, and A. F. Silva, “Pirates Without Borders: The Propagation of Cyberattacks Through Firms’ Supply Chains,” *J. Financ. Econ.*, 2023, doi: 10.1016/j.jfineco.2022.12.002.

[5] B. Balawejder, R. Dankiewicz, and A. Ostrowska-Dankiewicz, “The Role of Insurance in Cyber Risk Management in Enterprises,” *Humanit. Soc. Sci. Q.*, 2019, doi: 10.7862/rz.2019.hss.33.

[6] X. Lu, D. Niyato, H. Jiang, P. Wang, and H. V. Poor, “Cyber Insurance for Heterogeneous Wireless Networks,” *Ieee Commun. Mag.*, 2018, doi: 10.1109/mcom.2018.1700504.

[7] D. Woods, “A Turning Point for Cyber Insurance,” *Commun. Acm*, 2023, doi: 10.1145/3545795.

[8] V. O. OGBEIDE, O. OMOROGIUWA, and E. E. Salami, “An Empirical Survey to Substantiate the Need for a Cyber Security Framework for Smes in Nigeria,” *Ijrp*, 2023, doi: 10.47119/ijrp1001281720235221.

[9] B. Dupont, “The Cyber-Resilience of Financial Institutions: Significance and Applicability,” *J. Cybersecurity*, 2019, doi: 10.1093/cybsec/tyz013.

[10] M. Bentley, A. Stephenson, P. Toscas, and Z. Zhu, “A Multivariate Model to Quantify and Mitigate Cybersecurity Risk,” *Risks*, 2020, doi: 10.3390/risks8020061.

[11] B. Sheehan, F. Murphy, A. N. Kia, and R. Kiely, “A Quantitative Bow-Tie Cyber Risk Classification and Assessment Framework,” *J. Risk Res.*, 2021, doi:

10.1080/13669877.2021.1900337.
[12] S. O. Dawodu, A. Omotosho, O. J. Akindote, A. O. Adegbite, and S. K. Ewuga, “Cybersecurity Risk Assessment in Banking: Methodologies and Best Practices,” *Comput. Sci. It Res. J.*, 2023, doi: 10.51594/csitrj.v4i3.659.
[13] S. Shete, “Interface Design for Cybersecurity Risk Quantification With Monte Carlo Simulation,” *Des. Single Chip Microcomput. Control Syst. Stepping Mot.*, 2023, doi: 10.47363/jaicc/2023(2)171.
[14] T. Fagade, K. Maraslis, and T. Tryfonas, “Towards Effective Cybersecurity Resource Allocation: The Monte Carlo Predictive Modelling Approach,” *Int. J. Crit. Infrastruct.*, 2017, doi: 10.1504/ijcis.2017.088235.
[15] M. Sun and Y. Lu, “A Generalized Linear Mixed Model for Data Breaches and Its Application in Cyber Insurance,” *Risks*, 2022, doi: 10.3390/risks10120224.
[16] R. Pal, L. Golubchik, K. Psounis, and T. Bandyopadhyay, “On Robust Estimates of Correlated Risk in Cyber-Insured IT Firms,” *Acm Trans. Manag. Inf. Syst.*, 2019, doi: 10.1145/3351158.
[17] A. Granato and A. Polácek, “The Growth and Challenges of Cyber Insurance,” *Chic. Fed Lett.*, 2019, doi: 10.21033/cfl-2019-426.
[18] S. Romanosky and E. L. Petrun Sayers, “Enterprise Risk Management: How Do Firms Integrate Cyber Risk?,” *Manag. Res. Rev.*, 2023, doi: 10.1108/mrr-10-2021-0774.
[19] K. Awiszus, T. Knispel, I. Penner, G. Svindland, A. Voß, and S. Weber, “Modeling and Pricing Cyber Insurance,” *Eur. Actuar. J.*, 2023, doi: 10.1007/s13385-023-00341-9.
[20] S. Mamanazarov, “Insuring Data Risks: Problems and Solutions,” *Irshad J Law Policy*, 2024, doi: 10.59022/ijlp.166.
[21] S. A. Talesh, “Data Breach, Privacy, and Cyber Insurance: How Insurance Companies Act as ‘Compliance Managers’ for Businesses,” *Law Soc. Inq.*, 2018, doi: 10.1111/lsi.12303.
[22] G. W. Peters, M. Malavasi, G. Sofronov, P. V. Shevchenko, S. Trück, and J. Jang, “Cyber Loss Model Risk Translates to Premium Mispricing and Risk Sensitivity,” *Geneva Pap. Risk Insur. Issues Pract.*, 2023, doi: 10.1057/s41288-023-00285-x.
[23] B. Zhao and W. Chen, “Data Protection as a Fundamental Right: The European General Data Protection Regulation and Its Exterritorial Application in China,” *Us-China Law Rev.*, 2019, doi: 10.17265/1548-6605/2019.03.002.
[24] C. Tikkinen-Piri, A. Rohunen, and J. Markkula, “EU General Data Protection Regulation: Changes and Implications for Personal Data Collecting Companies,” *Comput. Law Secur. Rev.*, 2018, doi: 10.1016/j.clsr.2017.05.015.
[25] C. Elendu, E. K. Omeludike, P. O. Oloyede, B. T. Obidigbo, and J. C. Omeludike, “Legal Implications for Clinicians in Cybersecurity Incidents: A Review,” *Medicine (Baltimore)*, 2024, doi: 10.1097/md.0000000000039887.
[26] T. Lindén, R. Khandelwal, H. Harkous, and K. Fawaz, “The Privacy Policy Landscape After the GDPR,” *Proc. Priv. Enhancing Technol.*, 2020, doi: 10.2478/popets-2020-0004.
[27] N. Saqib, V. Germanos, W. Zeng, and Λ. Μαγλαράς, “Mapping of the Security Requirements of GDPR and NISD,” *Icst Trans. Secur. Saf.*, 2018, doi: 10.4108/eai.30-6-2020.166283.
[28] Y. Hong and M. Xu, “Autonomous Motivation and Information Security Policy Compliance,” *J. Organ. End User Comput.*, 2021, doi: 10.4018/joeuc.20211101.oa9.
[29] A. Alkhalifah, A. Ng, P. Watters, and A. S. M. Kayes, “A Mechanism to Detect and Prevent Ethereum Blockchain Smart Contract Reentrancy Attacks,” *Front. Comput. Sci.*, 2021, doi: 10.3389/fcomp.2021.598780.
[30] N. Jevtić and I. Alhudaidi, “The Importance of Information Security for Organizations,” *Serbian J. Eng. Manag.*, 2023, doi: 10.5937/sjem2302048j.
[31] R. Mai, “Using Information Technology to Quantitatively Evaluate and Prevent Cybersecurity Threats in a Hierarchical Manner,” *Int. J. Appl. Inf. Manag.*, 2023, doi: 10.47738/ijaim.v3i1.51.

[32] C. Liu, C. Wang, H. Wang, and N. Bo, “Influencing Factors of Employees’ Information Systems Security Police Compliance: An Empirical Research in China,” *E3s Web Conf.*, 2020, doi: 10.1051/e3sconf/202021804032.
[33] W. He *et al.*, “Improving Employees’ Intellectual Capacity for Cybersecurity Through Evidence-Based Malware Training,” *J. Intellect. Cap.*, 2019, doi: 10.1108/jic-05-2019-0112.
[34] A. M. Y. Chu and M. K. P. So, “Organizational Information Security Management for Sustainable Information Systems: An Unethical Employee Information Security Behavior Perspective,” *Sustainability*, 2020, doi: 10.3390/su12083163.