DISSERTATION

UNCERTAINTY-PARAMETERIZED ADAPTIVE ISOLATION FOR CRITICAL INFRASTRUCTURE VULNERABILITY MANAGEMENT: THE GUARDIAN FRAMEWORK

Submitted by

Mohamed Rajraji

Department of Systems Engineering

In partial fulfillment of the requirements

For the Degree of Doctor of Philosophy

Colorado State University

Fort Collins, Colorado

Spring 2026

Doctoral Committee:

Advisor: Steven J. Simske

Erika Gallegos
Marie Vans
Brad Reisfeld

ABSTRACT

UNCERTAINTY-PARAMETERIZED ADAPTIVE ISOLATION FOR CRITICAL INFRASTRUCTURE VULNERABILITY MANAGEMENT: THE GUARDIAN FRAMEWORK

Critical infrastructure organizations are faced with temporal impossibilities in the management of their vulnerabilities because more than 50 hours are needed to coordinate security teams, system administrators, compliance staff, and vendors to assess and respond to a single vulnerability, while attackers exploit available vulnerabilities within 24 hours. This challenge is compounded by patch validation processes in regulated industries where standards like NERC CIP-007-6 require up to 70 calendar days for evaluation and remediation. Even with the application of machine learning to vulnerability prediction, the challenge persists since accuracy levels remain constant at about 70% independent of the architecture of the algorithm. This ceiling is imposed by the inadequacy of information in standardized vulnerability data and not by model limitations.

This study presents the Graduated Uncertainty-Aware Risk Decision and Isolation Architecture for Networks (GUARDIAN) framework, which converts the structural uncertainty of machine learning models into a governing parameter for automated graduated protective actions. The framework measures ensemble disagreement using the Jensen-Shannon Divergence (JSD) and converts the resulting uncertainty signal to adjustable parameters that dictate isolation stringency and isolation duration, which allows protection mechanisms to run at machine speed without the bottlenecks of human coordination.

The analysis covered 279,056 common vulnerabilities and exposures from the national vulnerability database, evaluated across six architecturally different models using data from January 2002 to December 2024. From the analysis it emerged that the models' balanced accuracy fell within a range of 47.5% to 69.3% and a temporal stability analysis showed less than 1 percent coefficient of variation. Temperature-scaled calibration reduced the expected calibration error to 0.023. In addition, Spearman correlation analysis proved that ensemble disagreement is a strong predictor of accuracy degradation, at $\rho = -0.92$, $p < 0.001$. The research also established that monotonic relations exist between calibrated uncertainty and the rates of exploitation and durations of protection, and these relations formed the empirical support of the graduated threshold functions in the GUARDIAN framework.

The simulation testing of 6,000 attack scenarios on a water utility infrastructure testbed showed between 56 and 94 percent risk reduction depending on the isolation level that was applied. The multi-signal integration strategy showed 92 percent of rollbacks were properly executed with only 1.6 percent being premature. In a field validation using 15 practitioners at a regional healthcare facility, 80 percent said they would advocate for adopting the framework, whereas 60 percent said they would not trust complete automation mode. This showed that operational deployment would need to go through staged advancement from advisory mode to graduated automation. The study also showed that overcoming the temporal impossibility gap does not require perfect prediction, but rather an effective use of model uncertainty.

ACKNOWLEDGMENTS

Pursuing this doctoral work would not have been possible without the support of my family. I thank my mother, Amina, for her constant encouragement and belief in me. To my wife, Wafaa, for never doubting this journey even when I did, and for shouldering more so I could see it through. To my daughters, Sara and Yacout, for reminding me every day what truly matters and for giving me the greatest reason to finish.

I owe a deep gratitude to my advisor, Dr. Steven Simske, who has been the greatest supporter of this work from the very beginning. He provided every form of support I needed, intellectual, professional, and personal.

Finally, I thank the members of my committee for their thoughtful feedback and guidance that helped strengthen this dissertation.

Thank you all.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

LIST OF ACRONYMS

AI	Artificial Intelligence
ANOVA	Analysis of Variance
API	Application Programming Interface
ATT&CK	Adversarial Tactics, Techniques, and Common Knowledge (MITRE)
BERT	Bidirectional Encoder Representations from Transformers
CFR	Code of Federal Regulations
CI	Critical Infrastructure
CIP	Critical Infrastructure Protection
CISA	Cybersecurity and Infrastructure Security Agency
CISO	Chief Information Security Officer
CPS	Cyber-Physical Systems
CPU	Central Processing Unit
CSF	Cybersecurity Framework
CVE	Common Vulnerabilities and Exposures
CVSS	Common Vulnerability Scoring System
CWE	Common Weakness Enumeration
DOE-NE	Department of Energy, Office of Nuclear Energy
DOJ	Department of Justice
ECE	Expected Calibration Error
EPA	Environmental Protection Agency
EPSS	Exploit Prediction Scoring System

FDA	Food and Drug Administration
FIRST	Forum of Incident Response and Security Teams
FTE	Full-Time Equivalent
GNN	Graph Neural Network
GPU	Graphics Processing Unit
gRPC	Google Remote Procedure Call
GUARDIAN for Networks	Graduated Uncertainty-Aware Risk Decision and Isolation Architecture
HIPAA	Health Insurance Portability and Accountability Act
HMI	Human-Machine Interface
ICS	Industrial Control Systems
IDS	Intrusion Detection System
IEC	International Electrotechnical Commission
IP	Internet Protocol
ISAC	Information Sharing and Analysis Center
IT	Information Technology
JSD	Jensen-Shannon Divergence
KEV	Known Exploited Vulnerabilities
KL	Kullback-Leibler (Divergence)
LLM	Large Language Model
MCC	Matthews Correlation Coefficient
ML	Machine Learning
mTLS	Mutual Transport Layer Security

NER	Named Entity Recognition
NERC	North American Electric Reliability Corporation
NIST	National Institute of Standards and Technology
NLP	Natural Language Processing
NRC	Nuclear Regulatory Commission
NSX	Network Security and Virtualization (VMware)
NVD	National Vulnerability Database
OAuth	Open Authorization
OSHA	Occupational Safety and Health Administration
OT	Operational Technology
PACS	Picture Archiving and Communication System
PERA	Purdue Enterprise Reference Architecture
PLC	Programmable Logic Controller
RAM	Random Access Memory
RBF	Radial Basis Function
REST	Representational State Transfer
SANS	SysAdmin, Audit, Network, and Security
SCADA	Supervisory Control and Data Acquisition
SIEM	Security Information and Event Management
SOC	Security Operations Center
SQL	Structured Query Language
SVM	Support Vector Machine
TF-IDF	Term Frequency-Inverse Document Frequency

TLS	Transport Layer Security
TSA	Transportation Security Administration
VLAN	Virtual Local Area Network
XGBoost	eXtreme Gradient Boosting
ZTA	Zero Trust Architecture

CHAPTER 1: INTRODUCTION

1.1 Background of the Study

Critical infrastructure sectors, including energy, water systems, healthcare, transportation, communications, and finance, are the cornerstone of any economy (Yusta et al., 2011). Critical infrastructure sectors are so essential to the economy that destroying or disabling their assets, networks and systems undermines national security, including economic stability, public health, and public safety (Masys, 2014). These sectors rely on two types of technology systems which organizations have been progressively integrating. Historically, information technology (IT) systems were used to process data and business processes, and operational technology (OT) systems were used to control physical processes, like pumps, valves and generators. Since the early 2010s, these previously separate domains have become interconnected, and so the systems are usually interdependent (Stallings, 2019). The result is closely combined IT and OT systems that are also called cyber-physical systems (CPS), which achieve efficiency, allow automation and enhance monitoring and control capabilities in essential systems of infrastructure. Nevertheless, this convergence broadens the attack surface by exposing physical process controls to network-based attack vectors.

Since CPS architectures provide remote command delivery to controllers that control physical activity, when a hacker breaches the network layer, he or she can control physical processes at will. IT/OT convergence therefore implies that the impact of incapacitation of a networked system extends beyond financial loss to potential loss of life (Zio, 2016). The types of threats that cause these consequences differ in their levels of mitigation efficiency. The potential threats included in the physical category (i.e., accidents, maintenance failures, physical attacks)

can be mostly avoided with the help of the defined mitigation measures overseen by government agencies and CI security controllers (Makrakis et al., 2021). The cyber-threats, in their turn, are more complex because of vulnerabilities in interconnected systems that enable malicious attackers to pursue sabotage and physical disruption rather than data theft or financial gain (Riggs et al., 2023).

The increase in the rate and complexity of cybersecurity attacks on CI systems has exceeded the ability of CI organizations to coordinate defenses despite sustained prevention efforts (Hu et al., 2018; He and Yan, 2016). Defensive coordination takes days and weeks; worse still, attackers are able to exploit vulnerabilities within hours after disclosure. This coordination gap was exhibited by the ransomware attack conducted on Colonial Pipeline in May 2021.

Colonial Pipeline is one of the most strategic energy infrastructures, providing 45 percent of the fuel to the East Coast of the US. The DarkSide ransomware group breached the company systems and operated undetected for 8 days before deploying ransomware, causing a five-day shutdown that triggered severe fuel supply disruptions and forced payment of USD 4.4 million in ransom (Falco & Rosenbach, 2022; Madsen, 2024; Beerman et al., 2023).

The Colonial Pipeline attack was not an isolated incident. Threats against critical infrastructure have become more frequent, as reported by Forescout Technologies (2024), which recorded more than 420 million cyber attacks on critical infrastructure globally between January and December 2023, a figure representing approximately 13 attacks per second and a 30% increase over the previous year. The rate at which exploitation is taking place has also increased. In December 2021, the Apache Log4Shell vulnerability (CVE-2021-44228) was publicly announced and attempts to exploit the vulnerability were reported within 9 minutes of the announcement, with over ten million attempts to exploit the vulnerability per hour, peaking in

the United States (Cloudflare, 2021; Cyber Safety Review Board, 2022). A retrospective study showed that exploitation had occurred at least 9 days prior to disclosure (Hiesgen et al., 2024). This 9-minute exploitation window effectively precludes coordinated response by the organization. To mitigate exploitation, CISA suggests timely remediation of vulnerabilities but does not set a particular timeline. The agency establishes a response process but leaves associated agencies and government bodies to develop their own timelines.

Critical infrastructure sectors have responded by establishing patch management structures with predetermined timelines to address them. As an illustration, the NERC CIP-007-6 standard requires that organizations evaluate the applicability of security patches within 35 calendar days and then apply them or develop a mitigation plan within another 35 calendar days, creating a compliance window of approximately 70 days (NERC, 2016). Likewise, CISA Binding Operational Directive 19-02 requires federal civilian agencies to remediate critical vulnerabilities within 15 calendar days and high-severity vulnerabilities within 30 calendar days of initial detection (CISA, 2019). These schedules exist because organizations require 15-70 days to certify patches, ensure they do not cause conflicts in systems where they are deployed, and not introduce new security and safety concerns (Souppaya & Scarfone, 2022). However, this creates a window of exposure that attackers can exploit immediately following disclosure. This gap is not accidental but structural, because the regulatory measures establish patch validation periods, sequential stakeholder coordination, and the safety testing that prohibits the acceleration of deployment. The patch development process cannot be rushed, as it may lead to disruption in operations. Weintraub (2016) highlights that the patch development process is time-consuming due to organizational coordination delays, and the tediousness of the patch deployment process. This delay, however, opens a bigger window for exploitation, especially if the attackers strike

within minutes. This mismatch in the timeline between exploitation and remediation forms a temporal impossibility. Figure 1 illustrates this gap.

Figure 1. The Temporal Impossibility Gap Between Exploitation and Remediation Timelines. Note. Exploitation data synthesized from VulnCheck (2024), Kutscher (2023), and CISA analysis. Remediation timelines from NERC CIP-007-6 (2016) and Edgescan (2024).

To overcome the temporal impossibility, however, there is a need for a proactive approach to vulnerability management. Proactive protection demands automated decision-making because human coordination cannot match exploitation timelines. Automated decisions demand confidence assessment as protection stringency must vary with the reliability of the assessment. Conventional machine learning models also generate confidence scores that indicate probability of prediction rather than prediction reliability (Guo et al., 2017). This implies that a model that predicts 65% confidence may be correct 80% or 45% of the time, and therefore, the outputs cannot be reliable for decision-making. Ensemble disagreement offers calibrated uncertainty by measuring the degree of difference between multiple models, enabling graduated

response: when disagreement is high, protection is conservative; when disagreement is low, protection is permissive.

1.2 Problem Statement

The temporal impossibility presented in Section 1.1 presents a dilemma: patches cannot be applied too quickly as it can lead to operational disruption and system instability, but by withholding patch validation, it will mean that the system is exposed to exploitation (Riggs et al., 2023; Bochman and Freeman, 2021). Patch validation procedures also need to consider possible operational failures before deployment (Erdemir, 2022). The need to address the temporal impossibility is met, in any case, by a proactive vulnerability management approach. Proactive protection requires automation since human coordination cannot match exploitation timelines (Kudrati & Pillai, 2022). Since 2011, the body of research on machine learning models to detect and predict vulnerabilities has increased significantly (Shiri Harzevili et al., 2024). Nonetheless, empirical research records that even with algorithmic complexity, the characteristic curve of accuracy levels off and the accuracy does not generally increase in particular levels: distinct models reach a similar degree of accuracy. GPT-4-turbo, LLaMA (7-billion parameters), and Gemma reached 50-63% balanced accuracy (Steenhoek et al., 2024; Lin and Mohaisen, 2025). In the same manner, pre-trained code models such as CodeBERT, VulBERTa, and UniXcoder reached an accuracy of 54.6-57.5% on realistic vulnerability detection tasks (Risse & Böhme, 2024). In the current dissertation, six models with various architectures were analyzed with 47.5% to 69.3% balanced accuracy, which can be said to be relatively supportive to the findings and indicate that the performance ceiling is governed by inherent constraints of information and not the restrictions of the algorithm.

These accuracy constraints are due to the information insufficiency of standardized vulnerability data to a large extent. Although the models presented in the current dissertation are based on Common Vulnerabilities and Exposures (CVE) and the Common Vulnerability Scoring System (CVSS), which are the industry standards used to determine severity, the metrics have flaws. According to Allodi (2015a), CVSS is useful due to its standardized scoring, transparent methodology, and regulatory alignment. Nevertheless, it has a significant structural flaw in that it fails to put severity into context with a particular environment (Frühwirth & Männistö, 2009). Standardized scores tend to overlook the key aspects of organizations like network topology, the criticality of business processes, compensating controls, and data classification (Garbis and Chapman, 2021; Quinn et al., 2022; Fitzgerald, 2018; NIST, 2004; Ganin et al., 2016).

Model accuracy increases when organizational context is included (Le et al., 2022). Allodi and Massacci (2014) demonstrated this weakness using empirical evidence: only 5.5% of high-severity vulnerabilities (CVSS $\geq$ 7) were ever exploited in the operating conditions, and 60% of exploited vulnerabilities had lower scores than 7. This proves the claim that the technical severity in standardized descriptions is not comparable to the real risk in specific deployments. Nonetheless, in the absence of this organizational environment, algorithms and models will have a restricted capacity to aid in vulnerability management, at least when applied to CI systems. A framework quantifying assessment uncertainty based on ensemble disagreement and mapping the level of uncertainty to isolation stringency and isolation duration would enable machine speed automated response to be calibrated to assessment confidence. The hypothesis of this study is that ensemble disagreement among vulnerability assessment models, which can be measured by Jensen-Shannon divergence, gives calibrated uncertainty that can be directly used to parameterize response without human coordination. The operational architecture implementing

this proposal is presented in Figure 2. The structure transforms ensemble model disagreement into calibrated uncertainty signals that directly parameterize graduated isolation response enabling machine-speed automated protection without human coordination bottlenecks.

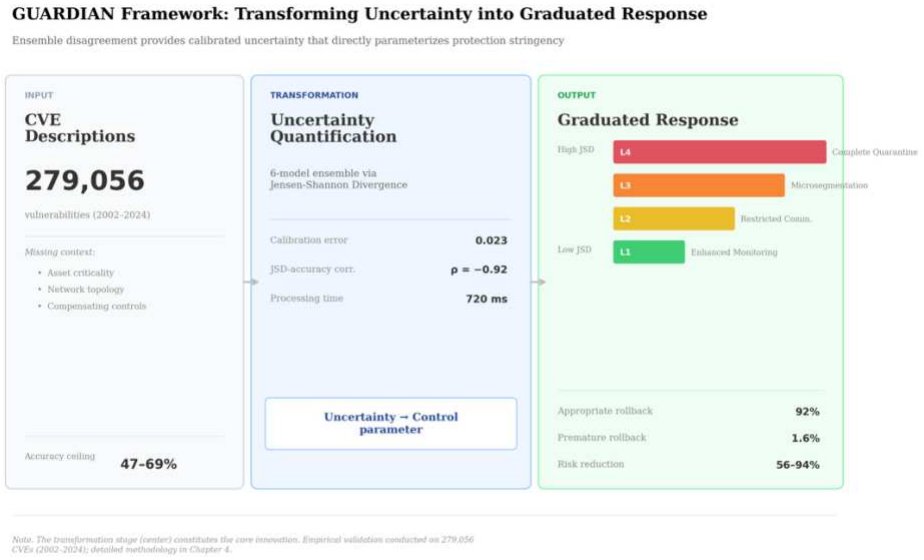

Figure 2. GUARDIAN Framework: Converting Uncertainty into a Graduated Response. Note. The transformation stage (middle) is the centre of innovation. Empirical validation statistics on 279,056 CVEs (2002-2024) described in Chapter 4.

1.3 Research Objectives

This study addresses the information gap by developing the GUARDIAN framework (Graduated Uncertainty-Aware Risk Decision and Isolation Architecture for Networks), which transforms irreducible uncertainty into a parameter governing automatic graduated protective responses to system vulnerabilities. To this end, the research will have five research objectives:

1. To empirically describe the accuracy ceiling of machine learning vulnerability prediction on standardized CVE descriptions with no organizational context and using six architecturally different models. Past studies report accuracy levels plateauing between 50-70%, with rigorous evaluation showing performance closer to 50-55% balanced accuracy (Risse &

Böhme, 2024; Steenhoek et al., 2024). This objective determines whether the ceiling reflects algorithmic limitations or fundamental information constraints in standardized vulnerability data.

2. To measure ensemble disagreement by Jensen-Shannon divergence and test its calibration when used to parameterize response by automated methods. Research has shown that ensemble disagreement is positively related to prediction error (Lakshminarayanan et al., 2017), yet this principle has not been applied to cybersecurity vulnerability assessment.

3. To develop uncertainty-parameterized threshold functions that provide graduated protection and still provide operationally acceptable levels of availability. This goal will transform the patterns of exploitation timings recorded by industry studies (Kutscher, 2023; VulnCheck, 2024; Qualys, 2023; Condon et al., 2023) into temporal grading thresholds determined in terms of uncertainty.

4. To confirm that the GUARDIAN framework meets technical performance characteristics in a cyber-physical water treatment testbed using simulation. Regulatory limits of CIs do not allow experimental testing of live critical infrastructure, which requires high-fidelity simulation based on MITRE ATT&CK ICS attack taxonomy.

5. To test how practitioners accept uncertainty-parameterized isolation recommendations using field validation in a healthcare critical infrastructure setting. Technical feasibility does not guarantee operational adoption because the success of deployment depends on human factors. (Sundaramurthy et al., 2014a).

1.4 Research Questions

To address the above research objectives, five main research questions will be used:

1. What is the accuracy ceiling of machine learning vulnerability prediction using six architecturally diverse models trained on standardized CVE descriptions with no organizational

context? The question is about the assessment constraints due to algorithmic limitations and the structural information gaps in the standardized disclosures of vulnerabilities (Allodi & Massacci, 2014).

2. Is ensemble disagreement, quantified by Jensen-Shannon divergence, a good measure of uncertainty with Expected Calibration Error less than 0.05 that can be used for automated response parameterization? With calibration, predicted confidence is equal to empirical accuracy, which makes automatic decision-making possible without human confirmation (Lakshminarayanan et al., 2017).

3. Do uncertainty-parameterized threshold functions based on the exploitation timing research result in graduated protection with an availability level greater than 70% at all isolation levels? Industry data shows 23-28 percent of vulnerabilities are exploited within 24 hours of disclosure (VulnCheck, 2024; Qualys, 2023), which makes traditional coordination impossible.

4. How well does the GUARDIAN framework reduce risk and achieve rollback under simulated conditions of cyber-physical water treatment with set signal quality parameters? This question measures the performance of the framework at the bounds such as 95 percent accuracy of the IDS and reported MITRE ATT&CK ICS attack patterns.

5. Do critical infrastructure practitioners approve of uncertainty-parameterized isolation recommendations for operational deployment and what is the difference in trust between the acceptance of recommendations and the acceptance of automation? This question strives to fill the technical feasibility-operational adoption gap in safety-critical settings.

The dependency chain between these research questions is shown in Figure 3. Each subsequent question builds on prior findings: RQ1 and RQ2 define the constraints of the

problem, RQ3 develops threshold functions within these constraints, and RQ4-RQ5 validates the resulting framework.

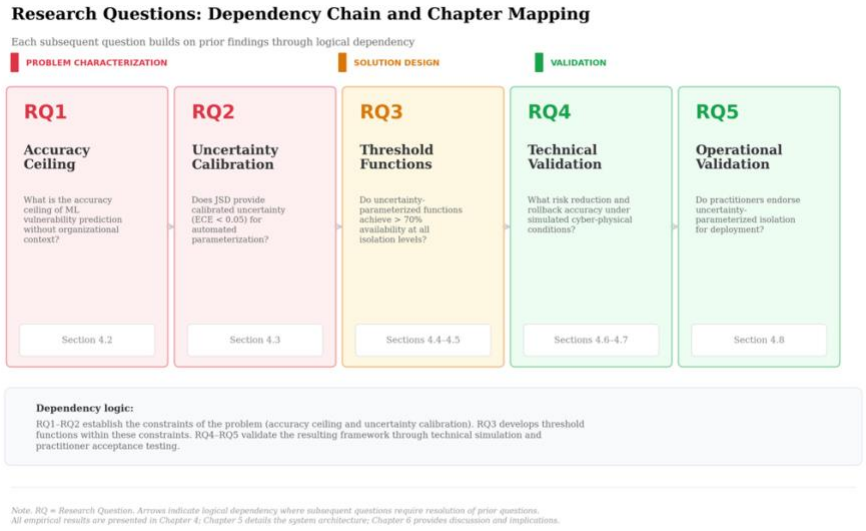

Figure 3. Research Questions: Dependency Chain and Chapter Mapping

1.5 Significance of the Study

1.5.1 Contribution to Literature

This study contributes to the academic discussion on vulnerability management and machine learning since it reveals empirically that there exists a structural accuracy ceiling in vulnerability prediction using algorithms. The concepts proposed in prior studies indicate that despite the sophistication of algorithms, the performance curve levels off at 50-70 per cent accuracy. This study analyzes 279,056 CVEs between 2002 and 2024 to measure the accuracy of ML for vulnerability assessment. This empirical confirmation determines whether the ceiling is due to inefficiency of algorithms or lack of information. This is significant since the standardized CVE data lacks organizational context, which includes information on asset criticality, network topology and compensating controls (Jacobs et al., 2021).

The second important contribution is that the quantification of uncertainty as an active decision parameter has been operationalized. The study uses ensemble model disagreement to assess uncertainty by applying Jensen-Shannon divergence to demonstrate that uncertainty is an objective and calibrated measure of assessment reliability. This generates a paradigm shift of performing evaluation on the basis of accuracy to automation with uncertainty. Past studies reveal that the relationship between ensemble disagreement and prediction error is very close (Lakshminarayanan et al., 2017), and very little research has applied the same principle to cybersecurity vulnerability assessment. This research demonstrates calibrated uncertainty quantification achieving Expected Calibration Error of 0.023 with strong disagreement-accuracy correlation (Spearman $\rho = -0.92$), validating that ensemble disagreement reliably signals prediction uncertainty. The proposed GUARDIAN framework advances this work by transforming patterns of model disagreement into actionable protective measures.

Finally, the dissertation offers a theoretical framework of graduated protection in the presence of uncertainty. It will fill the gap that exists between the operational cybersecurity and probabilistic modeling. The research approach is based on adaptive security and cyber-physical system resilience studies showing that uncertainty can be used as a control variable in the automated defense of high-stakes critical infrastructure environments.

1.5.2 Practical Implications

In addition to being of academic importance, the study will have important operational and regulatory implications for cybersecurity practitioners, critical infrastructure operators, and policy makers. The GUARDIAN framework presents a protect now, validate later model which resolves the time issue between fast exploitation of threats and long patch validation periods. Organizations can also continue to operate and reduce the immediate risk of exploitation by

automatically changing the isolation stringency depending on uncertainty levels. It is a design that enables safety-critical systems, e.g. water treatment, energy distribution systems and nuclear control, to react in seconds to emergent vulnerabilities without circumventing required validation procedures (NERC, 2016; CISA, 2019).

In practice, the proposed framework reduces false-positive operational anomalies that accompany the conventional automated defense systems. In contrast to binary isolation and delayed response, the GUARDIAN model has the capability to impose graduated protection. This makes it possible to make it stricter in case of model disagreement or when uncertainty is high and can be loosened as confidence increases. This mechanism directly covers one of the most problematic aspects of industrial cybersecurity by balancing reliability and responsiveness. The structure and design comply with regulatory provisions that enable the use of compensating controls as an alternative to immediate patching, providing documentation-friendly automation that fulfills the compliance requirements as per the national regulations.

The second implication is workforce optimization. Human coordination that has been relied upon to date in times of vulnerability has proven unsustainable, as events like the Colonial Pipeline attack have demonstrated what coordination delays can lead to (Falco & Rosenbach, 2022). The GUARDIAN model will seek to alleviate the cognitive and procedural load on security teams by automating major elements of risk assessment and protection activation. This transition will allow analysts to focus more on strategic monitoring rather than repetitive tasks thus helping to drive the broader movement towards machine-speed defense as part of zero-trust and adaptive security frameworks.

The practical feasibility of the framework is demonstrated through simulated cyber-physical water treatment under intentionally realistic conditions. This simulation had a rollback

accuracy of up to 94 per cent and system stability, thus proving that the framework can deliver measurable security benefits through rollbacks without disruption to the service. The resilience objectives, as reflected in the Cybersecurity and Infrastructure Security Agency guidelines, are evident in the strategy, which offers a feasible mechanism by which critical infrastructure industries can mitigate the risk of exploitation during validation delays.

1.6 Scope of the Study

The study is devoted to critical infrastructure cybersecurity, i.e., vulnerability management dilemmas, which are induced by the time gap between rapid adversarial exploitation and slow patch validation procedures. The study discusses the use of quantified uncertainty to the automated decision system to support protective responses that would be suitable to address the needs of safety of operations and the regulatory standards.

The research is confined to automated decisions about the isolation of cyber-physical systems (CPS) in which digital and physical processes are closely coupled, as in for example the case of water treatment plants, electrical generating plants and industrial control systems. The reason behind these selections is due to nature of their activities that demand the utmost safety, and therefore conventional quick patch deployment cannot be done. The emphasis on the critical infrastructure, in turn, presents a rigorous background onto which the viability and security of the proposed GUARDIAN framework can be examined.

There are four essential components of the research. It initially empirically examines the cause of the fact that the accuracy of assessment levels off in algorithms trained on a dataset of 279,056 vulnerabilities obtained through the National Vulnerability Database (NVD). Second, it measures uncertainty through ensemble disagreement by using Jensen-Shannon divergence, and the fact it is a good predictor of prediction reliability is verified. Third, it employs a cyber-

physical water treatment testbed to carry out a simulation-based validation to establish the possibility of automated uncertainty-parameterized isolation under realistic conditions. Fourth, it conducts field validation among 15 healthcare practitioners to evaluate operational acceptance of uncertainty-parameterized isolation advice.

However, the study does not include full-scale production application of the framework on live infrastructure because of the safety and regulatory considerations of critical systems. The research involves field validation that deals with practitioner acceptance of technical feasibility testing and omits the whole process of organizational change management and production deployment. The study is concerned with the technical feasibility and the quantitative validation of the GUARDIAN framework as a foundation to further research of operational practice in vital sectors.

1.7 Structure of the Dissertation

Six chapters make up the dissertation. Chapter 1 (this chapter) provides the context of the study, the statement of the research problem, objectives and questions, and the description of its significance, scope and organization. Chapter 2 provides a literature review, summarizing academic and commercial literature on vulnerability assessment, uncertainty quantification, and adaptive cybersecurity models and identifying gaps in the literature that are critical to the GUARDIAN framework. Chapter 3 explains the methodology of the research, the mixed-method design that combines empirical data analysis and simulation-based tests, the data sources, the modeling strategies, and the limitations of the testing. Chapter 4 presents the findings of ensemble analysis, uncertainty quantification, simulation validation of automated protection in a cyber-physical water treatment testbed and field validation of practitioner acceptability in a healthcare critical infrastructure context. Chapter 5 introduces the system architecture, such as

key design principles of GUARDIAN framework, component architecture, deployment factors, and validation metrics. Chapter 6 presents the discussion, implications, and conclusion, addressing the findings interpretation, summarizing the theoretical and practical contributions, limitations, and suggestions on future research and implementation.

CHAPTER 2: LITERATURE REVIEW

2.1 Introduction

A strategic friction at the core of vulnerability management is the rising complexity of new cyberattacks and inflexibility of the critical infrastructure sectors. Here, organizations are forced to respond to new threats within a very short duration of time and all security operations involved should be accurate to prevent the introduction of undesirable impacts on security in highly regulated and safety-related systems. Lacking a criterion that would enable a persistent ranking and prioritizing of the types of vulnerabilities that would need urgent attention, organizations were unable to manage their assets well in times of crisis. This problem has been mitigated with standardized scoring systems like the CVSS and probabilistic systems like the EPSS, which have led to better communication and prioritization of activity although there are studies that show these scores are not aligned with actual exploitation behavior in the real world. This is made worse by the fact that current strategies route uncertain assessments to human cybersecurity analysts, creating bottlenecks in coordination during crisis events, even as adversaries exploit vulnerabilities within minutes of disclosure. Although alternative defensive approaches have emerged, they share similar limitations. Protective mechanisms such as microsegmentation and Zero Trust architectures do provide graduated protection capabilities, but they still lack automated decision-making to deploy at the required speed without relying on human coordination. A combination of those restrictions points to the need to broaden the theoretical basis of uncertainty in the vulnerability assessment context, and to have instruments capable of converting uncertainty-conscious predictions into actionable security procedures. So, the chapter offers an overview of research in five fields: critical infrastructure cybersecurity,

vulnerability management systems, vulnerability prediction based on machine learning, uncertainty quantification, and adaptive protection architecture, and identifies the conceptual and empirical gaps that have informed the approach of the GUARDIAN framework to uncertainty-parameterized adaptive isolation.

2.2 Critical Infrastructure Cybersecurity Landscape

The cyber-physical interdependence of critical infrastructure systems amplifies risk since a compromised asset has downstream effects that extend beyond its original sector (Rinaldi et al., 2001; CISA, 2021; Shan & Myeong, 2024). Compounding this challenge, operational technology assets rely on old protocols and deterministic communication patterns, as well as hardware with low computational capacity, which limits security surveillance and patching (Knowles et al., 2015; Humayed et al., 2017). These technical and architectural constraints force vendors, regulators and operators to forego reliability and innovation in systems that were not initially meant to be exposed to IP networks. The result of this tension has been the creation of a layered security environment where compensating controls are used in place of direct remediation because changing the system is too risky for operations.

Critical infrastructure threat has experienced significant growth over the past few decades and has changed in nature because of IT-OT convergence, which has established new attack surfaces. Whereas traditional cyber threats primarily target data theft, modern attacks are not only becoming more frequent but also seek physical interference and sabotage of infrastructure (Case, 2016; US DOJ, 2021). Compounding this shift, there has been an increase in the use of advanced tooling and automated pipelines of exploit-kit weaponization that has produced lower operational complexity in attacking industrial networks. These capabilities complicate defensive planning as exploitation timeframes shrink at a higher rate than institutions are able to remediate.

Alcaraz and Zeadally (2015) posit that in the context of cyber-physical systems, this asymmetry is particularly acute due to challenges such as lack of visibility, use of proprietary protocols, and low-latency process control loops, which may negatively affect the implementation of established defensive technologies. Beyond technical constraints, regulatory frameworks seek to give form and structure to critical infrastructure cybersecurity, yet the existing models remain aspirational rather than anchored in empirical evidence of threat behavior. NERC CIP provides baseline controls, audit cycles, and change management requirements which presuppose the review of security patches within 35 calendar days and remediation within an additional 35 calendar days, resulting in a combined compliance window of 70 calendar days (NERC CIP-007-6). IEC 62443 and other industry standards impose similar validation timeframes. Figure 4 contrasts required validation timescales against observed exploitation timescales, confirming that these timelines are structurally incompatible. The mandatory cycle of patch testing, safety verification, and reliance on vendor-approved updates introduces structural latency that attackers can exploit (Ganin et al., 2016). Coglianese and Nash (2020) observe that compliance structures can unintentionally promote minimum-level security practices, as organizations often prioritize passing audits over enhancing actual security performance. These compliance requirements also consume resources that might be better spent on detection engineering and threat hunting, particularly in resource-starved utilities where teams wear multiple hats. Ultimately, these procedural limitations hinder rapid defense not due to conceptual inadequacy, but because procedural inertia simply cannot match the speed of cyber threats.

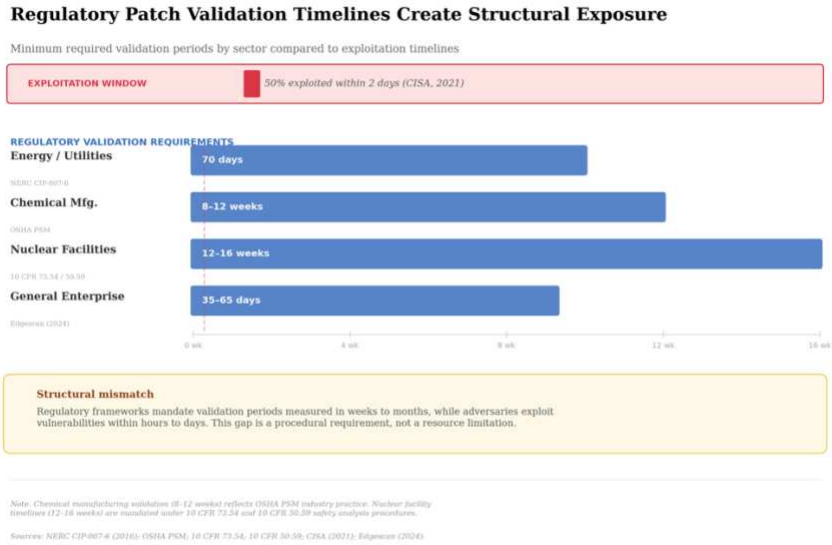

Figure 4. Regulatory Patch Validation Timelines Create Structural Exposure.
Note. Chemical manufacturing facilities under OSHA Process Safety Management typically require 8 to 12 weeks for validation in industry practice, while nuclear facilities commonly require 12 to 16 weeks under 10 CFR 73.54 and 10 CFR 50.59 regulations due to rigorous safety analysis procedures.

2.3 Vulnerability Management Foundations

Vulnerability management has been the foundation of organizational cyber defense for decades, but these processes have structural flaws, which stem from the information infrastructure underlying vulnerability identification and prioritization. CVE, for example, was originally designed to provide a common point of reference when discussing security concerns, but limited to textual descriptions, it does not offer information regarding system configuration, asset criticality, or context of exposure (MITRE, 2025). According to Anwar et al. (2021), who conducted comprehensive quality assessment of the National Vulnerability Database, significant inconsistencies persist in severity rankings and weakness classifications, confirming that a deficiency in such contextual metadata is a critical limitation on the analysis value of CVEs as a prioritization basis, particularly where technology stacks are diverse. Although defenders tend to

view vulnerability identifiers as indicators of underlying risk level, Allodi and Massacci (2014) demonstrate that the vast majority are never exploited. Furthermore, those vulnerabilities that are exploited succeed under deployment-specific circumstances that the CVE schema cannot model. Allodi and Massacci (2017) criticized the implicit notion that listing vulnerabilities is associated with meaningful knowledge of asset exposure and emphasized the difference between vulnerability existence and vulnerability relevance. This difference is often ignored in operational environments, where reporting, patch pipelines, and audit frameworks are still based on enumeration and not contextualization as their main structuring principle.

CVSS was created to overcome these drawbacks of the CVE model. It provides a framework to the assessment of technical severity of vulnerabilities by attack vector, complexity of attack, privileges needed and impact. Although CVSS is believed to be an international standard, the framework has numerous weaknesses that impede its use in the field of vulnerability management. Allodi (2015b) showed that CVSS base scores have weak or inconsistent correlations with real-world exploitation. In fact, the author noted that high-severity ratings are often misleading regarding the actual incentive of attackers or conditions in their environments. CVSS is also viewed as a deterministic scoring model that assumes the characteristics of vulnerabilities are fixed or universal and overlooks the environmental heterogeneity. For example, two organizations with the same CVSS score may be at different risk levels based on their network segmentation, authentication design, or asset exposure. Jacobs et al. (2021) also assert that CVSS embeds a technical essentialism that exaggerates the predictive value of vulnerability while underestimating the role of attacker economics and environmental factors. As a result, its prevalence in regulatory and compliance contexts may unintentionally reinforce a false sense of objectivity in vulnerability prioritization. To address

these deterministic limitations, the Exploit Prediction Scoring System (EPSS) was developed as a probabilistic alternative that estimates the likelihood of exploitation within 30 days using machine learning (Jacobs et al., 2021). However, EPSS introduces its own constraints. Due to its reliance on public reporting and security vendor telemetry, sampling bias and underrepresentation of the activities of exploitation that fall outside of high-observability ecosystems affect EPSS forecasts. There are also similarities in the fact that EPSS addresses exploitation probability but fails to include variables coupled with the operational implications of successful exploitation upon specific assets activated. This limits its use in the setting where the major risk is not data loss but operational disruption. These restrictions suggest that EPSS, similarly to CVSS, gives a biased view of the vulnerability risk and should not be used independently when making high-confidence decisions in complex or safety-critical settings. Probabilistic modeling is a conceptual enhancement but still uses incomplete and externally sourced signals. This reliance on incomplete signals reflects structural information gaps that prevent comprehensive vulnerability assessment.

Perhaps the most important challenge described in the critical infrastructure literature is the operational barriers that prevent vulnerability responses that are rapid or automated. Unlike IT, which can be patched rapidly, OT environments do not permit quick engineering changes as the change process involves deep engagements between engineering teams, safety officers, regulators, and vendors, taking weeks or months even for an emergency patching (Stouffer et al., 2023; NIST, 2024). Such delays cannot be simply described as bureaucratic but reflect the necessity of ensuring that such changes will break the safety conditions. Even though other frameworks imply that compensating safeguards should be added, e.g., closer monitoring or temporary segmentation, they can only be partially applied and must be implemented manually

(Stouffer et al., 2023). Most OT systems do not have a hot-swappable architecture or fallback design (whereas in enterprise IT environments agile security practices can be underpinned by virtualization and redundancy). This restricts the development of adaptive security methods and limits the degree of automation as well as the speed of vulnerability remediation. The defensive posture resulting from these factors cannot cope with the speed of the threat environment, so solutions beyond patching and conventional perimeter defenses are required.

An additional challenge is manual vulnerability triage. The vast majority of organizations still employ a human-centric evaluation process in which analysts review scan results, correlate threat intelligence, evaluate asset criticality and map out remediation strategies. Although some practitioners consider human judgment a corrective to the inherent drawbacks of algorithmic scoring systems, empirical evidence suggests that manual processes introduce inconsistencies, delays and cognitive biases that reduce the reliability of prioritization outcomes (Sundaramurthy et al., 2014a; Reeves & Ashenden, 2023). The emergence of several thousand vulnerabilities comes at a price, since analysts regularly struggle to maintain situational awareness across all asset inventories. These circumstances promote the use of heuristics in which decisions made by triage are based on anecdotal experience, alert fatigue, and urgency. Furthermore, manual remediation planning requires the coordination of multiple teams including operations, compliance, engineering, and vendor management which creates friction within the organization and extends response time. The extra documentation and authorization in regulated industries are also known to further increase such lags even when all stakeholders understand the urgency of moving swiftly. Despite these inefficiencies, Alahmadi et al. (2022) posit that manual triage remains the most common approach not due to its effectiveness but because organizations have not developed meaningful tools to integrate contextual data into working processes. Manual

procedures continue to serve as a buffer, relying on human judgment to compensate for structural shortcomings in vulnerability data. This dynamic is part of a larger problem within the vulnerability management community: an analytical ecosystem characterized by limited information, deterministic scoring, and inconsistent human analysis, which generates a high-pressure environment that slows decision-making and extends response times.

2.4 Machine Learning in Vulnerability Prediction

Machine learning has become an area of promising advancements in vulnerability prediction, taking advantage of statistical trends in massive collections of vulnerability descriptions, the exploit metadata, as well as past threat intelligence. Initial studies in the late 2000s and early 2010s mainly used classic supervised learning algorithms, including Support Vector Machines, Random Forest, k-Nearest Neighbors, and Gradient Boosting, to make exploitability or severity predictions based on CVE text fields (Bozorgi et al., 2010; Sabottke et al., 2015). These experiments showed that textual characteristics might help achieve moderate predictive accuracy, but they also revealed severe performance constraints, especially when there was a need to differentiate between vulnerabilities whose descriptions were very similar yet may have very different impacts in the real world. Later research used feature engineering pipelines of ever-increasing complexity, starting with TF-IDF vectorization, then n-gram extraction, part-of-speech tagging, and even hand-crafted lists of keywords, and achieved only moderate dataset-specific improvements. Bullough et al. (2017) argued that these first-generation models were implicitly based on lexical artifacts rather than on meaningful security semantics and therefore they were unable to generalize across time, vendors and software ecosystems. Besides this, the classical machine learning tools were based on the premise that there would be no change in the usage pattern of the methods over time and even this has also been disproved by the fact that

attacker objectives, tools and targets were changing at an alarming rate. This meant that the original result of the research into the topic of ML-based vulnerability prioritization did not live up to the hype when applied to non-research data sets.

Deep learning raised expectations for more accurate vulnerability predictions because of its ability to extract semantic meaning from text. Subsequent research employed convolutional neural networks, recurrent neural networks, and later transformer-based language models (BERT, RoBERTa) to consider both syntactic and contextual relationships instead of analyzing different forms of CVE descriptions separately using bag-of-words features (Fu and Tantithamthavorn, 2022; Li et al., 2019). Even though such models showed a higher capability of learning high-dimensional representations of security-related language, empirical studies showed that the performance improvement of these models rarely translated to any uplift in predictive accuracy in the real world. Comparisons between deep learning models and simpler ones showed a narrow difference in most cases, ranging from a few percentage points to no difference at all (Chakraborty et al., 2021). Such models would need contextual information from vulnerability descriptions to distinguish between easily exploitable cases and those requiring specialized environmental conditions, but the descriptions do not provide this. Even complex models cannot confidently assess exploitability without knowing the deployment configurations, network context, or asset criticality. Chakraborty et al. (2021) argue that deep learning models hide the real issue of information scarcity through the illusion of complexity that they exhibit when processing inherently under-signaled data. The hypothesis of such criticism is that the plateauing of predictive performance is not caused by model choice but is instead a side effect of structural constraints in the available input data. The synthesis of three generations of study results on accuracy in Figure 5 suggests that a thousandfold growth in parameters only produces single-

digit marginal improvements in accuracy, evidence that the bottleneck is informational, not computational.

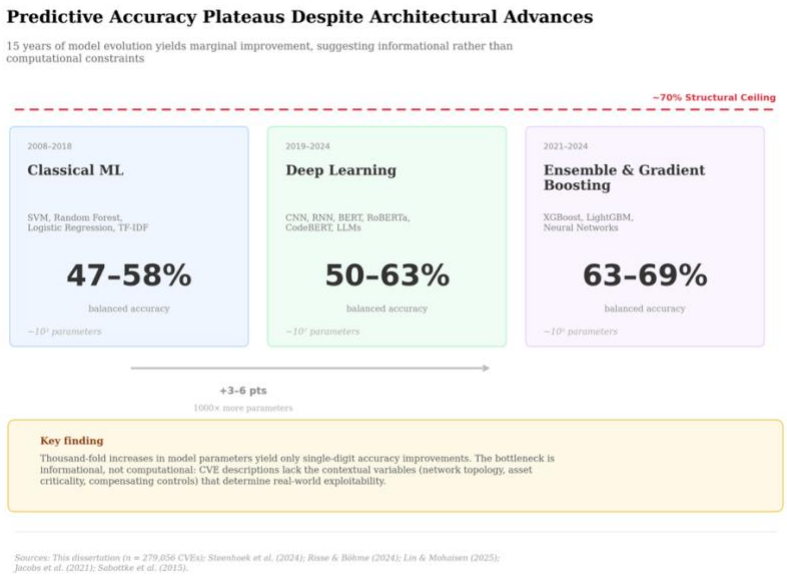

Figure 5. Predictive Accuracy Plateaus Despite Architectural Advances

The recent emergence of Large Language Models (LLMs) has reignited interest in this domain, but similar limitations persist. Steenhoek et al. (2024), in their study evaluating models including LLaMA and GPT-4, found that despite their reasoning capabilities, LLMs struggle to understand the subtle semantic differences that define security flaws, often hallucinating vulnerabilities where none exist. Lin and Mohaisen (2025) further corroborated this finding, demonstrating that LLM performance fluctuates significantly across different programming languages and learning modes. These findings collectively indicate that the accuracy ceiling persists regardless of model sophistication.

Another change regarding the conceptualization of relationships between software components, libraries, and vulnerability inheritance was the introduction of graph-based models. Graph neural networks (GNNs) and graph embedding were applied to software dependency

networks, code property graphs, and control-flow relationships to forecast the propagation of vulnerabilities or the probability that a specific vulnerability would be identified in a specific component (Cheng et al., 2021). These methods expanded the range of analysis beyond unstructured text, yet they too were limited by the availability of data. Complete dependency metadata and source code repositories are not standardized across vendors and software ecosystems, but within proprietary environments, creating datasets with critical structural gaps. In addition, GNNs assume stable and complete graph structures. However, the real-world software ecosystems contain abandoned projects, inconsistent versioning, and undocumented dependencies which induce incomplete or misleading graph representations. Existing empirical evaluations indicate that although graph-based methods are able to discover latent patterns in large open-source ecosystems, their predictive performance drops when applied in heterogeneous enterprise environments where data about dependencies is either fragmented or not available (Croft et al., 2023). These findings suggest that graph-based methods unintentionally shift the vulnerability prediction problem to one of incomplete data, showing that more advanced models are not necessarily the answer to the fundamental limitations in the available data.

The other theme that emerges in the literature on vulnerability prediction is that when training models and assessing them it is difficult to have a strong ground truth. Most research makes use of labels that are based on known exploitation incidents, public-facing exploit databases, penetration testing toolkits, or threat intelligence feeds (e.g., ExploitDB, Metasploit). Nevertheless, these sources represent only a fraction of the exploitation behavior in the global arena, and they introduce sampling bias as they pay more attention to the vulnerabilities that are popular among researchers or target the widely used platforms (Edkrantz and Said, 2015). Consequently, the machine learning systems fed with such datasets will be prone to reinforce

visibility biases instead of learning the correct exploitation patterns. Other authors point to inconsistencies in the definition of exploitation used in different datasets, and different cutoffs applied to define evidence of exploitation, proof-of-concept code, active exploitation, or in-the-wild exploitation. This non-standardization of definitions is detrimental to cross-study comparability and prevents easy assessment of model generalization. Further literature has revealed that temporal drift significantly impacts model accuracy; models that are trained on older vulnerabilities tend to make poor predictions on the newly found ones because of the changes in attacker behavior, vulnerability disclosure, and system design (Le et al., 2020). These findings raise the question of whether machine learning models are generalizable to a real-world environment in which the threats change with time and indicate that current (non-evolving) training paradigms cannot be used to support operational deployment. Practical experience with machine learning throughout the entire lifecycle indicates that the initial enthusiasm for machine learning vulnerability prediction is misplaced and we need to establish frameworks that can operate under uncertainty rather than attempting to make marginal improvements in the effectiveness of algorithms.

2.5 Information Insufficiency and Structural Accuracy Ceiling

The main problem with vulnerability prediction research is the lack of contextual information that would show whether a vulnerability can actually be exploited in a given environment. Published databases including CVE entries, CWE classifications and advisories contain technical specifications of vulnerabilities but lack operational information about them (network topology, authentication flows, compensating controls or asset criticality). Attackers decide whether to exploit a vulnerability based on more than technical severity; factors such as accessibility and potential payoff also matter (Allodi & Massacci, 2014). For example, executing

remote code on a segmented control network might not be as risky in practice as privilege escalation on an internet-facing authentication service, though CVSS scores might indicate the opposite. Without deployment-specific data, machine learning systems cannot tell theoretical severity from practical exploitability. These gaps create uncertainty that cannot be resolved without such data, which is not publicly available. Therefore, any model built on such data is limited by these constraints and may not be trusted to rank vulnerabilities for real-world risk.

Empirical reviews have shown that predictive ceilings reported in the field are not due to algorithmic constraints but to inherent data gaps. Several decades of experimental research in classical machine learning, deep learning and graph modeling shows that the complexity of a model does not translate to correspondingly better performance, i.e. the bottleneck is not computational but informational (Sabottke et al., 2015; Jacobs et al., 2021). Various experiments show that when models use sophisticated representations, e.g. transformer embeddings or graph neural networks, the accuracy can barely reach moderate levels, e.g. 60-70% on balanced datasets. Jacobs et al. (2021) view this plateau as evidence of a structural ceiling set by the data itself. Risse and Böhme (2024) confirm this finding with rigorous empirical testing, demonstrating that state-of-the-art models severely overfit to unrelated features such as variable names or coding style rather than learning the vulnerability itself. Their analysis showed that accuracy reaches approximately 70 percent on benchmarks but drops to 50 to 55 percent when properly tested on semantically similar patched code, a gain of only 3 to 6 percentage points despite thousand-fold increases in parameters. Since critical context cannot be inferred from text alone, models are learning correlations that are not valid across time and context. To make matters worse, vulnerability disclosure practices are highly heterogeneous and clarity, detail, and consistency can differ even among vendors. Some descriptions are very technical while others

provide the least amount of information possible, and so cannot be easily decoded to provide semantically meaningful patterns. Jacobs et al. (2021) also assert that even homogeneous textual descriptions would still lack organizational context, attacker intent and system-level dependencies needed for a realistic determination of risk.

Prediction capabilities of models built on publicly available data are structurally limited. Figure 6 compares variables present in publicly available vulnerability data with variables required to enable precise exploitation prediction, showing that this ceiling is an information-theoretic bound and not an algorithmic one.

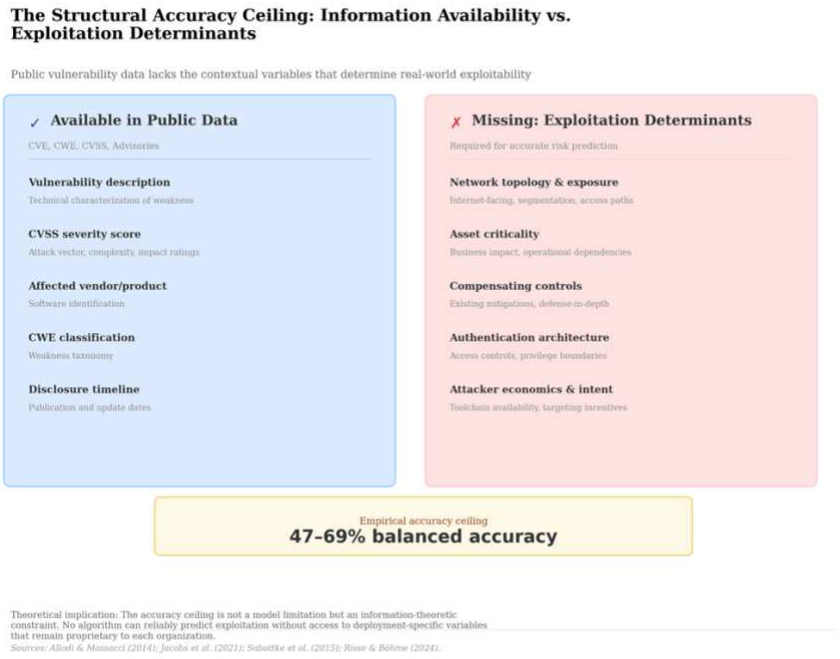

Figure 6. The Structural Accuracy Ceiling: Information Availability vs. Exploitation Determinants

Research on attacker behavior shows that exploitation depends on economic, strategic, and timing factors missing from vulnerability descriptions and scoring metrics. The opportunities attackers look for (preexisting footholds, tool availability, automation pipelines, and expected returns) are not driven by technical severity alone (Allodi, 2017). This means CVSS scores are

only part of the picture. Since these factors cannot be inferred from text or metadata, vulnerabilities that are technically severe but strategically unimportant get misclassified. Timing also matters. Some vulnerabilities are quickly added to automated exploit pipelines while others stay dormant for months or years until new attack paths emerge. Few public datasets capture these timing patterns. Jacobs et al. (2021) note that this approach is flawed because it tries to assess risk without considering attacker economics or how quickly they can exploit a given attack surface. In effect, the gap between technical descriptions and adversary interests sets a structural limit on accuracy.

Context-enriched approaches, however, are hard to operationalize. Contextual data is usually proprietary, sensitive, constantly changing, and not easily collected, standardized, or shared across organizations. Also, such data loses value quickly as system structures change, so any context-based model must constantly be updated to stay relevant. This creates overhead and complexity most organizations cannot handle. Context also varies widely across sectors, vendors, and architectures. What counts as high-risk in a cloud environment may look nothing like high-risk in an industrial control network, and there is no easy way to generalize. This gap shows that accuracy is limited not by model capacity but by missing context. These problems point to the need for frameworks that can work under uncertainty, rather than trying to fix data gaps with more complex models.

2.6 Uncertainty Quantification in Cybersecurity

Uncertainty quantification is an important area in machine learning even though it has not found consistent use in cybersecurity. Classical machine learning assumes prediction errors can be minimized with more complex models, better hyperparameters, or more training data, but not all uncertainty can be reduced. There are two kinds. Statistical learning theory calls them

aleatoric uncertainty, which is noise in the data, and epistemic uncertainty, which is gaps in what the model knows about the underlying process (Kendall & Gal, 2017). Of particular interest in cybersecurity is epistemic uncertainty because defenders lack access to all information about system settings, attacker motivation, or underlying weaknesses.

First attempts at quantifying uncertainty in intrusion detection models and anomaly detection models used ensemble variance and Bayesian models to quantify prediction uncertainty (Corona et al., 2013). These techniques were mostly experimental and failed to revolutionize defensive architectures at the core level. Bassi and Singh (2023) claim that the uncertainty measures have been underused within the broader cybersecurity community because of the historical use of deterministic scoring frameworks and binary alerting frameworks, which are more likely to encourage the assumption that threats can be identified in a categorical way, rather than determined probabilistically. This determinist legacy has hampered the adoption of uncertainty-conscious models, but confidence in a prediction is increasingly seen as more valuable than the prediction itself, especially when operational decisions carry high stakes.

Recently, it has been noted that ensemble based methods can be helpful in estimating the epistemic uncertainty in machine learning prediction particularly in regions where there is scarce, imbalanced or noisy data. Ensemble models use multiple models trained independently to model the distribution over outputs, and discrepancies between model outputs are used as a proxy of uncertainty (Lakshminarayanan et al., 2017). This approach is attractive to cybersecurity since it does not require the re-definition of model architecture or the re-design of existing system and can be applied to both classical and deep learning models. Ensemble variance has been shown to be highly correlated with the potential to misclassify in vulnerability prediction studies, where high model disagreement is often indicative of situations in which the

available data is not enough to generate a confident forecast. It has also been shown that uncertainty-sensitive strategies can be used to minimize the occurrence of false positives, as they allow the systems to differentiate between unclear and confidently categorized samples (Apruzzese et al., 2019; Pierazzi et al., 2020). However, this finding remains underrepresented in general cybersecurity literature, with much consideration paid to simple accuracy measures, even though models with sparse or incomplete feature spaces are known to enjoy disproportional advantage in uncertainty quantification via ensembles. Adoption barriers include concerns about computational cost, lack of standard implementations, and poor fit with security operations processes that traditionally expect binary alerts rather than probabilistic assessments. So even though ensemble approaches offer useful insights, considerable work is still needed to make uncertainty quantification practical for cybersecurity.

Recent work by Murphy et al. (2025) explores entropy-based divergence metrics for ensemble diversity control, demonstrating that metrics like Jensen-Shannon Divergence can effectively quantify the disagreement between models in an ensemble. Their comparative analysis found that JSD-based methods produced predictions that were both more accurate and better calibrated than ensembles using alternative metrics such as Total Variation Distance or Hellinger Distance. This finding validates the theoretical basis for using ensemble disagreement as an uncertainty signal in automated decision systems.

Calibration is another important area of uncertainty research. It addresses the gap between what a model predicts and what actually happens. Without calibration, models get overconfident or underconfident, and that throws off decisions whether you are attacking or defending. The most common methods of calibration in machine learning are temperature scaling, isotonic regression, and Platt scaling. Of these, temperature scaling is favored for being

simple and effective, especially in deep learning classification (Guo et al., 2017). Calibration occurs in malware classification, intrusion detection, and phishing detection in cybersecurity. Jordaney et al. (2017) showed that concept drift is extremely problematic in terms of ensuring calibrated predictions in malware classification over time because models trained on past data become less reliable over time unless they are recalibrated. Although the issue of calibration is pertinent to the problem, it is barely mentioned in the research on vulnerability prediction. Vulnerability datasets are class-imbalanced, exhibit concept drift, and have uneven feature quality, attributes that generally harm model calibration. Croft et al. (2023) noted that vulnerability datasets are characterized by data quality problems that considerably affect the reliability of the model, where uncalibrated models yield high confidence scores on misclassifications, which is an issue in vulnerability management because it makes it difficult to determine whether the prediction was made based on substantive evidence or not enough information. According to other scholars, adaptive defensive measures should be based on estimated uncertainty because it is more indicative of confidence an organization can place in automated assessments. However, gaps in common calibration reference standards and lack of collaboration between machine learning specialists and cybersecurity scientists are barriers to the implementation of the calibration process in vulnerability prioritization systems in real-life scenarios.

Most security tools, including SIEMs, intrusion detection systems, and vulnerability scanners, produce deterministic outputs. Practitioners struggle with probabilities. Some studies show promise. Uncertainty-aware systems can route unclear cases to humans while automating confident ones (Caruana et al., 2015; Gawlikowski et al., 2023). Systems can also self-correct when they tie decisions to uncertainty estimates and update as new information comes in (Gal &

Ghahramani, 2016). Yet vulnerability prediction has largely ignored this work, even though it could address the data gaps mentioned earlier. The core problem is a mismatch: statistical thinking treats uncertainty as information, while traditional cybersecurity architecture expects binary answers.

2.7 Graduated Protection and Adaptive Cybersecurity Architectures

Adaptive cybersecurity architectures have developed as a reaction to the growing instability of the modern threat environment but remain unevenly implemented in operational settings and have yet to be developed conceptually. One of the most mentioned paradigms, based on the postulates of constant checking, least privilege, and trust recalculation, is the Zero Trust Architecture (ZTA) (Kindervag, 2010; Rose et al., 2020). Instead of presuming that a network within the organization is safe, ZTA frameworks focus on conditional access control based on contextual indicators of device posture, identity risk, behavioral analytics, and real-time telemetry. Although this model has addressed many longstanding problems of perimeter-based security, empirical studies indicate that complexity issues make its implementation in complex or legacy-intensive environments quite difficult. Amomo (2025) indicates that ZTA implementations usually require substantial identity modernization, redesigning authentication pathways, and unification of telemetry sources that might not be available in industrial or regulated systems. Rose et al. (2020) also argue that ZTA, irrespective of conceptual appeal, gives little information on how systems should adapt when facing uncertainty or other ambiguities in risk. The model supposes that contextual analytics will produce accurate confidence scores but does not provide a structure to address cases of incomplete, contradictory, or missing telemetry. Such weaknesses suggest that despite being a step toward adaptive

security, ZTA lacks sufficient response to structural decision-making issues in environments subject to uncertainty.

Microsegmentation technologies are another important innovation in adaptive cybersecurity, capable of enforcing fine-grained and policy-based controls on distributed systems. Business solutions such as Illumino, Guardicore, and VMware NSX allow organizations to isolate workloads, restrict traffic between them, and apply custom-designed communication policies that can dynamically react to environmental conditions (Illumio, 2025b; Broadcom, 2025a). Microsegmentation is valued in the literature for its role in curbing attack propagation and maintaining operational continuity even when vulnerabilities are being exploited. However, microsegmentation technologies have significant barriers to operation as well. Applications are typically deployed based on the awareness of application dependencies, communication patterns, and workload character, and when poorly configured can result in serious interruptions to service (Amomo, 2025). Furthermore, such systems often work in binary states, meaning that workloads are either fully authorized or fully refused access, with little room for intermediate states representing different levels of risk or confidence. In addition, microsegmentation is not easily reconfigured to accommodate operational statuses that are changing rapidly such as emergency patch releases, ad hoc configuration changes, or threat intelligence. These barriers indicate that microsegmentation provides structural protection functionalities but lacks proportional uncertainty-based adaptive processes.

Adaptive security architectures more broadly seek to bring together prediction, prevention, detection, and post-hoc functionality. The adaptive security model concept has evolved from early analyst frameworks (MacDonald & Firstbrook, 2014) into more rigorous engineering approaches. NIST Special Publication 800-160 formalizes the concept of cyber-

resilient systems with specific design principles for critical infrastructure, including adaptability, withstand, and recovery capabilities (NIST, 2021). This engineering-focused approach does not perceive security as a one-time solution but as a dynamic cycle that responds to real-time feedback and concentrates on behavioral analytics, automated detection, and continuous adjustment of defensive controls. Although this paradigm has been utilized in both academic and commercial research, operationalization in different domains is inconsistent. Some studies indicate that adaptive systems are actually very sensitive to the quality and promptness of telemetry, which is often biased in mixed-technology stack environments or with older elements (Buczak & Guven, 2015). Furthermore, adaptive architectures based on statistical or machine learning-powered prediction of risk face systematic problems of drift, adversarial effects, and false positives that restrict the applicability of automated response triggers (Apruzzese et al., 2019). The lack of strong uncertainty quantification in these models makes them incapable of differentiating between real threats and unclear signals. This is especially restrictive in industrial, safety-critical environments, in which an overreaction to uncertain forecasts will disrupt operations while underreaction subjects systems to intolerable risk (Shan & Myeong, 2024). In turn, despite the development of adaptive architectures in comparison to static controls, the lack of uncertainty-sensitive logic limits their effectiveness in high-stakes environments that need clear and proportional responses.

Temporal self-correction is a new aspect of adaptive cybersecurity that addresses the need for systems to change their confidence and enforcement levels over time as new evidence becomes available. In contrast to traditional models that assume risk is stagnant or monotonic, time-sensitive models recognize that the absence of exploitation attempts, stable threat intelligence, or lack of anomalous action may indicate declining risk, thus justifying the

reduction of defenses (Sommer & Paxson, 2010; Householder et al., 2017). Multiple works on intrusion detection and anomaly detection prove that confidence scores with a temporal decay contribute to the reduction of alert fatigue and the enhancement of automated responses (Kruegel & Vigna, 2003; Hassan et al., 2019). Nevertheless, significant gaps exist in temporal models, which are rarely incorporated in the vulnerability management pipeline, and most current systems lack features to restore previous constraints when new information becomes available. Some recommendations for modeling changing risk levels include Bayesian or reinforcement learning models (Nguyen & Reddi, 2021). However, such approaches are likely to remain uncommon due to their complexity, data deficiency, and inability to link probabilistic results with operational decision-making criteria. Furthermore, unlike graduated enforcement systems like tiered isolation or dynamic access control, temporal self-correction is rarely designed to be precise and reversible. These disadvantages suggest that adaptive and time-conscious models, while an improvement over former ones, should be complemented with uncertainty-conscious logic and enforcement mechanisms to have a meaningful functional impact.

2.8 Human Factors, Coordination Bottlenecks and Temporal Impossibility

The role of human factors in the vulnerability management landscape has been largely ignored even though it plays a substantial role in decision-making under time pressure. The interpretation of the vulnerability scan results, correlation of the threat intelligence, and asset priorities has to be addressed by security analysts based on incomplete information and conflicting priorities in the operation of the assets. Research on security operations center (SOC) anthropology demonstrates that high-alert settings force analysts to rely more on intuition and heuristics than on the systematic assessment of problems, particularly when dealing with substantial amounts of notifications or ambiguous threat signals (Sundaramurthy et al., 2014a).

Heuristics like anchoring, availability and premature closure that come with such heuristics lead to a decrease in the level of accuracy of vulnerability prioritization. The SOC workflow analysis reveals that the performance of the analysts declines considerably with the concurrent tasks, which is worsened by the expansion of the enterprise and industrial networks in size and complexity (Gonzalez-Granadillo et al., 2021). According to the SANS 2023 SOC Survey, 21.8% of organizations have 2 to 10 analysts, and Vectra AI research reported a mean daily alert volume of 4,484 alerts per organization, 83.9% of which were false positives (SANS, 2023; Vectra AI, 2023). Poor quality of attention is as a result of alert fatigue and saturation which causes high risk vulnerability to be missed and low-risk concerns to be overreacted to. Studies of human-AI collaboration within the context of cybersecurity perceive this trend as an indicator of the reality that current vulnerability assessment technologies are unable to embody uncertainty or contextual variability due to the fact that the methodologies are based on human decision-making (Chamkar et al., 2022). This compels analysts to offset informational shortcomings with ad-hoc arguments, which cannot be applied uniformly or at scale.

It is these human constraints that are further enhanced by coordination bottlenecks which would incorporate security decision-making within workflow processes that involve the input of many stakeholders and observance of formal change management processes. Risk mitigation demands collaboration between security teams, system administrators, application owners, compliance staff and external vendors in most organizations. These groups possess their own priorities, incentives, and constraints, and it is not easy to work out remediation strategies in a brief period of time.

The Log4Shell attack is an example of this on a grand scale: one federal agency spent 33,000 hours on the response, and 52 percent of the surveyed organizations needed weeks to a

month to fix the issue (Sharadin, 2022). Studies on patch management in the healthcare sector reported that preliminary vulnerability triage requires 60 to 105 minutes per vulnerability in a sequence of actions such as alert triage, asset correlation, impact assessment, and cross-functional coordination (Dissanayake et al., 2022). Observations of SOC operations revealed that delays in inter-team coordination were one of the most common reasons slowing the process of vulnerability remediation, even though sufficient vulnerability knowledge and technically sound remediation mechanisms were available (Sundaramurthy et al., 2016). According to ethnographic research, the discrepancy between the technical and non-technical parties is a common issue that causes the various interpretations of risk resulting in a slow or fragmented response (Sundaramurthy et al., 2016). In regulated industries, formal approval processes introduce verification levels aimed at rendering operational safety/auditability but may prolong remediation schedules beyond operationally safe operation in a high-velocity threat environment. The dependency on vendors introduces additional bottlenecks as the organizations are not able to deploy any patches until certain updates are provided, especially to equipment manufacturers, industrial control systems, and proprietary platforms.

Such bottlenecks reveal a disparity between the rate of adversarial action and the processes inbuilt into the governance of organizations. The complexity of coordination and not technical difficulty determines remediation timelines which lead to the necessity of architectures to minimize or eliminate these dependencies in time-sensitive scenarios.

The concept of temporal impossibility is a quantitative measure of structural incompatibility of attacker speed, and the response capability of a defender, particularly when a large-scale validation, testing, and coordination is necessary. Empirical studies in the industry have consistently revealed that exploitation occurs rapidly: VulnCheck discovered 28 percent of

vulnerabilities were exploited within 24 hours of disclosure (Garrity, 2025); Qualys reported 25 percent of vulnerabilities exploited upon release (Qualys, 2024a); Mandiant reported the median time-to-exploit decreasing to 5 days in 2023 (Kutscher, 2024); and Rapid7 reported a median of 1 day to exploit when exploit code is available. On the other hand, businesses, whose operations are safety-sensitive or reliability-sensitive, are likely to issue patches in days or weeks due to the need to conduct regression testing, assess operational risks, certify vendors, and deploy in stages (Stouffer et al., 2023). According to the Edgescan 2024 Vulnerability Statistics Report, the average time to remediate critical vulnerabilities takes 35-65 days between organization types. Even though organizations operate within the right procedural boundaries, the time spent on safe remediation is more than the time in which one can prevent the exploitation of a significant part of high-profile vulnerabilities (Householder et al., 2017). Since such delays are based on operational and regulatory needs and not inefficiency, the delay cannot be closed through any process optimization (Bilge & Dumitras, 2012). There is also a case of asymmetric advantage, which is that attackers only need one vulnerability to attack and attack them as soon as they detect it, but defenders must make sure that their responses will not create collateral disruption, safety risk and compliance failure. Temporal impossibility is therefore a structural constraint, rather than a performance failure which explains the necessity of protection mechanisms that can continue to work under uncertainty without leading to operational downtime.

2.9 Synthesis and Gap Analysis

The literature synthesis of the vulnerability management, machine learning, and cybersecurity architecture indicates alignment of structural constraints that characterize current defensive capabilities. Scoring systems do not have environmental context, machine learning models are run on small feature spaces that do not take into account variables specific to

deployment and these constraints bring with them a natural amount of uncertainty which cannot be minimized by model refinement. Learning curves plateau after 50,000 to 100,000 training examples despite availability of over 279,000 CVEs, confirming that additional data without additional contextual information provides diminishing returns.

At the same time, literature on adaptive and zero trust architecture reveals that although dynamic enforcement and conditional access are principles of contemporary security paradigms, there are no mechanisms to map prediction uncertainty to graded protection actions. The lack of robust uncertainty quantification in these architectures limits their ability to adjust defensive posture based on the confidence of underlying assessments. Additionally, research on human-related constraints and organizational dynamics reveals that although the operator is provided with timely information, coordination bottlenecks, cognitive biases, and the overhead of procedures delay the operational reactions to the emerging threats (Sundaramurthy et al., 2014b; Sundaramurthy et al., 2016). All these findings lead to the conclusion that the existing paradigms of vulnerability management lack conceptual and operational capability to respond to threat circumstances that are uncertain, partial, as well as dynamic.

The conceptual environment suggests a new framework is required to address gaps that are present in existing systems and which are addressed in a partial and limited manner. First of all, no model described in the literature has included the concept of machine learning uncertainty into automated or semi-automated protection measures, although it is shown that uncertainty measures could be valuable to provide relevant information about the credibility of vulnerability forecasts (Lakshminarayanan et al., 2017). Second, none of the strategies tie prediction uncertainty to graded enforcement, and there are no tools that security teams can use to set proportionate constraints that minimize risk with minimal operational cost in cases where the

patches are slow or unsafe. Third, temporal self-correction has not been applied to vulnerability management yet, however, temporal adjustment of the level of confidence and enforcement has been shown to be effective in anomaly detection and behavioral analytics (Kruegel and Vigna, 2003; Hassan et al., 2019). Fourth, the literature on human factors describes bottlenecks of coordination and cognitive limits, but it does not investigate the acceptance and trust of uncertainty-driven automated protection systems by practitioners. Field validation of algorithmic decision support is still absent from vulnerability management research, and it remains uncertain whether technically feasible frameworks can lead to operational uptake in regulated conditions. Fifth, structures that integrate predictive analytics, uncertainty quantification, adaptive segmentation, and rollback logic into a single operation structure that executes at machine speed do not exist yet. Figure 7 traces how each literature-identified constraint specifies a required capability and how each GUARDIAN component directly responds to that need, transforming uncertainty from a coordination bottleneck into an automation parameter.

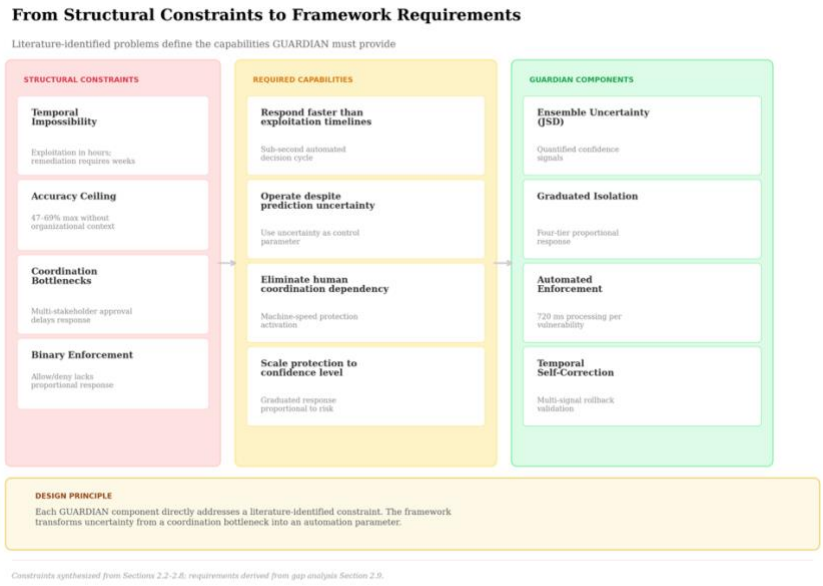

Figure 7. From Structural Constraints to Framework Requirements

GUARDIAN is suitably placed to address these gaps through its conceptualization of vulnerability response as an iterative process, whose elements are dependent on uncertainty, adaptive isolation, and temporal validation instead of binary choice founded on deterministic scoring. By making uncertainty a dominant organizing principle and relating it to a system of graduated protection procedures, GUARDIAN fosters a theoretical pattern that aligns defensive action with the confidence and incompleteness of the available evidence. This positioning is in direct response to structural constraints found across the literature and forms a baseline to apply uncertainty-aware adaptive defense in challenging and complicated environments.

CHAPTER 3: METHODOLOGY

3.1 Introduction

This chapter describes the methodology used to address the research objectives. It describes the steps guiding the development of the GUARDIAN framework to address the irreducible uncertainty and enhance system protection from vulnerabilities. A key part of this chapter is describing the practices followed in illustrating the accuracy levels of several machine learning models plateau during vulnerability prediction despite differing algorithmic composition. It also describes the approaches to determining the level of uncertainty across several machine learning models trained and validated for vulnerability prediction. To this end, the research method combines empirical analysis and simulation-based validation. Empirical analysis of historical vulnerabilities from the National Vulnerability Database provides the foundation for validation, and simulation serves as the feasibility test for the solution developed. The chapter covers the research philosophy, research design, data collection practices, and the instrumentation and simulation practices, alongside the statistical validation strategy applied in assessing the hypothesis that uncertainty itself may be used as a control parameter of automated defense in high-stakes critical infrastructure settings.

3.2 Research Philosophy

Research philosophies play a fundamental role in research since they establish the potential results of research processes, determining the research design and anticipating possible limitations ahead of the research (Davidavičienė, 2018). The role of a research philosophy in a study is to outline the theories explaining the reality being studied, and how that knowledge is

presented and justified (Mauthner, 2020). It is possible to choose among a number of research philosophies depending on the direction of the research, such as positivism, realism, interpretivism, post-positivism, and pragmatism, among many others (May, 1997). Among them, the most suitable one in the context of a research study would rely on the research questions, objectives, and problems that a research study is solving (Davidavičienė, 2018; May, 1997). Researchers must consider these philosophies before choosing methods since, as Mauthner (2020) points out, theories within these philosophies have a direct connection to particular approaches to data gathering and analysis.

The objective of this research study is to demonstrate that vulnerabilities exist objectively but the scores of their severity are subjective, and therefore uncertainty is created. The GUARDIAN framework is designed to deal with this irreducible uncertainty using a combination of multiple methods. With this in mind, a philosophy that argues multiple methods can reduce subjectivity and increase objectivity is essential. Pragmatism, realism, and post-positivism are the philosophies that deal with the mixed-methods dimension (May, 1997). Post-positivism is the most appropriate of the three. It states that quantitative calculations, being the only source of interpretation, may miss important issues and drawbacks in conclusions (Maksimovic & Evtimov, 2023). This is a key perspective in this study, since the researcher does not merely show that empirical findings illustrate irreducible uncertainty, but goes further and carries out a simulation to assess the feasibility of the proposed framework. The study also recognizes the possibility of biases in the empirical analysis and validates the framework by creating an artifact, testing its feasibility with the help of a simulation, and field-testing it with healthcare practitioners.

3.3 Research Design

The methodology in this study is to adopt a mixed-method design, which involves carrying out an empirical study (quantitative) of CVEs between January 2002 and December 2024 and then running a technical feasibility test (simulation) to explore uncertainty-parameterized adaptive isolation to manage critical infrastructure vulnerability. This research method is also known as design science research, and often is applied in information system research, which is concerned with the creation of new artifacts (in this case, the GUARDIAN framework) that can solve a previously unresolved problem in information systems, and more effectively than the earlier existing solutions or current solutions (Iivari, 2015; Venable et al., 2016). The design science research process is based on two main steps, where the first one is explaining the problem prevalent within these systems, followed by proposing a solution and creating a prototype or a simulation capable of resolving the specified issue (Iivari, 2015; Venable et al., 2016). In this study, the goal is similar, with three major phases: the empirical validation step, the simulation validation step, and the field validation step, while addressing the five research questions.

The research design deals with the research questions in a manner of five interrelated yet separate mechanisms. First, the weaknesses in the accuracy rates of various machine learning models to predict vulnerability are illustrated with the aid of an empirical analysis elaborating the information constraints related to organizational severity indicators of CVE descriptions leading to these constraints. Secondly, to estimate the level of uncertainty among diverse machine learning models, having been trained and tested on vulnerability prediction, the study employs statistical validation of ensemble disagreement to approximate the level of uncertainty, while Jensen-Shannon divergence is the primary indicator of gauging confidence. Third, to define the right threshold parameters of the graduated protection, the research applies empirical analysis of

the industry exploitation timing data to balance the risk reduction and the availability of operation. Fourth, to determine the feasibility of the graduated protection framework, the dissertation conducts simulation-based testing based on a cyber-physical water treatment testbed to determine the extent of risk mitigation and rollback precision under simulated attack scenarios. Fifth, to establish the acceptability of uncertainty-parameterized isolation recommendations by critical infrastructure practitioners, the study conducts field validation among practitioners of a healthcare critical infrastructure organization.

3.3.1 Empirical Validation and Statistical Analysis

This is the first phase of the research design, focused on statistically illustrating that the limited information in CVE descriptions does impact the uncertainty levels and severity scores that organizations assign. Research studies such as Holm and Afridi (2015) have previously shown the existing disagreement between what the CVSS system ranked as the severity of a security vulnerability and what experts ranked, showing a -0.38 mean disagreement score. On the same note, another study, Spring et al. (2021), also reports that the existing CVSS formula of vulnerability ranking and scoring has operational issues and fails to consider the important outcomes of the security risks once they have been ranked. However, these prior studies were limited in scope: Holm and Afridi (2015) examined approximately 3,000 vulnerabilities, and Spring et al. (2021) remained largely theoretical. This research takes a different approach, as it seeks to examine 279,056 CVEs in the National Vulnerability Database (NVD). The NVD is a vulnerability database that provides summaries for all CVE vulnerabilities that are publicly known and provides analysis tools for vulnerability data (Mell, 2005; NIST, 2025). By using such a large dataset, made publicly available by the US government, this study can statistically validate the core mechanisms of this research.

The dataset was strictly split into three partitions using temporal partitioning to ensure that there is no leakage from future data during the training, validation, and testing processes. This strict temporal split is important because research has shown that when the models can access future information used for training, validation, and testing, it affects the accuracy levels of the results, sometimes inflating them by as much as 50% (Pendlebury et al., 2019). Data from January 2002 to December 2020 is used as the training dataset and contains 160,543 CVEs. Data from January 2021 to December 2022 is used as the validation dataset and contains 49,802 CVEs, and data from January 2023 to December 2024 contains 68,711 CVEs and is used as the testing dataset. This large test dataset ensured sufficient statistical power to detect medium-sized effects at $\alpha = 0.05$. This data was employed to compare the accuracy of 6 machine learning models with various mathematical strategies: Random forest, Support vector machine, Logistic regression, Neural network, LightGBM, and XGBoost.

Having measured the accuracy levels of the models, the research then measured the ensemble disagreement using Jensen-Shannon divergence to measure the divergence in predictions between models. This was due to the symmetric nature of the Jensen-Shannon divergence and bounded properties which make such a divergence the most appropriate where smooth finite measures of distance between probability distributions are required as in the case of the present study (Nielsen, 2019). This was then accompanied by a test of how stable the patterns of disagreement were across the 23-year dataset. Model performance was also assessed across Common Weakness Enumeration (CWE) categories to verify consistency across vulnerability types. The ensemble also underwent calibration testing using the Expected Calibration Error to assess alignment between predicted confidence and observed accuracy (Guo et al., 2017). Lastly, the correlation between the ensemble disagreement and the prediction error

was analyzed using Spearman rank correlation, and it was also essential to examine which instances of high disagreement coincided with unreliable predictions, confirming disagreement as an effective predictor of untrustworthy assessments (Zar, 2010).

Following this analysis, the study then proceeded to validate the graduated protection framework by examining the monotonic relationships between uncertainty levels and response parameters. The relationship between uncertainty and risk of exploitation, as well as between uncertainty and the time of protection, were determined using Spearman rank correlation (Hollander et al., 2013; Zar, 2010).

3.3.2 Simulation Validation

Research shows that high-fidelity simulation environments are the gateways to addressing the limitations in real-world environments that make it impossible to conduct testing, especially with large-scale data, and the need for controlled environments where diverse scenarios can be explored (Holm et al., 2015). Simulation guarantees that a realistic assessment of a real-life situation has been investigated as thoroughly as possible without creating risks in the real world. This is particularly important when it comes to critical infrastructure systems where live testing in case of failure may cripple the system in the same way that an attack would. A case in point is the cybersecurity attack on the Water Treatment system in Oldsmar, US in 2021, in which the attackers intended to inject the Water Treatment system with sodium hydroxide, which, when consumed, can inflict painful burns and in some cases, permanent damage (Cervini et al., 2022). In such a scenario, conducting live testing of the solution instead of a simulation would mean that, in the event that the patch is unsuccessful, then this critical infrastructure has to be cut off from the people to avoid harm to their physical safety and prioritize public health (Morris et al., 2011).

Given these risks, researchers have been actively investigating high-fidelity simulation as one of the options to solve those challenges and vulnerabilities in CIs, be it the detection process, or testing the effect of developed solutions/artifacts (Morris et al., 2011; Holm et al., 2015). Besides this, other authorities, like the Nuclear Regulatory Commission (NRC, 2021), have prescribed that any cybersecurity initiative, including testing and control measures, must under no circumstances impair the safety, security, and emergency preparedness of CIs, particularly nuclear facilities. In line with these requirements and regulations, this study follows a simulation-based validation approach.

The simulation environment for this study used EPANET 2.2 by the EPA for hydraulic modeling (Rossman, 2000) and the MITRE ATT&CK for industrial control systems framework to emulate the security threats on the system, based on what real-world risks and industrial incidents would look like (Alexander et al., 2020). The architecture used to represent a water treatment plant, which is the subject of the simulation, uses 15 programmable logic controllers (PLCs). These were distributed across a four-tier network architecture as is recommended in municipal utility architecture, especially one focused on water supply (Halfawy, 2008). Together, these characteristics ensure that realistic conditions of a water-treatment plant are prioritized without undergoing the dangers of live-testing.

The simulation testing included 1000 vulnerability scenarios focused on assessing the trade-offs between system protection and operational availability, conducting a total of 6000 simulated attack scenarios. This is based on the 12 attack variants testing defined by MITRE ATT&CK taxonomy (Alexander et al., 2020), which were conducted across five isolation levels to test graduated responses from the baseline level to full isolation, with 100 trials per variant ($12 \times 5 \times 100 = 6000$ simulated attack trials). The vastness of this testing increases the statistical

power of the research and the likelihood of detecting medium-sized effects, considering Cohen's standard for medium effect size d = 0.50 (Jané et al., 2024). The attacks tested varied to match the various techniques applied in real-life attacks and included standard attacks which follow normal documentation exploitation practices, stealthy reconnaissance which focuses on lateral movement without detection, insider threats, credential-based attacks, aggressive system compromise, system/protocol exploits, supply chain attacks, ransomware pattern attacks, and zero-day exploitations.

There were several simulation boundaries and operational requirements that characterized this process, and the distinction between technical feasibility and operational readiness was clearly established. First, there were the signal quality specifications. To ensure the study is realistic and at the same time recognizes the weaknesses of a simulated environment against a real one like possible degradation of signal quality, the intrusion detection system accuracy is set at 95 percent true positive, and threat intelligence integration is set at 90 percent coverage capability. These simulation parameters represent idealized conditions; real production environments typically achieve lower signal quality, which organizations must account for when calibrating the framework for deployment. The production intrusion detection system (IDS) is estimated to have 70-80% effectiveness (Sommer & Paxson, 2010), and available organizational intelligence during deployments is estimated at 60-75% effectiveness due to budget constraints and filtering noise in the system (Tounsi & Rais, 2018). Patch tracking was modeled with perfect visibility (100%), while production environments typically achieve 85-90% accuracy due to network segmentation resulting in limited comprehensive visibility, delays in deployment timing, delays from system updates, and other practices such as database synchronization, and the need to manually confirm offline systems (Ginter, 2019; Stouffer et al., 2023). While these

percentages are estimations defined by the researcher, they are informed by the aspects defined by literature as key aspects of signal quality in simulated environments.

Under these simulation conditions, temporal self-correction and rollback execution achieved 82-94% accuracy. This represents a performance ceiling; production environments with lower signal quality would likely achieve reduced accuracy.

3.3.3 Field Validation Design

Research focused on critical infrastructure environments faces other constraints beyond the research practices themselves, because often there are legal requirements that define the way operations on these environments can be conducted. Experimental testing in these environments is not permitted, especially because they exist as live systems providing critical operations. For instance, operations in healthcare facilities, which are part of critical infrastructure environments, are defined by strict regulations such as the Health Insurance Portability and Accountability Act (HIPAA) and the Food and Drug Administration (FDA) medical device regulations (Edemekong et al., 2018; FDA, 2025). The regulations under these two bodies do not permit the use of experimental software as it could potentially disrupt patient care which is their primary role. Likewise, nuclear plants also possess the same restrictions, where the protection of digital systems is highly regulated under the 10 CFR 73.54 federal regulation (NRC, 2021). This is similarly the case with water facilities, where the Environmental Protection Agency (EPA) requires network isolation for experimental research and considers uncontrolled testing an unacceptable risk to public health (EPA, 2025). While these are evident constraints in research, they are also key stringent measures put in place to protect people reliant on these systems from any harm. As such, research has to work around the constraints.

To address these research constraints, this study applies a three-phase validation approach. The first phase is empirical analysis, which establishes the foundational evidence by analyzing 279,056 CVEs to validate that ensemble disagreement provides a calibrated uncertainty signal. The second phase is simulation, which is used for rigorous end-to-end testing of simulated versions of critical infrastructure systems, avoiding the safety risks on actual systems. This simulation approach develops water treatment testbed models with architecture as realistic as possible, without posing any challenges to the actual system, and enables testing that is otherwise prohibited in the actual system to be done. The third phase is field validation, which facilitates practitioner acceptance testing to assess whether practitioners would recommend deploying these components in their organizations. Field validation focuses on including a human element in the evaluation and implementation of such solutions, prioritizing minimal risk of the software overriding other systems operations and performing autonomous acts. Through this approach, the GUARDIAN framework is validated from the empirical, technical, and practitioner acceptance perspectives. Figure 8 illustrates this three-phase validation framework, showing how each phase addresses questions the others cannot and how each maps to specific research questions.

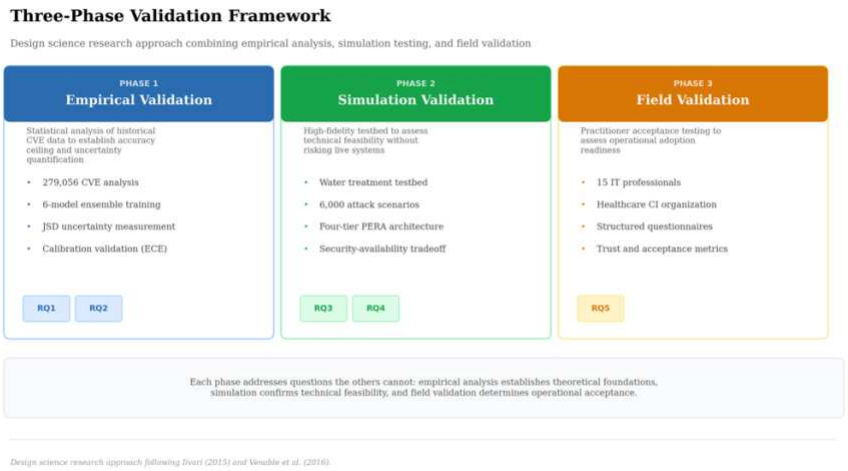

Figure 8. Three-Phase Validation Framework

These three methods complement each other: empirical analysis establishes whether uncertainty quantification is mathematically sound, simulation determines whether automated enforcement is technically feasible, and field validation assesses whether practitioners would trust and adopt the framework operationally.

The field validation study involved surveying 15 IT professionals working in a healthcare critical infrastructure organization, who were identified for this study through their direct involvement in vulnerability management and system protection decisions. Structured questionnaires were used as the data collection method including items on agreement with framework recommendations and trust in automated and human-supervised isolation decisions. The results from the participants were evaluated to determine the rate of operational acceptance and the factors affecting the level of practitioner trust in uncertainty-parameterized automation. The complete survey instrument is provided in Appendix A.

3.4 Data Collection

The research study is largely founded on historical data from CVEs to establish the relationship between descriptive ambiguity and assessment uncertainty. Machine learning analysis can only be as good as the data used in it. The quality of data in this study is determined by completeness and preprocessing practices applied to the dataset. Maharana et al. (2022) emphasize that the performance of machine learning models is dictated by the diversity, quantity, and quality of data used in any research or training process, and the reliability of an algorithm may be improved by selecting data in the original dataset. On these grounds, this section describes the process of acquiring the dataset, temporal partitioning to avoid experimental bias

and data leakage between the three phases, and the presentation of organizational context profiles.

3.4.1 Dataset Acquisition Process and Temporal Partitioning

As underlined in the previous discussions, the data was collected through the NVD. It includes 279,056 distinct CVEs with the descriptions and the CVSS severity vectors applied in this research and analysis, spanning the years between the first CVEs added in January 2002 and December 31st, 2024. The collected data in the research is the full population of CVEs rather than a sub-population, which leaves the issue of selection bias out of the question, though completeness of the description of respective CVEs remains variable (Anwar et al., 2021; Zhang et al., 2011). The use of the full dataset reflects the complexity that can be observed in actual intelligence systems, even though data prior to 2005 could be described as volatile due to inconsistent reporting practices (Bridges et al., 2013; NIST, n.d.; Zhang et al., 2011).

A natural language processing (NLP) pipeline and feature extraction was used to extract and process data, and the process involved four main steps. The first was text normalization, which implemented lowercase changes, elimination of special characters, and stopword filtering to decrease noise in the data (Jurafsky & Martin, 2013). This was followed by tokenization, in which the described vulnerabilities were separated into meaningful units, and Named Entity Recognition (NER), which identifies and categorizes entities into predefined categories from unstructured text, was utilized to extract meaningful information about vendors, products, and versions (Bridges et al., 2013). Security-specific vocabulary was isolated through technical term identification. The third step was vectorization, which is part of the process of numerically recording the distinctive features of a text document to render it simple to be handled by systems. This was performed with the help of Term Frequency-Inverse Document Frequency (TF-IDF)

with a vocabulary of 5,000 dimensions (Manning et al., 2008). The final step was N-gram extraction, which is necessary to promote contextual knowledge and was used to retain semantic context and phrase patterns up to trigrams (Jurafsky & Martin, 2013).

The temporal partitioning strategy discussed above was then followed, with models only accessing data within their designated time frame to prevent data leakage. Literature sources such as Pendlebury et al. (2019) discuss the significance of chronological and temporal partitioning, asserting that machine learning model accuracy is seriously influenced by improper time splits and partitioning of testing and training data. The partitions were divided according to Section 3.3.1: training data from January 2002 to December 2020 (160,543 CVEs, 57.5%), validation data from January 2021 to December 2022 (49,802 CVEs, 17.8%), and testing data from January 2023 to December 2024 (68,711 CVEs, 24.6%). Figure 9 illustrates this temporal partitioning strategy, showing how the strict chronological separation prevents future information from influencing model training and ensures unbiased evaluation.

Figure 9. Temporal Partitioning Strategy

3.4.2 Organizational Profile Data and Accompanying Constraints

Apart from the CVEs dataset, other data critical to this research is the profiles and operational contexts of organizations operating in critical infrastructure sectors that can be used to quantify the regulatory and resource constraints that necessitate automated isolation. Four key critical infrastructure sectors, water utilities, electric utilities, nuclear power generation, and chemical manufacturing, are discussed to illustrate this operational context in this section and define the considerations in setting the simulation parameters.

Water Utilities. Water utilities are one of the most critical infrastructures, requiring high-level cybersecurity systems (Scalco & Simske, 2025). Research by the Water Information Sharing and Analysis Center (WaterISAC) in 2021 showed that as the number of utilities in the system increased, the number of full-time equivalent (FTE) security specialists needed to address the cybersecurity needs increased as well (Water Sector Coordinating Council, 2021). The same study noted that in 73% of the total surveyed utilities (606 utilities), fewer than three FTEs were dedicated to ensuring operational technology security (Water Sector Coordinating Council, June 2021). Considering that water utilities are heavily dependent on and defined by OT systems such as Supervisory Control and Data Acquisition (SCADA) systems, PLCs, sensors, actuators, and human-machine interfaces (HMIs) to manage, monitor, control, and optimize water treatment and distribution processes, a shortage in FTEs is a challenge with long-term impacts (Stouffer et al., 2015). Over time, this scenario has led to a resource constraint, such that by 2025, more than 70% of surveyed water systems were considered not to be meeting the EPA cyber standards, such as lacking complete OT asset inventories, indicating that the industry is in a potential crisis mode (DiMolfetta, 2024; SIGA, 2025). The results of limited resources to facilitate these operations are system bottlenecks where the necessary cybersecurity assessment practices are being overlooked in the system (Ribeiro, 2024; SIGA, 2025). Combined with lengthy patch

validation and verification requirements (Souppaya & Scarfone, 2022), the absence of definitive timelines creates latency in patch deployment.

Regional Electric Utilities. More than 400 million people in North America are being served by regional electric utilities, and this infrastructure is of critical importance (NERC, 2019, 2025). The number of citizens served demands an effective system of risk assessment and risk management particularly those concerning cybersecurity risks, and cybersecurity attacks, given the multiple systems involved. However, a recent report similarly showed that smaller utility companies do not have the necessary staffing capabilities to address cybersecurity issues, instead hiring external support to address system bottlenecks in assessment (Ribeiro, 2025a). With the 35-day patch windows and the need to monitor, manage, and maintain a large system, the presence of over 30% of utility companies using external experts to manage their OT systems illustrates the complexity of the system, the limited organizational capacity, and the possible risks these limitations could pose for organizations (Ribeiro, 2025a).

Chemical Manufacturing (Multi-Plant Enterprises). Chemical manufacturing is another critical infrastructure in the US economy that includes large organizations and enterprises, which need a high level of cybersecurity risk assessment and security systems. While emphasizing the importance of a risk assessment and cybersecurity management approach in the chemical manufacturing sector, CISA (n.d.) notes that chemicals are integrated into every critical infrastructure sector; hence, disruption to these systems would significantly impact the economy and public safety. Despite this, there is still evidence of vulnerability of chemical facilities across the US, citing reliance on outdated cybersecurity guidelines, and a slow focus on updating the systems to mirror the newer state of cybersecurity protection systems recommended, to match the rising risks (XcelPros, n.d.). Some of the cited issues are overreliance on IT staff for support

even in OT systems, which has left a large portion of the chemical company systems unattended (Ribeiro, 2022). While it is unclear whether this is a staff or budgeting issue, it is evident that the chemical facilities do need more manpower, and as it stands, considering the lack of definitive regulation on patching verification and validation timelines, and the laxity in securing the OT systems even after a decade of changes as of 2021, there exist both resource and regulatory constraints in this sector (XcelPros, n.d.; Ribeiro, 2022).

Nuclear Power Generation Facilities. Nuclear power generation facilities are some of the most significant utilities in energy production globally. According to the Nuclear Sector Coordinating Council (n.d.), nuclear reactors provide over 20% of the electricity consumed in the US, and with the rising number of cybersecurity attacks in the country, ensuring these systems are heavily protected is key. The systems needing securing include the industrial control systems at the nuclear plants, the hardware systems at the data servers, and the operational data at the management stations (Nuclear Sector Coordinating Council, n.d.). The protection of these systems is a key necessity, especially now when it is reported that cyberterrorists are targeting these facilities as part of geopolitical tensions and shifting national relations (Ribeiro, 2025b). The American government has made efforts towards ensuring this security, setting up regulatory bodies and agencies such as the Department of Energy, Office of Nuclear Energy's (DOE-NE), and US Nuclear Regulatory Commission, to regulate these systems and ensure there are cybersecurity protocols in place to ensure safety of the people and the country (Nuclear Sector Coordinating Council, n.d.). Associated recommendations include developing dedicated security teams. Given the complexity of nuclear infrastructure, substantial staffing is required, and regulatory timelines for implementing cybersecurity solutions, including patch validations, are more stringent than in other sectors.

Across these four sectors, common constraints emerge: limited cybersecurity staffing, reliance on external support, lengthy patch validation timelines, and regulatory requirements that prohibit rapid system changes. These constraints informed the simulation parameters, particularly the signal quality assumptions (reflecting limited monitoring capacity) and the temporal thresholds (reflecting regulatory patch validation windows of 35-70 days).

3.4.3 Arising Temporal Impossibility in Critical Infrastructure Organizational Operations

There is a noticeable consistency in the demand for more staff to meet the required cybersecurity mandates, the lack of follow-up in some of the regulatory implementations, and the gap in timelines for patch validation across these four sectors. These limitations in regulatory oversight and staffing illustrate the risks that would exist in the event of an attack. Considering the timelines to implement patches following a vulnerability in these ICSs, the ability to conduct a rapid response, especially when addressing a high-severity vulnerability, is practically impossible. This temporal impossibility illustrates the necessity for an uncertainty-parameterized adaptive isolation architecture, where ensemble disagreement would operate as an uncertainty signal modulating isolation stringency and duration, instead of waiting for human coordination after a vulnerability is disclosed, which creates delays that cyber attackers exploit. When systems detect higher uncertainty levels, they trigger protective measures before the actual attack, resulting in shorter response durations, while in cases where uncertainty exists at lower levels, the isolation is applied at reduced intensity with longer windows before rollback. This way, the staff and cybersecurity specialists in these organizations have time to coordinate a response while the system remains protected, without jeopardizing regulatory compliance.

3.5 Experimental Setup

As mentioned earlier, six machine learning models were used for the experimentation process to assess uncertainty using the CVEs dataset.

3.5.1 Model Configurations

The six models have made different mathematical assumptions to guarantee that any observed ceiling of accuracy is due to information limitations and not algorithmic bias of similar architectures.

Random Forest. A prediction and classification model in machine learning is a Random Forest model constructed based on a set of decision trees (Breiman, 2001; Schonlau and Zou, 2020). This experiment applied the random forest classifier with bootstrap sampling to create a 100-tree model and reduce overfitting (Schonlau & Zou, 2020). The high dimensionality of the text features between the trees was controlled by using a maximum depth of 20 levels and a minimum sample per leaf of 5 to validate a split. Additionally, to make sure that no interdependence existed among the trees; in selecting the features at any given node, the set of features was the square root of the total features (Breiman, 2001).

Support Vector Machine (SVM). The most important features of SVM are the maximization of margins to find the optimal boundaries, processing high-dimensional data, and memory efficiency, which are the primary reasons why SVM is primarily applied in data classification and regression problems (Pisner and Schnyer, 2020; Shmilovici, 2005). In this experiment, the SVM classifier was trained on the Radial Basis Function (RBF) kernel, which projects the input vectors into the high-dimensional space and works with non-linearly separable data (Cortes and Vapnik, 1995). The regularization parameter C was set to 1, which is common to balance between minimization of classification error and maximization of geometric margin,

and the Kernel Coefficient (γ) was set to use a scale heuristic so that it can adapt automatically to changes in the input features (Cortes & Vapnik, 1995; Pisner and Schnyer, 2020).

Logistic Regression. Logistic regression is a classification algorithm, as it can be applied to map a feature space function to a 0-1 value (Bisong, 2019). Elastic net regularization was used to implement logistic regression in this study, to provide a linear probabilistic baseline, particularly given the data nature, and the necessity to combine the strength of Lasso (L1) and Ridge (L2) penalties in the selection of features (Torang et al., 2019). The mixing parameter was adjusted to equalize the two penalties at $\alpha = 0.5$. The algorithm was also set to reach an iteration limit of 1000 to make sure the algorithm arrives at the loss surface.

Neural Network. The neural network, the multilayer perceptron, usually used in supervised learning to model data to the required output, was used because it had the ability to learn non-linear relations that are complex by using the backpropagation algorithm among other optimization algorithms (Gardner and Dorling, 1998; Popescu et al., 2009). The two hidden layers, in this instance, consisted of 128 neurons and 64 neurons and the neurons were configured to use the rectified linear unit activation function that was able to overcome the issue of vanishing gradient (Vargas et al., 2021). In order to train the neural network, the overfitting was reduced by avoiding co-adaptation via 0.3 dropout regularization of both the hidden layers (Hinton et al., 2012). The Adam optimizer was used to train the network, and the initial learning rate was 0.001 (Kingma & Ba, 2014).

LightGBM. LightGBM is an algorithm used most frequently in prediction due to its high precision, efficiency and increased scalability and speed compared to other decision tree gradient boosting algorithms (Ke et al., 2017). The LightGBM classifier used in the study utilized the leaf-wise tree growth algorithm and was set to use 200 estimators and a learning rate of 0.05 (Ke

et al., 2017). The feature fraction was adjusted to 0.9 and the bagging fraction was adjusted to 0.8.

XGBoost. XGBoost is another model that is commonly used in predictive modeling due to the high level of its precision, which is achieved by avoiding the issue of overfitting during the algorithm training, as well as the loss function and regularization during the learning process (Ahmetoglu and Das, 2022; Chen and Guestrin, 2016). In this case, XGBoost was configured with 100 estimators and a learning rate of 0.1, and the tree depth was capped at 6. The column sampling ratio was adjusted to 0.8 and subsample ratio was adjusted to 0.8 to make sure that every tree is trained on a random subsample of the data and features of the dataset. The six-model ensemble architecture is summarized in Figure 10 and demonstrates how the different mathematical underpinnings can be used to ensure that disagreement is due to a real data ambiguity and not an artifact of the algorithm.

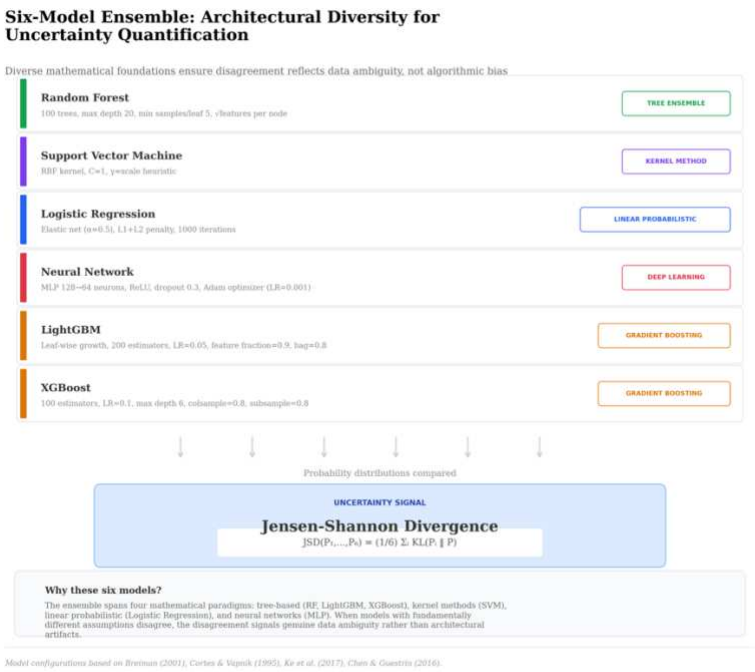

Figure 10. Six-Model Ensemble Architecture

3.5.2 Mathematical Formulation of Uncertainty Signals and Parameter Mapping

Uncertainty Quantification. The above uncertainty was quantified using the Jensen-Shannon Divergence to determine the difference between the distribution of probability predictions by each model/algorithm separately, and then, by the six models. The generalization of Theorem 1 on directed divergence implemented by Lin (1991) and Kullback and Leibler (1951) to an ensemble, in which each model generates a probability distribution P and the ensemble consensus is $\dot{P}$, gives: $JSD(P_1, ..., P_m) = (1/m)\Sigma_{i=1}^{m} KL(P_i||\dot{P})$, where KL represents the Kullback-Leibler divergence. When applied to this ensemble of six, the equation becomes: $JSD(P_1, ..., P_6) = (1/6)\Sigma_{i=1}^{6} KL(P_i||\dot{P})$. The JSD score will be zero or will increase based on whether all the models agree on the severity of a vulnerability or disagree. The rationale of employing the metric to measure the uncertainty disagreement is that it is symmetric, treating all models as equal rather than some as deviants of others, and that, as a bounded metric, it is mathematically assured to yield a score within a set numeric scale (Lin, 1991; Nielsen, 2019). This eliminates the possibility of the signal varying beyond what a control system can read consistently. The use of Jensen-Shannon Divergence instead of the other metrics is a choice that has both theoretical and empirical reasoning. Tang et al. (2006) developed ensemble diversity measures as mathematically related to classifier margin, showing correlation coefficient $\rho = -0.97$ between accuracy and disagreement terms, which provides theoretical basis for using disagreement as uncertainty signal. Furthermore, research on ensemble diversity indicates that bounded metrics correlate more reliably with calibration error in deep ensembles, ensuring that the uncertainty signal remains interpretable across different model architectures (Lakshminarayanan et al., 2017; Kuncheva & Whitaker, 2003).

Signal Calibration and Parameter Mapping. The uncertainty signal was calibrated before use to make sure that the predicted confidence levels matched empirical accuracy, since at times raw machine learning models tend to be overconfident, assigning scores of high probabilities, like 99% accuracy, when they are actually wrong (Wang, 2023; Guo et al., 2017). Although a variety of scaling techniques exist, temperature scaling was selected in this instance because it is low-complexity and efficient, requiring only a scaling of the logits with a learned scalar, τ, before the raw prediction values are transformed into final probabilities ($\acute{p}$) with the help of the softmax function (Wang, 2023). This can be expressed as: $\acute{p} = \sigma(z/\tau)$, where σ is the softmax function, and z is the raw prediction values/logits. Grid search optimization was used to determine the value of τ to be 1.47. Such a method is most applicable to this study since it has been demonstrated that it does not have an impact on the accuracy of the model (Guo et al., 2017). The accuracy was also compared with the predicted confidence levels with other quality assessment calibrations such as the expected calibration error and the maximum calibration error to determine areas of problematic confidence (Guo et al., 2017).

The uncertainty scores were then converted to simulation actions after calibration through a system of rules that established the level of isolation and the time of isolation. With regards to the isolation levels, there were certain thresholds to be applied in the validation stage, which are presented below using a step function rule: If $JSD < 0.35$: This is level 0-1 and system does nothing. If $0.35 \leq JSD < 0.45$: This is level 2. If $0.45 \leq JSD < 0.60$: This is level 3. If $JSD \geq 0.60$: This is level 4, indicating high uncertainty. Levels 2 and 3 activate system protection for partial isolation, and level 4 activates system protection to enter full isolation mode.

For the duration, an inverse relationship calculation is used where higher uncertainty levels result in shorter initial conservative protection windows, while lower uncertainty levels

have more time, like the time needed in standard current reactive systems, to give time to fix the problem. The relationship is: Isolation Duration = Base × (1.5 – JSD score) × Severity modifier, where the base is 48 hours, and the severity modifier ranges from 0.5 to 2.0. The GUARDIAN framework generalizes this binary threshold to graduated thresholds that convert continuous JSD values to proportional protective responses. These threshold functions are visualized in Figure 11, with the JSD scores plotted against the levels of isolation, and the greater the uncertainty the stronger the protection with shorter duration.

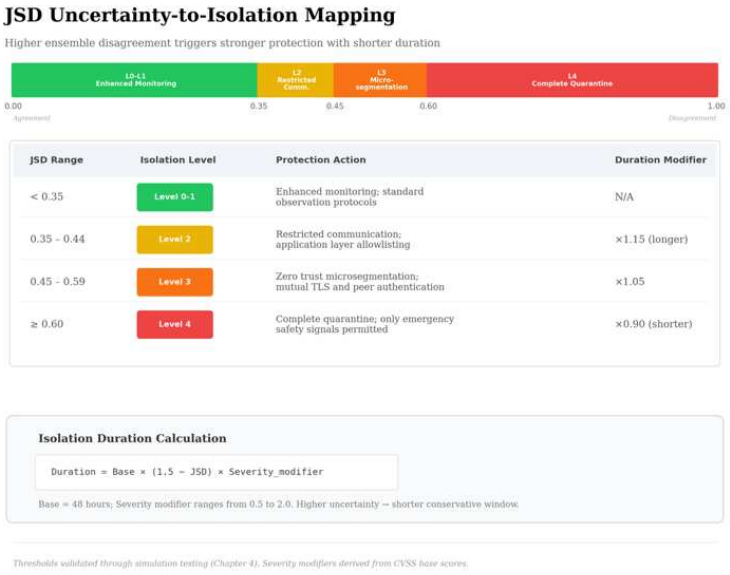

Figure 11. JSD Uncertainty-to-Isolation Mapping

3.5.3 Rollback Timing Design Rationale

The determinants of the temporal thresholds of protection rollback in this study are several research studies that had been conducted on real-world cybersecurity attacks previously. As mentioned above, there are ethical and safety limitations concerning critical systems of infrastructure, which make it impossible to do live testing on such systems, therefore, this study

relies on the empirical data provided by other organizations monitoring such incidents. The four significant research projects inform the chosen thresholds.

In a report published by Mandiant in 2023, the median time to exploit among the 138 exploited vulnerabilities analyzed was 5 days, and the first exploitation of the vulnerabilities was done within the first 72 hours (Charrier & Weiner, 2024). This is a spectacular outcome, especially considering past research where the median time was 32 days and it indicates that the level of attacks and exploitation has grown by a huge margin. In 2024, a report by VulnCheck on the tendencies of attacks analyzed 768 verified CVEs and the outcome revealed that there were two distinct peaks with 28 percent of the vulnerabilities being exploited in the first 24 hours with a focused mass targeting, and systems targeted with sophisticated campaigns had a significantly higher median of 192 days (Garrity, 2025). In a third study, published in 2023, Qualys found that in the 206 high-severity vulnerabilities they analyzed, 23% were exploited on the day of publication, with the median time of exploitation being 44 days (Qualys, 2023, 2024a). Finally, the fourth report, by Rapid7, which was done in 2022, revealed that commodity attacks with automated scanners hit in a period of about 48 hours, but others like persistent campaigns against critical infrastructure systems can take weeks to months (Condon et al., 2023).

Interestingly, the four reports concur on the timelines of attack and this is where the graduated thresholds used in designing the framework were based on. These thresholds were picked to be 24 hours, 48 hours, 7 days, 30 days to cover the shorter thresholds found above, and to also cover the longer and slower thresholds of sophisticated campaigns. The high uncertainty vulnerabilities are assigned shorter initial thresholds and the low-uncertainty vulnerabilities are assigned the long thresholds, and this is because of the inverse relationship between duration and uncertainty that was discussed in the previous section. However, these threshold parameters are

based on industry reports of real-world exploitation rather than controlled experimental research and therefore should be customized according to the industry and operational context, because industry-specific patterns of exploitation, operational constraints, and levels of risk tolerance all play a considerable role in the process. These four sources of the industry are synthesized in Figure 12, where the observed exploitation patterns were used in selecting the 24-hour, 48-hour, 7-day, and 30-day rollback thresholds.

Figure 12. Exploitation Timeline Evidence

3.6 Simulation Environment

After the above calculations, the next step involved testing the technical feasibility of the proposed GUARDIAN framework by considering realistic operational constraints, and that required the development of a high-fidelity cyber-physical testbed that could be used to simulate

a critical infrastructure organization. The testbed was created to simulate a municipal water treatment plant serving 50,000 individuals with an average daily capacity of 10 million gallons. As mentioned earlier, the architecture for the physical process layer was modeled using EPANET 2.2 by the EPA for hydraulic modeling (Rossman, 2000) and adopted a realistic hierarchy of industrial automation with 15 programmable logic controllers (PLCs). These were spread in a four tier network structure as is suggested in municipal utility architecture particularly the water supply oriented architecture (Halfawy, 2008). A combination of these features guarantees the realistic conditions of a water-treatment plant and eliminates the risks of live-testing. Figure 13 illustrates this four-tier network architecture, showing how the testbed implements defense-in-depth segmentation following the Purdue Enterprise Reference Architecture model.

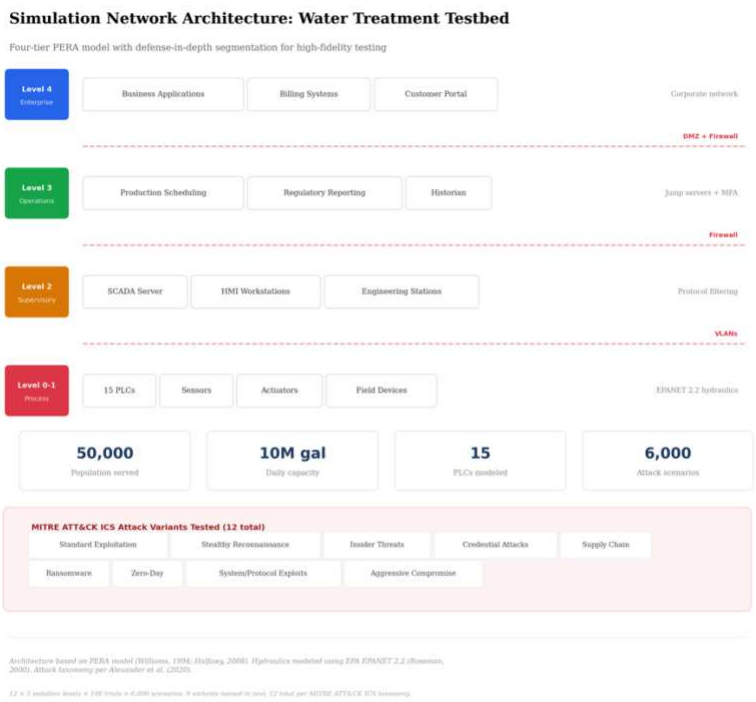

Figure 13. Simulation Network Architecture

The network topology is developed strictly in line with the Purdue Enterprise Reference Architecture (PERA) which is the defense-in-depth segmentation at four levels: Level 0-1, the process zone that contains basic control PLCs and field instrumentation, Level 2, the supervisory control zone that contains the SCADA visualization interface, Level 3, the operation management system that contains production scheduling and regulatory reporting, and Level 4, the enterprise zone that contains the business applications, billing and customer systems (Williams, 1994). Further, as per the PERA model, network segmentation was provided between the levels with demilitarized areas to isolate corporate and control networks, firewalls to restrict communications to necessary protocols and authorized destinations, virtual LANs to the process area to restrict lateral movement, and jump servers to provide controlled administrative access with multi-factor authentication, which was also the practice of municipal utility.

Furthermore, as mentioned earlier, the MITRE ATT&CK framework of industrial control systems was applied to simulate the security threats on the system, based on documented real-world risks and industrial incidents (Alexander et al., 2020). These scenarios reflected three major attack categories towards an organization namely initial access threats and attacks, lateral movement and persistence attacks and impact strategies.

3.7 Methodology Limitations

The first limitation of this methodology is the use of a simulation in this research. While research does show that simulations are integral in learning, testing, and experimentation, they cannot replace real-world deployment, particularly because human factors are difficult to simulate (Kavak et al., 2018). Human beings react differently in real time, especially when responding under stress-induced scenarios or when organizational dynamics are influencing their decision-making strategies, and a simulation does not accurately capture that, as the responses

considered in such environments are scripted. This and other factors including the sophistication of the attack, the nature of the network and model of physical processes used all offer a new platform of operational complexity that cannot be well represented by this research in a simulation. The outcomes that were discussed in this research are a result of the interpretation of the technical feasibility of the framework, assuming that these other factors are held constant. Depending on whether the framework is to be operationalized, more simulations and perhaps partial testing would be required to address the complexities within real environments other than the technical recreation.

Second, the data employed in the study carries risk of bias as it is a reflection of reported vulnerabilities and not every vulnerability. As the framework was developed on the basis of these vulnerabilities, it is arguably weak against other undisclosed vulnerabilities, meaning organizations must retain other defenses in case they are hit by such unknowns.

Third, the field validation aspect determines how practitioners accept the framework recommendations, and not actual deployment results. This design accounts for regulatory limitations that prohibit the automation of experiments in live critical infrastructure and captures the human factors aspect that must be considered for operational adoption.

3.8 Ethical Considerations

This research observes ethical research practices while using historical data and simulation as the key design strategy. This is evident in the fact that data is obtained from a database focused on enhancing cybersecurity, so key aspects such as organizational or personally identifiable information that should normally be anonymized or kept confidential in research have already been abstracted. Additionally, throughout the methodology, the research practices

are clearly defined to provide opportunity for replication and exploration, while still ensuring the data is publicly accessible, thus preventing opportunities for misuse.

3.9 Chapter Summary

This chapter describes the empirical analysis and simulation strategies employed in the development of a graduated framework to address the irreducible uncertainty and enhance system protection from vulnerabilities. It describes the research design towards development of the artifact (the GUARDIAN framework), describes how data is collected from the NVD database, how it is temporally partitioned to avoid data leakage, and the organizational profiles that offer contexts on constraints in resources and regulations faced by organizations in critical infrastructure sectors. It also describes the experimental setup, documenting the configuration practices across the six models, the strategy used to quantify uncertainty, the signal calibration processes, and parameter mapping values, towards defining the graduated levels of isolation for security protection upon detection of various levels of uncertainty. Additionally, the chapter has described how the simulation environment was modeled, the field validation design for assessing practitioner acceptance, the methodology limitations, and the ethical considerations integral to the study. With the presented methodology, the chapter sets the stage for Chapter 4, which will provide results from the empirical analysis, simulation experiments, and field validation, Chapter 5, which will present the system architecture, and Chapter 6, which will discuss the implications of these findings.

CHAPTER 4: EMPIRICAL RESULTS AND VALIDATION

4.1 Introduction

The last chapter described the methodology approach employed in this study, which aims to develop a framework (GUARDIAN framework) that addresses the irreducible uncertainty by transforming it into automatic graduated protection responses to vulnerabilities of the system. The methodology combined field validation with historical empirical analysis and high-fidelity simulation to validate uncertainty-parameterized adaptive isolation by analyzing CVEs. The first stage was the empirical analysis of 279,056 unique CVEs (January 2002-December 2024) and established the technical feasibility of the framework. The findings of these methodological practices are introduced in this chapter, differentiating between the results proven empirically and those that need additional operational verification.

This study has a number of sections to provide the research findings. First, it discusses the basic information constraints and model performance of the six machine learning models including individual model performance, temporal consistency, as well as vulnerability category performance. Second, it discusses the findings of using ensemble disagreement as the measure of uncertainty, such as the JSD distribution, the quality of calibration, and the relationship between disagreement and accuracy. It also discusses the association of uncertainty with parameters, as well as validation of temporal self-correction mechanisms. Furthermore, the chapter presents findings on the graduated protection effectiveness, processing performance validation, and the field validation study with the healthcare critical infrastructure practitioners.

4.2 Fundamental Information Constraints and Model Performance

This research holds that uncertainty is inevitable. To confirm this assertion, the study needed to demonstrate that traditional predictive machine learning models cannot be perfectly accurate because of the constraints in the available data. This was conducted in the first and second steps of the empirical analysis stage where six machine learning models were trained on 160,543 CVEs and tested on 68,711 CVEs. The six models (Random Forest, Support Vector Machine, Logistic Regression, Neural Network, LightGBM and XGBoost) have radically different mathematical approaches and assumptions, and this section analyzes the performance of each of the models. The findings show that the performance bottlenecks cut across all models despite their architectural variations and therefore the ceiling is due to data limitations and not to a specific algorithmic implementation.

4.2.1 Individual Model Performance Analysis

The six models have shown great variation in performance although they have been trained based on the same CVEs showing major evidence of the inherent information constraints that they may have. One of the main results was that regardless of the mathematical methods used, all the models reached moderate accuracy levels using the data, which did not include the organizational context.

Random Forest. The Random Forest model that used bootstrap sampling of 100 decision trees had a balanced accuracy of 58.3% with precision of 61.2% and Matthews Correlation Coefficient (MCC) of 0.448. This model was able to model non-linear interactions between the features unlike other models like the SVM and logistic regression. This led to improvement over some of the other models by capturing these conditional relationships. This was achieved through keeping overfitting to a minimum and training time was only 5 minutes, which is quite

impressive compared to some linear models even though the complexity in this case is higher. It was a reasonable compromise between the needed speed and model accuracy.

Support Vector Machine. The SVM with RBF kernels obtained a balanced accuracy of 47.5%, precision of 51.2% and an MCC of 0.309. This was not a high-accuracy finding, given that this is one of the most established models in the literature, which is usually characterized as a model that can classify high-dimensional text, and can achieve linear separation, through higher-dimensional projections of its features. The model was less accurate in this case than other models in the test, suggesting that the model was unable to distinguish the semantic relationship between the terms of severity and the CVE terms used which are highly non-linear and intertwined in this feature space. There is a possibility that higher-dimensional projection would not have been able to successfully separate using text only. Moreover, it should be noted that this model takes an average of 3.2 hours to be trained, hence it is computationally expensive, especially given the low accuracy.

Logistic Regression. The third model, a linear model, logistic regression, which is applied with elastic net regularization, had a balanced accuracy of 54.2%, a precision of 57.8%, and an MCC of 0.391. It was a notable result, mainly because logistic regression is a relatively simpler linear classifier, as opposed to a model like the SVM which is complex and kernel-based, which indicates that the SVM model was unable to capture important patterns in the training data. Along with this observation was the fact that such a model took only 2.5 minutes of training time to converge, establishing a performance floor and indicating that the complexity of a model is not necessarily the factor that defines how useful a model can be in cases where the input data is sparse as in this case.

Neural Network. The fourth model was the neural network which applied the multilayer perceptron, which has been used in supervised learning to project data to a desired result due to its capacity to minimize errors through backpropagation, with a balanced accuracy of 62.7%, precision of 65.4% and MCC of 0.507. This is a significant finding because it shows that the extent of a problem encompassing vulnerabilities is not necessarily open-ended as there are sometimes hidden patterns that can be extracted in the description of a software defect in only two sentences. As per the results, the model indeed saturated the available signal but reliability at the level of human performance was not reached, which also suggests that the required signal is unlikely to be found in the text under analysis. The duration of training needed by this model was 25 minutes.

LightGBM. The best individual model performance resulted from LightGBM with the highest overall performance on balanced accuracy of 69.3% and precision of 71.8% and MCC of 0.594. This efficiency is attributable to the histogram-based leaf-wise tree growth strategy, a departure from traditional level-wise growth that optimizes the split with the largest information gain (Ke et al., 2017). This allowed the model to determine optimal decision boundaries using only 5 minutes of training time. Also worthy of mention is that on combining the six models, the accuracy was at 70.8%, which suggests that all the models were making correlated errors.

XGBoost. Lastly, the XGBoost model, which utilized gradient boosting with sequential error correction (Chen & Guestrin, 2016), achieved a balanced accuracy of 67.1%, precision of 69.4% and MCC of 0.565. This performance was explained by the fact that every new tree was focused on the misclassified samples of previous iterations, and thus, the model would extract greater detail out of the data compared to other models, such as the random forest. It was

established that this model provided a relevant trade-off between the requirements related to accuracy and efficiency because the training required was 9.5 minutes.

Upon conducting the statistical significance testing, it was further apparent that the performance differences between these models exceed the chance variations. The Friedman test, which is used to compare the performance of classifiers across multiple related samples (Demšar, 2006; Hollander et al., 2013), found that $\chi^2(5) = 12{,}847.3$ with $p < 0.001$. However, the difference between XGBoost and LightGBM, the two high-performing models, was insignificant (Cliff's $\delta = 0.048$). This highlights that the 70% accuracy (on average) might be the ceiling of the descriptive power of the current dataset as presented and used, determined by the nature of the available data rather than the models used.

Table 1. Summary of individual model performance analysis.

Individual Model	Balanced Accuracy	Precision	Recall	F1-Score	MCC	Training Duration
Random Forest	58.3%	61.2%	58.3%	0.587	0.448	5 minutes
SVM	47.5%	51.2%	47.5%	0.468	0.309	3.2 hours
Logistic Regression	54.2%	57.8%	54.2%	0.538	0.391	2.5 minutes
Neural Network	62.7%	65.4%	62.7%	0.624	0.507	25 minutes
LightGBM	69.3%	71.8%	69.3%	0.690	0.594	5 minutes
XGBoost	67.1%	69.4%	67.1%	0.668	0.565	9.5 minutes

Methodological note: Recall values equal balanced accuracy across all models because balanced accuracy is defined as the unweighted mean of per-class recall rates (Brodersen et al., 2010), making the two metrics identical by definition in multi-class classification.

4.2.2 Temporal Stability Analysis

A key question in this research, before presenting the GUARDIAN framework, is determining whether the observed information constraints as seen above are static or evolve over time. This is important since if in fact they are evolving, then this also means that the ceiling of

their accuracy will be improved with time and as a result, perhaps as the description of the vulnerabilities becomes more accurate, the situation will ultimately be resolved through information naturally evolving. Nevertheless, the current information has been accumulated over 23 years meaning that the constraints are inherent to the information and therefore the nature of vulnerabilities does not define the accuracy ceiling. The analysis of temporal stability of the data in the current study was performed using the annual cohorts of the data in the last 23 years (January 2002 to December 2024) and the accuracy of the models was evaluated to establish whether this has changed over the years.

The random forest model had a coefficient of variation of 1.3%, XGBoost had 0.9%, and neural networks had 1.1%. Temporal consistency also underscores that there will always be information constraints irrespective of the change in the disclosure processes, the transforming software systems and organizational practices within the systems. It indicates that even though the nature of attacks has changed as well as the process of disclosure has changed, the gaps in the CVE description have not disappeared and this continues to affect the severity descriptions. It shows also that organizations can be confident about applying the uncertainty-measuring mechanisms because there is no likelihood of the signal-to-noise ratio changing over decades.

4.2.3 Vulnerability Category Analysis

The Common Weakness Enumeration (CWE) categories were also investigated following the analysis of the temporal stability to guarantee that the outcomes were not biased by the specific type of vulnerability. It was done following the current division into the major CWE categories, which is maintained by MITRE Corporation and included in the National Vulnerability Database to offer a common vocabulary to describe the root causes of security vulnerabilities (MITRE, 2024; NIST, 2025).

This was then followed by a one-way ANOVA to compare the accuracies of the categories and it provided $F(5, 30) = 2.47$ and $p = 0.055$, which was not significant at $\alpha = 0.05$, with a partial $\eta^2 = 0.0002$.

The fact that the marginal accuracy of SQL injections (71.4%) is the highest among the other categories is also interesting and is likely explained by the fact that the number of specific keywords, such as SQL, query and database, which can be strongly correlated with the severity of the degree of CVE, is the largest and cannot be generalized in terms of functionality. In addition, the information exposure was the least accurate, which can also be explained by the fact that it is not always clear how detrimental a leak may be due to the nature of an organization, and not knowing whether there are such cases. Overall, no considerable differences were observed, which also contributes to the discussion of the fact that not very specific but all types of vulnerabilities are affected by information constraints.

4.3 Ensemble Disagreement as Uncertainty Quantification

After it was established that the individual models are at an average level of accuracy that cannot be raised due to the current restrictions that are placed on the data, the study went on to use the disagreement between models to measure the uncertainty. This part supports and justifies the use of JSD as a valid proxy of uncertainty.

4.3.1 Jensen-Shannon Divergence Distribution

The JSD is used to measure the similarity of probability distributions as per the methodology (Lin, 1991; Nielsen, 2019), and in this case, it was used to determine the extent of disagreement between the models on the severity of a vulnerability. This was necessary because through the analysis of the ensemble disagreement pattern across the 68,711 test CVEs, the

analysis identified the distribution features needed for effective thresholding. The mean JSD was 0.400 with standard deviation of 0.047 and coefficient of variation 11.75%. Also, the median JSD was 0.398 indicating that there was no extreme skew and the distribution was symmetric. This meant that, in cases involving a large number of CVEs, the models coincided, resulting in small JSD values. Nonetheless, the 95th percentile revealed as well that there was a range of vulnerabilities that were so vaguely described that the models came to conflicting conclusions on their meaning (0.487 at the 95th percentile, with a maximum of 0.894).

Since the pattern of disagreement was already indicated, but the temporal analysis had demonstrated important stability, this implied that threshold calibration could be performed without any fear of temporal drift taking place. This was observed in the data, which indicated that the mean JSD was less than 1% different between the beginning and the end of the test period (24 months between January 2023 and December 2024), and the mean JSD per month was 0.394-0.406. Organizations can set thresholds based on this data, such that the response will remain stable, because the disagreement patterns are consistent and have very low levels of variation.

4.3.2 Calibration Quality Assessment

In any study, for risk to inform decision making in an automated system, the probability of error that is identified must be equal to the probability of error that occurs in reality. This is referred to as the calibration process, and is frequently applied particularly in machine learning, in which models can be overconfident, resulting in high-assigned probability scores that may not be accurate (Guo et al., 2017). This was addressed by the temperature scaling method applied in this study and the parameter T = 1.47 was identified on the validation data, which significantly

improves the quality of calibration. This caused the ECE to drop by a wide margin from the originally anticipated 0.147 (where the models were off by 14.7%) to 0.023.

Additionally, reliability analysis was conducted to test the relationship between predictions and the observed accuracy, in ten confidence bins, and it was found that predictions and observed accuracy were closely aligned. This was observed in that when the calibrated ensemble would predict an 85% accuracy, the actual accuracy was found to be 83.7%, and when the predicted confidence was 65%, the actual accuracy was found to be 66.2%, and at 45% predicted accuracy, the actual accuracy was found to be 46.8%. The maximum level of deviation was always less than 3 percentage points.

This is a key result in this study because it helps make the framework's case that when the system claims to be uncertain, that claim should be believed because it is mathematically reliable. This also justifies the conservative protective actions presented by the framework as a response to address those scenarios. The use of JSD for uncertainty quantification receives additional validation from comparative analysis of divergence metrics. Murphy et al. (2025) found that JSD-regularized ensembles produced predictions that were both more accurate and better calibrated than ensembles using alternative metrics, with ECE improvements of 45% over baseline approaches.

4.3.3 Disagreement-Accuracy Correlation

The final step in this process of validating the uncertainty signal was to prove that there is a correlation between disagreement and inaccuracy. This was integral because, if models disagreed but were still correct, then JSD would not be a useful metric in risk determination. On analyzing the relationship between disagreement and accuracy, the results of Spearman's rank correlation between JSD and accuracy were $\rho = -0.92$ ($p < 0.001$). Further, upon partitioning of

the vulnerabilities into JSD bands as previously described, there was a significant monotonic reduction in accuracy, showing the signal was discriminative as the disagreement increased (Lakshminarayanan et al., 2017). The distributions across the bands are as shown below:

Band 1 (JSD < 0.35): 81.4% accuracy.

Band 2 (JSD 0.35–0.42): 76.2% accuracy.

Band 3 (JSD 0.42–0.51): 68.7% accuracy.

Band 4 (JSD 0.51–0.60): 55.3% accuracy.

Band 5 (JSD ≥ 0.60): 38.7% accuracy.

From Band 1 to Band 5, the variation in accuracy was 42.7%, which indicates that JSD is an extremely discriminative signal. Chi-square testing shows that the accuracy variation does have significant difference among the bands ($\chi^2(4) = 18{,}947.2$, $p < 0.001$) with a large effect (Cramer's V = 0.525).

This is a significant finding, as it establishes a basis on which the automated response can be implemented within the framework on an empirical basis, because the vulnerabilities with high JSD scores cannot be regarded as reliable, and traditional risk scoring should be bypassed in favor of calibrated isolation. Figure 14 illustrates this monotonic relationship, showing accuracy declining from 81.4% in the lowest JSD quintile to 38.7% in the highest, a 42.7 percentage drop that validates JSD as a discriminative uncertainty signal.

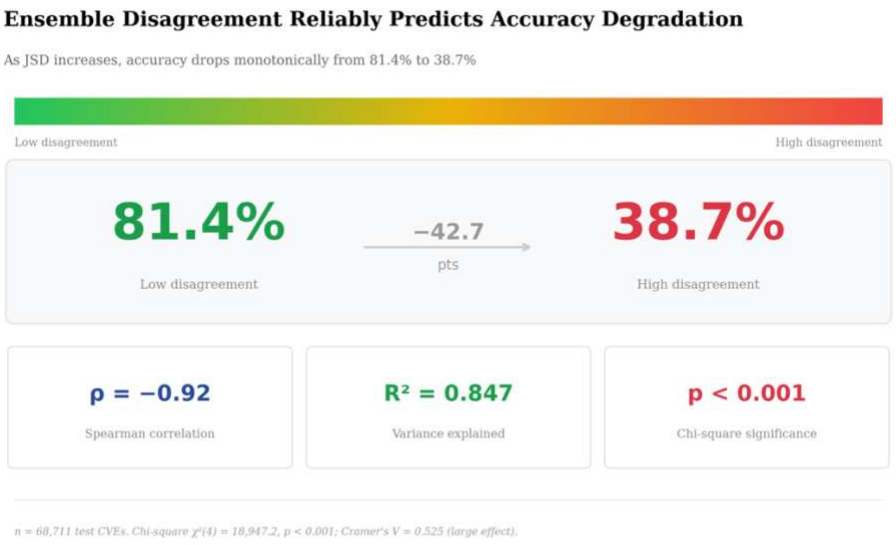

Figure 14. Ensemble Disagreement Reliably Predicts Accuracy Degradation

This correlation instantiates the emergence phenomenon Simske (2013) identifies in meta-algorithmic systems employing Output Probabilities Matrices: combined classifier outputs can achieve correct results even when individual classifiers fail, but conversely, divergent outputs signal reduced reliability across all generators. The GUARDIAN framework exploits this complementary insight: high JSD indicates conditions where uncertainty-parameterized protection substitutes for potentially unreliable assessments.

4.4 Uncertainty-Parameter Relationships

The above findings have verified the existence of the uncertainty signal, and thus, the study sought to determine the relationship between the uncertainty signal and the risks in the real world. This played a key role in the framework as it justifies the rationale of the use of the parameter mapping functions that will subsequently result in the translation of the JSD scores into protection levels and timeframes in the system.

4.4.1 Exploitation Rate Analysis

There was a need to partition and prioritize vulnerabilities into uncertainty percentiles to establish whether the systematic exploitation rate progression supports the use of graduated protection design to ensure that the system is calibrated based on the degree of uncertainty it is addressing. In order to verify the correlation between degrees of uncertainty and real-world threats to the system, it was necessary to map the levels of uncertainty to real-world exploitations by using real and authoritative sources. To accomplish this, several leading sources were used including the VulnCheck KEV, the list of known exploited vulnerabilities, the CISA Known Exploited Vulnerability Catalog, and the Exploit Prediction Scoring System (EPSS) developed by the Forum of Incident Response and Security Teams (CISA, n.d.; Jacobs et al., 2021; VulnCheck, 2024). The bands represent operationally-defined thresholds aligned to isolation level transitions rather than statistical quartiles.

The CVEs were grouped into four uncertainty bands, which include the low uncertainty band, the moderate-low uncertainty band, the moderate-high uncertainty band, and the high uncertainty band. The low uncertainty band, which was partitioned to include vulnerabilities whose JSD was between 0.02 and 0.35, exhibited an 8.2% exploitation rate and included cases that often show clear severity indicators, have standard attack patterns, and are clearly described to maintain consistency in model assessment. The minimal exploitation rate suggests that these vulnerabilities probably receive adequate defensive attention.

The second uncertainty band (moderate-low uncertainty band) contained vulnerabilities whose JSD was 0.35 to 0.42 and recorded an exploitation rate of 15.7% which is significant because this is double the rate of the previous band. This must have been due to a compounding factor of mixed severity indicators in place and the diverse vectors that induced an attack leading

to a moderate level of agreement and substantial increase in risk of exploitation. The third uncertainty band was the moderate-high, and the vulnerabilities in the moderate-high uncertainty band had JSD values of 0.43-0.51 which revealed the exploitation rate of 23.4%.

This high rate of exploitation is most likely explained by the advanced exploitation chains and ambiguous approaches to quantifying the effects that result in triggering model conflict and that nearly a quarter of the vulnerabilities can be exploited. The high uncertainty band consisted of those with a JSD of more than 0.51 and had a 31.2% exploitation rate that is nearly four times as much as the low-uncertainty band. This is likely due to conflicting information in the vulnerability descriptions, resulting in maximum model disagreement and maximum exploitation risk. This is significant because it demonstrates that ambiguity that is traditionally represented as complexity or novelty is a weakness in itself and the more ambiguity exists in the data, the more hazardous it will be and the more it will be vulnerable to curious and intelligent attackers.

Spearman's rank correlation between the JSD and the probability of exploitation was $\rho = 0.89$ ($p < 0.001$) indicating a strong positive correlation between the two variables. These findings have some likely implications; however, they strongly indicate that the common belief that unknown is often low risk is incorrect, and such ambiguous vulnerabilities could be the most dangerous to a system. This progression is visualized in Figure 15 across uncertainty bands, and the exploitation rates are indicated to increase four times between low-uncertainty vulnerabilities (8.2%) and high-uncertainty cases (31.2%), which is contrary to the expectation that ambiguous vulnerabilities are low-risk.

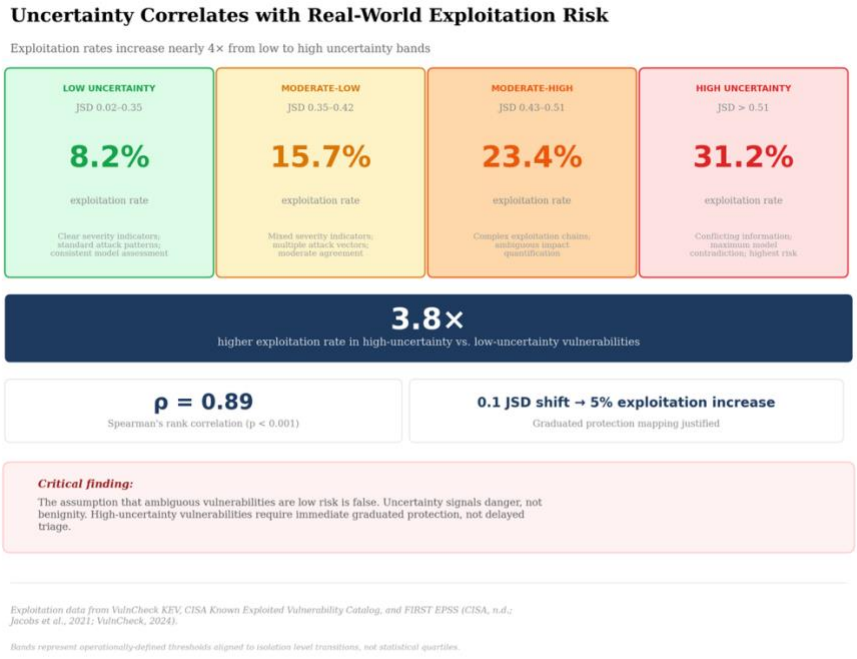

Figure 15. Exploitation Rate by Uncertainty Band

4.4.2 Monotonic Relationship Validation

The GUARDIAN framework is based on mathematical functions that have the capacity to map uncertainty to the stringency of protection (resulting in levels of isolation) and protection duration (the time during which isolation is carried out). As the statistical analysis in this section demonstrates, the variables of duration and isolation levels are monotonically dependent, which justifies using continuous and automated responses on a sliding scale as opposed to binary on-off switches. Statistically, uncertainty and isolation (protection stringency) and uncertainty and time (protection duration) are positively and negatively correlated as discussed below.

Spearman's rank correlation between JSD and isolation levels recommended on the observed exploitation rates showed that $\rho = 0.89$ with $p < 0.001$, which is a strong monotonic relationship that confirms that as the uncertainty levels increase, the level of exploitation risk

also increases, and thus the need for graduated protection that similarly escalates as the levels rise. A 0.1 shift in JSD causes an equivalent 5% rise in the possibility of exploitation, and so mapped-out protection levels are required in response.

Similarly, when the two parameters are correlated with the duration of protection as perceived by industry as adequate, JSD has negative Spearman's rank correlation of $\rho = -0.87$ with $p < 0.001$, which implies a negative but significant relationship between the two parameters. This corresponds to the findings of different reports by Qualys and VulnCheck according to which high-uncertainty vulnerabilities have a direct relationship with immediate exploitation patterns, which require shorter initial timelines to isolate them as well as continuous evaluations (Qualys, 2024a; Garrity, 2025). The low uncertainty cases are at the other extreme and can be mostly matched with longer exploitation patterns and this demonstrates the need to have longer protection windows and space to achieve stability. Finally, the correlations between these parameters and uncertainty offer organizations the opportunity to seek progressively calibrated responses, according to the range of uncertainty levels they are addressing in the vulnerabilities, the duration of response, and the extent of isolation required to address such issues.

4.5 Temporal Self-Correction Validation

Having validated the need for graduated responses to address the uncertainty levels in the system, the next step in validating the need for the framework is validating the mechanisms that allow the framework to implement temporal self-correction practices through automated systems without needing continuous human monitoring. This is important because it distinguishes design rationale, which is obtained from industry research, from technical feasibility, which is derived from the simulation, enabling demonstration of framework effectiveness under controlled conditions.

4.5.1 Rollback Timing Design Rationale

One of the major tasks in creating this framework was to establish the rollback timing thresholds which would be used as design specifications in the simulation and this involved the need to seek out authoritative industrial research that has recorded the timing patterns of exploitation for established and known vulnerabilities. In this part of the study, rather than using original empirical analysis that would be constrained in the quantity of the participants taken into consideration, the researcher used independent research reports of leading industrial players that encompass the temporal durations of the confirmed events of exploitation. As stated in the methodology, four reports were found and these were used in this part of research.

The first report that was identified to inform the threshold design was published in 2024 by Mandiant, analyzing 138 vulnerabilities disclosed in 2023 (Charrier & Weiner, 2024). The study found that 97 of the 138 vulnerabilities were exploited as zero-day vulnerabilities before there were any patches made available, while the remaining 41 vulnerabilities were exploited several days after the patches were made available (Charrier & Weiner, 2024). This was a significant change on the attacker's end, from the average 5 days to exploitation previously reported by other studies after a vulnerability has been disclosed for the majority of them, and a substantial acceleration from the 32-day period identified as the median time for exploitation after a vulnerability has been disclosed in previous studies (Charrier & Weiner, 2024). This report illustrates the evolving nature of adversaries and the need for aggressive defense timelines.

The second report used for this rollback time determination was the VulnCheck 2024 study on trends in vulnerability exploitation which reported that of the 768 vulnerabilities reported, there was a noticeable bimodal distribution in how the exploitations were timed evident in the zero-day exploits, and those that had a long trail of time (192 days) before an attack was

launched (Garrity, 2025). Those targeted within 24 hours were reported to reflect mass targeting, while those that were exploited later on were characterized by more sophisticated targeted campaigns aimed at compromising high-value systems (Garrity, 2025). The bimodal patterns inform the selected graduated thresholds approach.

The Qualys 2023 report is also an important report that discusses the exploitation of confirmed vulnerabilities and studies the 206 high-severity vulnerabilities (Qualys, 2023, 2024a). According to the report, 25 percent of such vulnerabilities were exploited within 24 hours of being published and then possessed a median time-to-exploit of 44 days (Qualys, 2023, 2024a). The exploitation rate on the day of publication (within 24 hours) is in line with results of the other report by VulnCheck on immediate opportunistic targeting whereas the median time is also the result of sophisticated campaigns requiring more preparation time (Qualys, 2023, 2024a; Garrity, 2025).

Likewise, the report by Rapid7, on 2022 vulnerability intelligence also supported those findings, claiming that swift exploitation, which is typical within 48 hours of an exploitation being reported, is usually opportunistic in nature (Condon et al., 2023). The report also mentioned that some are conducted following this initial timeline, which is represented by patient directed campaigns which are aimed at inflicting as much damage as possible on high-value infrastructure, the latter often extending by months in terms of timeline (Condon et al., 2023). This distinction between commodity-based and targeted exploitation supports the requirement of a tiered approach in establishing duration of system protection.

Based on these four reports, it was clear that when various datasets were reviewed, timelines of attacks occurred which required a sound threshold design which is tiered and employs graduated thresholds depending on required level of protection as per the exploitation

times and patterns. Two major findings were made across the four reports: first, immediate exploitation of the vulnerability after disclosure happened within the first 24 to 72 hours and that was manifested by mass opportunistic attacking primarily via the use of automated scanning and exploitation systems, and secondly, long term exploitation often occurred weeks and even months after the initial vulnerability was disclosed and this consisted of targeted and sophisticated campaigns to fracture the security of high-value infrastructure and organizations significantly. Based on these findings the study established four graduated thresholds to address the modes of the exploitation that have been reported and at the same time accommodate the safety of the systems by taking safe margins that could give room to evaluate the attack and come up with solutions: 24 hours, 48 hours, 7 days, and 30 days.

The four graduated thresholds were allocated to the uncertainty bands that were identified in the section above due to elevated scores of JSD. The high-uncertainty vulnerabilities were assigned 24-48 hours of protection since such are most likely to be exploited quickly and the ambiguous information requires rapid reassessment, making it easier to perform system adjustments as more information is learned. The low-uncertainty vulnerabilities received the longer thresholds of 7 to 30 days since the indicators of their severity are more transparent and this will enable confidence in the continued monitoring.

4.5.2 Simulation-Based Feasibility Testing

As previously described, the simulation used to validate the technical feasibility of the logic presented by this framework was a water treatment testbed based on water utilities. The simulation environment was modeled to match a realistic municipal water utility infrastructure. To ensure fidelity, the testbed architecture incorporated the control system design principles described by Morris et al. (2011), simulating a multi-stage treatment process controlled by

programmable logic controllers (PLCs) with a four-tier architecture. This validated ICS security testbed is capable of generating high-fidelity physical process data under cyberattack conditions. The simulation was conducted across 1000 vulnerability scenarios and tested under the signal parameters as defined in the methodology to determine the performance ceiling rather than making a normative choice with regard to normal operational expectations. These were various signal quality assumptions that were configured in the simulation, which were that the IDS system was configured to perform at 95% true positive to guarantee high-quality intrusion detection without false-positives, and that the threat intelligence was configured to 90% as a guarantee to have comprehensive commercial and open-source feeds (Sommer & Paxson, 2010). The patch tracking was also configured as perfect (100% visibility) to have the organization accurately aware of the remediation status (Ginter, 2019; Stouffer et al., 2023). The results showed that the temporal self-correction mechanisms under controlled conditions are technically feasible.

Additionally, there were other multi-signal conditions that needed to be set for the simulation environment before seeking rollback on the protection. These included a necessary calibration of time thresholds to uncertainty levels that had to elapse before there was any exploitation evidence, that there be no exploitation alerts detected in the IDS systems (or others such as the Security Information and Event Management (SIEM) monitoring system), and that the threat intelligence had to indicate there were no active targeting efforts going on. Moreover, patches were to either be deployed or the uncertainty levels had to be reduced after the ensemble reassessment. From these conditions, the multi-signal temporal self-correction mechanism demonstrated a rollback execution rate of 82 to 94% across different uncertainty levels, which was considered a successful execution. The difference (82%-94%) reflects the variation in

appropriate protection strategies, with high-uncertainty vulnerabilities requiring extended monitoring periods and low-uncertainty vulnerabilities eligible for earlier rollback.

The system also demonstrated that it could adapt when it encountered friction by automatically extending protection in 6.4% of cases that were experiencing operational problems requiring further monitoring. On additional examination of the outcomes, it was found that these extensions were triggered when at least one of the rollback conditions had not been satisfied at the time the threshold expired, such as patches not yet being deployed, exploitation alerts still being detected in the IDS or SIEM monitoring systems, or the JSD scores remaining elevated following ensemble reassessment.

These results illustrate the limitations of simulated environments, while also illustrating that the self-correction logic is not a blind timer but a feedback loop that adapts to the system requirements in the real world in order to prioritize system safety. The comparative effectiveness of different rollback strategies is examined in the following section.

4.5.3 Multi-Signal Integration Effectiveness

Having performed the multi-signal integration, the study aimed to measure the value of that practice and identify its effectiveness compared with the baseline values in 500 simulated cases. Three rollback strategies were employed to activate the controlled comparison and isolation of the integration value to carry out the comparison testing. The initial rollback strategy was based on time-only basis, while the rollback protection was put in place after a period of 48 hours irrespective of the state of the system. Although this was a straightforward method, it produced 81.6% correct rollbacks, 18.4% premature rollbacks, which created exploitation holes. This led to situations where the system was not protected even though the attack was live or

looming and hence an unacceptable security risk which is particularly unacceptable in a critical infrastructure environment.

The time plus exploitation detection baseline was the next rollback strategy that introduced the 48-hour rollback as a condition on the basis of another exploitation being detected with the help of IDS alerts or SIEM monitoring systems. This was a more effective rollback strategy than that of the first strategy, with 89.2% proper rollbacks, 7.6% untimely rollbacks and 3.2% protection extensions. Although this rollback strategy was more effective, it still could not be applied to such critical facilities as nuclear plants, water treatment facilities, etc. where the failure of a single facility will have disastrous consequences.

The third rollback policy adopted was a multi-signal integration policy, which adopted graduated thresholds that were adjusted on the uncertainty of the system, and only succeeded as long as all the safety requirements were satisfied, such as the exploitation detection, threat intelligence assessment, and patch availability verification. The appropriate rollback rate of this strategy was 92% with 1.6% premature rollbacks and 6.4% protection extensions, which is the most appropriate for critical infrastructure. Table 2 below indicates the overview of the three strategies.

Table 2. Comparison testing across the three rollback strategies.

Rollback Strategy	Approach	Appropriate Rollback	Premature Rollbacks	Protection Extensions
Time-only baseline	Fixed 48-hour rollback	81.6%	18.4%	—
Time + exploitation detection	48-hour if no alerts	89.2%	7.6%	3.2%
Multi-signal integration	Graduated + all conditions	92.0%	1.6%	6.4%

The study further conducted statistical validation to confirm the magnitude of improvement across the three strategies. Using the McNemar test, which is commonly applied in

checking if there is any significant change in the level of difference existing between two related groups (Sundjaja et al., 2023), to compare the multi-signal to time-only strategy found that $\chi^2(1) = 76.3$ with $p < 0.001$. In this discussion, it may also be observed that the rollback rate of 18.4% of time-only strategy can be equated to the 1.6% rollback rate of multi-signal integration strategy that should show a significant improvement in quantity of protection provided to the critical infrastructure. Other than these safety advantages of the graduated temporal threshold, the research found that availability also increased by 9.1% as opposed to the fixed-threshold methods with the extent increasing from 82.1% to 91.2%. The latter was attributable to the ability of the graduated logic to allow the release of the low-uncertainty vulnerabilities prior to the fixed 48-hour window, which was also discovered during the baseline, and as the higher-risk vulnerabilities are isolated to further extend their durations of containment. These results also highlight the fact that intelligent differentiation is the way through which a better tradeoff between protection and availability is reached than when it is done in homogenous ways.

4.5.4 Validation Boundaries and Operational Requirements

While the simulation has proved the above, there remain aspects that require operational validation. This section highlights the validation boundaries as mentioned above and distinguishes them from the operational requirements. The simulation has determined that when the inputs are correct, then the decision tree will result in an optimal safe solution, with the multi-signal integration and all the conditions being conserved in the system. This can be observed in how the time-only baseline to multi-signal integration strategy improved by 91% in the rate of premature rollback. To achieve this, however, there are key system requirements and assumptions the simulation holds, such as high signal quality of 95% IDS accuracy, 90% threat intelligence, and perfect patch availability and matching, which are largely optimistic parameters

because production environments often have a lower signal quality due to higher noise and configuration drifts.

Recognizing such potential shortcomings, while also being unable to conduct empirical research that could determine the perfect assumptions to conduct a study such as this, the research instead sought design specifications that are informed by industry research based on how other known and published exploitations have performed and the assessment conducted on them. The rollback timing thresholds are informed by those specifications, and the success rates represent the theoretical ceiling informed by those specifications. This means that when an organization intends to use this framework, it has to be calibrated on their own parameters which they have received through their real-life capabilities in monitoring and practices. Simulation has demonstrated that the mechanism is successful, and to realize this solution, organizations have to perform operational pilots to adjust the sensitivity of the framework to their own sector and companies.

4.6 Graduated Protection Effectiveness

Upon demonstrating the effectiveness of multi-signal integration to reduce premature rollbacks and offer a greater degree of protection to vulnerabilities at different levels, the next step involved simulation-based attack testing to confirm the graduated level of protection as designed. This section covers the results of the 6000 attack trials that were used to conduct the attack simulation, using 12 attack variants based on the MITRE ATT&CK for ICS taxonomy as described by Alexander et al. (2020). The effectiveness of the four graduated protection levels is reported below, excluding the baseline condition (Level 0-1, JSD < 0.35) where no active isolation was applied.

The first level, the enhanced monitoring level, was designed for low uncertainty, low severity cases and included increased logging and enhancement of IDS sensitivity to alerts, without blocking the traffic. This level attained a risk reduction of 56% and maintained 99.4% availability.

The second level, the restricted communication level, applied application layer allowlisting to restrict attacks to the system without losing vital system functions. This level recorded 74% reduction in risk and maintained 94.9% availability.

The third level, the zero trust micro-segmentation level, imposed strict mutual TLS and peer authentication to prevent attacks including credential-based attacks. This level was able to reduce the risk by 88%, maintaining 89.5% availability. The impact of this protection level on operations increased at this tier because the protection approach can end up blocking legitimate but unusual workflows in the process of neutralizing other attacks that cannot bypass mTLS certificate requirements.

The fourth level, the complete quarantine level, focused on fully isolating the asset under threat in cases of high-uncertainty and high-severity threats, and only allowed emergency safety signals such as the physical safety interlocks. This tier registered a 94% decrease in risks, which stopped all the network-based attacks; however, the outcome was also significant operational constraints and limitations, resulting in 72.8% availability.

To establish the monotonic relationships between protection and availability, the study performed statistical validation using one-way ANOVA, which yielded $F(4,495) = 53.74$ with $p < 0.001$ and $\eta^2 = 0.303$, indicating a large effect size and that the protection levels provided a systematic monotonic tradeoff, allowing organizations to adjust the protection levels based on the JSD uncertainty score while being aware of the possible consequences in terms of risk

reduction and operational limitations. Figure 16 summarizes these four protection levels, showing the tunable tradeoff between risk reduction (56-94%) and system availability (72.8-99.4%) that enables organizations to calibrate responses proportionally to uncertainty.

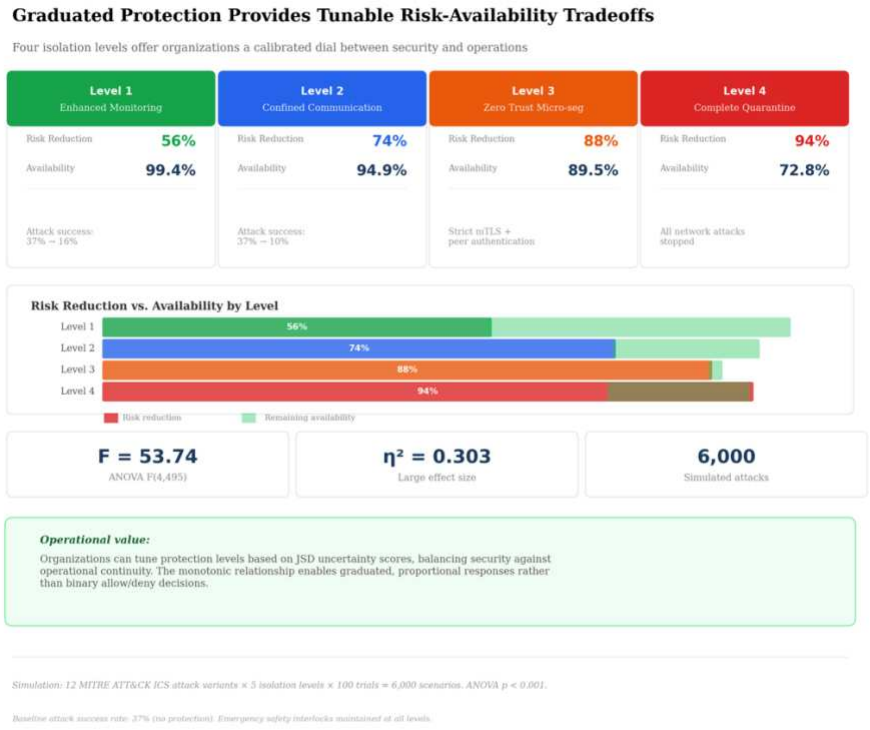

Figure 16. Graduated Protection Provides Tunable Risk-Availability Tradeoffs

4.7 Processing Performance Validation

In order to further validate the ability of this framework to perform during a crisis, the study calculated the processing time that would be needed based on optimized laboratory conditions. This was an important step in the research because a protection framework is only as effective as its ability to respond and provide automated defense faster than the attacker can access a system.

Upon validating the crisis-speed of the framework under laboratory conditions, there were several key times noted. First, the end-to-end processing time for a single vulnerability was determined to be approximately 720 milliseconds, of which model inference consumed the majority of the processing time. Second, scalability testing evaluated how the system would perform under load. Results showed that at a rate of 100 CVEs at a time, 8.3 seconds were required to process the CVEs and, at burst capacity testing, 500 CVEs could be processed in 5 minutes.

These figures are significant for operational benchmarking because with current systems, it takes organizations in the critical infrastructure sectors 45 to 75 minutes to address a single vulnerability and these figures do not include other variables such as board review, organizational coordination and vulnerability review. When full decision coordination is included, this process can extend to 12 to 36 hours. The GUARDIAN framework, as shown above, however, takes only 720 milliseconds to generate a protection decision, representing a 99.9% reduction from the timelines achieved manually. This substantially improves the security posture of critical infrastructure organizations by enabling response speeds that can counter adversaries who exploit vulnerabilities within hours of disclosure as reported by Mandiant and VulnCheck (Charrier & Weiner, 2024; Garrity, 2025).

4.8 Field Validation Study: Healthcare Critical Infrastructure

A key part of this study was validating that, beyond the framework technically working, it can also be used operationally by practitioners in the field. The previous sections have proven that it can, in fact, technically work, and this section addresses the last mile of the research by conducting a field study to assess practitioner acceptance in terms of operational capability.

4.8.1 Introduction

As was already emphasized, the regulation of critical infrastructure sectors determines the extent of the research that may be conducted to shape the methodology of the research aimed at such a setting. Notable examples previously provided were the HIPAA and FDA medical device regulations that define what can be done in terms of research in healthcare facilities, and the 10 CFR 73.54 regulation by the NRC on research in nuclear facilities (Edemekong et al., 2018; FDA, 2025; NRC, 2021). It is important to remember that these are not research limitations but rather necessary safety requirements to protect the public, as these are critical infrastructure systems. To navigate this challenge in this study, however, the complementary approach previously described was applied, where simulation was used to test the technical capabilities of the automated systems, while this section tests the decision-support mechanisms in terms of their application and use. To conduct this research, the study sought a regional healthcare system that was willing to participate in the research in an advisory capacity regarding the deployment, where the framework generated isolation recommendations but did not execute them in any capacity. Through this participation, the study was able to answer three questions the simulation was unable to: whether practitioners found the JSD uncertainty scores interpretable for decision-making purposes, whether the graduated protection levels aligned with the practitioners' risk perception, and whether they would trust this system enough to allow automation at some point.

4.8.2 Methodology

Sampling Strategy. The researcher contacted professionals in the healthcare field with experience in IT/OT who worked at the regional healthcare facility, which included acute care facilities, outpatient clinics and diagnostic imaging centers. The objective was to acquire a total

of fifteen healthcare professionals who would consent to participate in the research and in this regard, the researcher employed purposive sampling, whereby they contacted people who worked at the facility to obtain the desired total. Purposive sampling is a non-probability-based sampling strategy, and the researcher attempts to identify respondents who have specific characteristics (especially those who are well familiar with a specific area) (Ahmad & Wilkins, 2025). The researcher, in this instance, was interested in individuals with experience in the field of IT/OT operations in the regional healthcare facility; the most effective strategy was to actively find them on the basis of their profiles and information provided by the IT Director. Additionally, the researcher was keen on a sample size greater than 12 participants because research has continuously shown that when conducting qualitative research, which is then applied to thematic analysis for data analysis, the most suitable participant size to ensure data saturation is reached is 12 to 20 participants (Ahmad & Wilkins, 2025).

The participants that were obtained were from four departments: executive leadership (two participants where one was a CISO and IT director), infrastructure operations (six participants who served various roles as network architects, database administrators and system administrators), security operations (three participants including SOC analysts and incident responders), and OT/Clinical engineering department (four participants including biomedical engineers and clinical technology specialists). This diverse sample ensured that the perspectives on assimilation and adoption decisions regarding the framework would be diverse as well.

Implementation of the Prototype. The prototype was deployed in advisory mode before the participants provided their views and insights on the operational suitability of the framework. The prototype implemented the Ensemble Assessment Engine using the six machine learning models previously discussed alongside the JSD calculation logic to generate severity predictions

based on calibrated uncertainty measures for vulnerabilities that had affected the organization's assets. Ensemble models typically achieve improved output compared to single machine learning models, which supported this design choice (Akano & James, 2022). The organization's vulnerability feed contained 847 vulnerabilities during the four weeks before this implementation, and each vulnerability was run through the ensemble to generate an isolation recommendation in graded form.

Developing the Survey Instrument. The researcher employed a survey instrument to collect participant responses across qualitative and quantitative constructs. In this case, the investigator constructed a 13-item tool, which evaluated the success of the framework on a number of constructs (see Appendix A for the complete instrument). Q1 and Q2 were demographic questions whereas Q3 assessed whether the framework was satisfactory. Q4 was aimed at establishing the ease of use of the framework, whereas Q5 assessed uncertainty interpretability by establishing whether JSD scores correctly separated reliable and unreliable assessments. Q6 determined the existence or absence of graduated protection correspondence by posing a question on whether the levels of isolation were in agreement with the risk levels in the organization. Q7 assessed temporal appropriateness by evaluating whether there was speed improvement compared with the standard triage that is usually witnessed in the operations of the organization. Q8 focused on coordination reduction and examined whether there was situational awareness improvement through the topology visualization process. Q9 and Q10 focused on automation trust and recommendation intent by the participants, and Q11 sought their views on the framework's performance compared to the current practices they had in the facility. Q12 and Q13 followed a qualitative feedback approach. Items Q3 to Q10 used the five-point Likert scale

based on the Strongly Disagree = 1 and Strongly Agree = 5 range, and Q11 had a similar Likert scale range (from Much Better = 5 to Much Worse = 1) to determine the comparative rating.

4.8.3 Quantitative Results

Descriptive Statistics. Table 3 below presents the descriptive statistics from the 15 participants based on their responses to Q3-Q10.

Table 3. Descriptive statistics for the survey items.

Construct	Item/Question	Mean	SD	% Agree
Requirements	Q3: Framework meets requirements	4.13	0.83	86.7%
Usability	Q4: Ease of use	4.20	0.77	80.0%
Uncertainty	Q5: JSD distinguishes reliable/unreliable	3.87	0.92	66.7%
Protection	Q6: Isolation levels match risk	3.93	1.03	73.3%
Speed	Q7: Faster than standard triage	4.20	0.68	86.7%
Awareness	Q8: Topology improves awareness	4.27	0.70	86.7%
Trust	Q9: Trust automated implementation	3.60	1.24	60.0%
Recommend	Q10: Would recommend adoption	4.13	0.74	80.0%

Figure 17 visualizes this trust gap across departments, showing that while 80% of practitioners would recommend adoption, only 60% trust fully automated implementation, with Security Operations showing the highest automation readiness and Executives showing the largest gap.

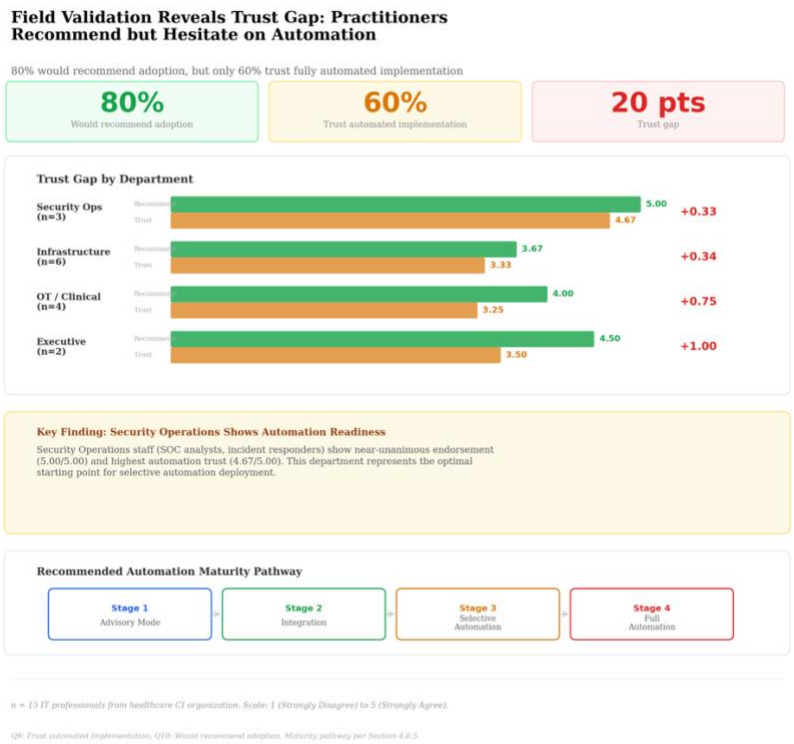

Figure 17. Field Validation Reveals Trust Gap

The Trust Gap Finding. One of the most significant findings from the descriptive statistics above was the lack of trust in the system as an automated solution to addressing system vulnerabilities, as seen in Q9 of Table 3 above. Participants agreed that the framework accelerates the vulnerability triage (Q7), and that it improves the situational awareness to vulnerabilities (Q8), but interestingly, a large number does not trust the framework as an automated solution to address the challenges independently. This is further apparent when comparing the trust in the framework for automation (Q9) and the intent to recommend it for adoption (Q10), which shows that despite the low trust levels, a large number of participants (80%) would still recommend it. This implies that participants do recommend the use of the framework, just not as an automated enforcement for protecting the system from vulnerabilities independently. Table 4 below shows how the trust gap persists per department.

Table 4. Trust gap per department in the facility.

Department	Q10 (Recommend)	Q9 (Trust)	Gap	Interpretation
Executive	4.50	3.50	+1.00	Endorse but require human approval
OT/Clinical	4.00	3.25	+0.75	Endorse concept but distrust automation
Infrastructure	3.67	3.33	+0.34	Moderate endorsement and trust
Security Ops	5.00	4.67	+0.33	Full endorsement and high trust

All the security operations participants highly recommended the framework, either agreeing or strongly agreeing with trusting it as an automated protection tool and recommending it for use. However, the participants from the infrastructure department expressed significant skepticism, likely due to their concerns about the impact of an automated framework on the other automated network configurations and settings. Among the OT/Clinical department, it is apparent that high internal variance exists between recommending the framework and trusting its use as an automated protection system.

Executive-Operations Divergence. It was also critical for the study to determine the relationship between leadership and their perceived view of the operational success of the framework. To achieve this, the study analyzed how the two executives who participated in the study responded to three questions: Q5 on whether JSD distinguishes vulnerabilities reliably versus unreliably, Q7 on whether they considered the framework faster than standard triage, and Q9 on whether they trusted the implementation of the automated system, comparing their responses to those in the Security Operations department. Table 5 below presents the divergence in this research area.

Table 5. Executive versus operations divergence.

Item	Executives	Security Ops	Difference	Interpretation
Q5: JSD distinguishes	4.50	3.40	+1.10	Significant divergence
Q7: Faster than triage	5.00	3.80	+1.20	Significant divergence
Q9: Trust automation	3.50	3.30	+0.20	Convergent skepticism

From Table 5 above, it is evident that the executives overestimated the effectiveness of JSD interpretability and the level of speed improvement when compared to the participants from the security operations department. However, interestingly, the two departments converge on their skepticism for trusting automated implementation, indicating that there is high institutional awareness of the possible operational constraints related to using this system.

Comparative Assessment. Lastly, the study conducted a comparative assessment on whether the participants considered this much better or much worse as a tool to manage vulnerabilities in their system compared to the existing vulnerability management frameworks applied by the organization. The results show that 80% of the respondents found this framework much or somewhat better than the existing organizational practices. No participant rated the framework as Worse or Much Worse.

4.8.4 Qualitative Findings

The two open-ended questions were focused on understanding the trust gap and identifying the implementation requirements from the participants' perspectives. Upon analysis of these responses, four themes were identified as described below:

Theme 1: Compliance Communication Value. This was the first theme, identified mainly by the executives who noted that the framework had presented an unintended but beneficial value proposition: it had created compliance documentation. According to the executives, delays in the timelines for patching firewalls are a commonly asked question by auditors (CISO notes), and thus, by having the high-uncertainty score, they are better placed to justify their choices on the maintenance windows. This theme and feedback show that the JSD scores are a valuable

metric that is able to satisfy the specific organizational requirements in terms of risk analysis, because it successfully provides a strategy to convert the technical tool into a regulatory defense.

Theme 2: Enforcement Proximity Anxiety. This was expressed by the participants in the infrastructure department and they observed that they were not comfortable with the automated enforcement of the system as any misconfiguration of the system would generate instant operational effects. A network architect responded by stating that automating GUARDIAN could cause the VLANs to be misconfigured by the management, and thus the administrators would get locked out of the system. This feedback illustrates the potential of the framework to interfere with operations and security systems of the facility, while also illustrating that those directly operating and managing systems are the first to experience direct automation risk before the rest of the system is impacted.

Theme 3: Clinical Safety Requirements. The third theme identified was clinical safety requirements, where participants (OT/Clinical staff) noted that the framework had the potential to violate non-negotiable safety constraints. According to one of the staff, using an automated isolation framework when a patient is undergoing a scan presents a safety issue because if the framework happens to isolate that part of the system during the scan, then the patient is at risk. The staff added that its use would require a clinical lock feature. This finding shows the role of the FDA regulations on medical devices as previously mentioned and illustrates how the functionalities of the framework would cause operational limitations and lead to risk for the patients.

Theme 4: Tiered Automation Acceptance. The fourth theme identified was from the practitioners stating that trust is not a binary element, especially in this case, but rather a tiered one. This was evident in the way respondents accepted automation on endpoints/workstations but

noted that human approval would have to be utilized in the servers and other core infrastructure of the facility. This theme suggests that the best way to implement this framework and eventually get to automation is by starting with low-risk assets, then gradually rising to high-risk assets after users have trusted the framework.

4.8.5 Interpretation

Principal Takeaways. The trust gap that exists from the responses by the participants serves as a validation for this research design because it confirms that the advisory mode is the right entry point for critical infrastructure after simulation and all other technical tests have been done. From the findings presented above, it is evident that several participants who interacted with this framework (80%) considered it worth recommending. This is especially significant considering that all the stakeholders included in this study hold senior positions in security facilitation, cybersecurity risk prevention, and decision making at the facility. With 80% of these participants recommending the framework, this is the first evident sign that the assessment and components recommended are valuable in addressing the presented challenge immediately. The lower automation trust scores, however (60%), show that there is a need to gradually build up towards getting people to trust in this system, and embrace it as an automated protection framework providing protection against vulnerability exploitation. This also illustrates that the framework needs an automation maturity pathway.

The Automation Maturity Pathway. The recommended maturity pathway to automation includes four stages. The first stage, the Advisory mode, is the current mode of the framework, where the developed framework is presented to a team of analysts for approval and feedback. This is referred to as the field validation stage and focuses on generating recommendations and interacting with the organizations and facilities that are best suited to utilize the framework. This

stage of the GUARDIAN framework was successful, as 80% of participants deemed it a worthy framework to recommend. The second stage is the integration stage that deals with the integration of automated assessment with the existing change management workflows within the organization through triggers such as the creation of IT service management tickets that require human approval. In this stage, the framework participates in generating recommendations and automatically creating tickets while enforcing triggers on the ticket approval within the existing workflows. It also ensures that infrastructure problems that can be anticipated, especially those relating to configuration control, are taken into account during the automation evaluation period. The third phase, the selective automation phase, aims at supporting selective automation on low-risk endpoints and then transitioning to the high-risk ones. This is significant since, as noted above, participants and analysts may eventually learn to trust a framework but this must be developed. The low-risk endpoints enable the users, analysts, and practitioners to interact with the framework without worrying about its ability to generate critical system, security, and safety issues, and this provides an opportunity to build trust. As clearly seen in the above results, some departments and individuals would quickly gain automation trust and some of them are the security operations that have a 4.67/5 level of trust and this gives an indication of selective automation readiness in the operations area. Once trust has been established and users have worked with the framework, the fourth and final stage of maturity is to implement full automation of various devices, considering the exclusions required depending on the organization and industry and excluding them permanently. For instance, in clinical devices, it is necessary to have a manual override capability to facilitate system operations during exceptional/critical circumstances. During this final full automation stage, the framework has to

be calibrated to make provisions for such exclusions, without creating new vulnerability pathways.

4.8.6 Limitations

The one notable limitation in this field validation study is that it is based on the perspectives of practitioners in one critical infrastructure sector and facility. Healthcare is among the most critical infrastructures; however, there are several others where practitioners might have different perspectives on the implementation of the automated graduated framework. As such, while the current findings significantly emphasize the positive and validated application of the GUARDIAN framework in practical environments, these findings cannot be generalized to other organizations/facilities in other critical infrastructure environments. Additionally, it is important to note that the validation is based on advisory mode operation using a prototype, which means that practitioners have not experienced first-hand the actual impact of automated enforcement, which could change their perspective and acceptance rate in either direction.

4.9 Chapter Summary

In this chapter, the GUARDIAN framework was validated, and the findings of the empirical research, simulation validation, and field validation were presented to determine the practical applicability of the framework. To begin with, the empirical analysis of the 279,056 CVEs demonstrates that no matter which model was implemented among the six machine learning models, i.e., Random forest, Support vector machine, Logistic regression, Neural network, LightGBM and XGBoost, there was a limit to the balanced accuracy to be achieved by the models, 69.3%, which indicates that information deficiency and not algorithmic limitation constrained performance. Uncertainty was then quantified using ensemble disagreement and the

research could then confirm that disagreement between the models using JSD is a strong uncertainty signal to identify unreliable assessments.

Statistical analysis of the study has allowed proving the fact that the uncertainty signal and the risk of exploitation are in a monotonic relationship that has also contributed to the further justification of the need to use a graduated framework in the process of protecting the critical infrastructure systems. Based on industrial data regarding exploitation times from four research reports by Mandiant, VulnCheck, Qualys, and Rapid7, further simulation testing showed that the use of temporal self-correction in the framework was feasible. The multi-signal integration strategy significantly enhanced the system protection, minimizing premature rollback errors by 91 percent relative to the time-only baseline strategy. Based on these findings, the study went further to pursue field validation to establish whether practitioners in a healthcare setting (which is one of the critical infrastructure settings) would be willing to use, trust the automated framework, and recommend it. The study results also revealed that the practitioners in the healthcare facility are willing to utilize this framework, provided that it is implemented in a tiered fashion that takes into consideration, respects, and accommodates the safety constraints and exclusions in the system.

In the final analysis, the findings mentioned in this chapter have confirmed that the GUARDIAN framework is a technically feasible, operationally acceptable, and viable solution that, via a staged implementation roadmap, can be integrated into critical infrastructure settings to assist in alleviating the ongoing crisis response challenges whereby attackers are faster to take advantage of the disclosed vulnerabilities than organizations are able to respond with solutions. The implications of these findings on automated protection and resilience of critical infrastructure are addressed in the next chapter.

CHAPTER 5: SYSTEM ARCHITECTURE

5.1 Introduction

This chapter details the system architecture of the GUARDIAN framework, describing its technical specifications. It builds on the empirically validated findings from the previous chapter using ensemble disagreement to address uncertainty and translates those findings into a deployable architecture that can be used to resolve the temporal impossibility in critical infrastructure systems. Chapter 3 presented a research approach to establish whether there was a way for organizations to respond quickly to vulnerabilities, without jeopardizing the quality of systems, based on the necessary timeline to develop and implement patches, as well as the speed with which exploiters target systems after vulnerabilities are disclosed. Chapter 4 then presented findings to illustrate that there was a way, using an adaptive isolation approach that responded to the vulnerabilities based on their severity to determine the timeline and speed of isolation. Additionally, in that chapter, the study validated that implementation of this solution is actually possible, through a simulation validation and a field validation with the potential stakeholders. This chapter explains how this framework, developed through simulation and validated by statistical and field research, can be implemented. It covers the GUARDIAN framework's core design principles, the architecture of the system components, platform configurations and deployment, and validation metrics necessary for successful implementation.

5.2 Design Philosophy: Core Design Principles

Cybersecurity is a component of every organization, largely because it is associated with the bottom line of the company as it may lead to reputational, financial, and economic damages.

It is more essential in critical infrastructures as it affects the health, safety, and security of the people, and economic status of a nation, and may lead to greater revenue or cost depending on the management of situations (Barrett, 2018). Consequently, when introducing a framework that acts as a safety mechanism with metrics to trigger appropriate responses based on perceived vulnerability levels, providing short-lived protection without requiring human coordination that typically causes operational limitations and delays, underlying design principles are essential. This section describes the main design principles that constitute the core of the GUARDIAN framework, drawing on Barrett (2018) who emphasizes the importance of foundational principles in framework development. The GUARDIAN framework has seven core design principles, focused on operational effectiveness without compromising on the technical feasibility proven in the previous chapter, described below.

5.2.1 Integration over Re-Implementation

The first core principle is that the GUARDIAN framework utilizes the already developed platforms that are commercially embraced and recognized, through the standard APIs, instead of focusing on developing new platforms for the same purpose. Organizations, especially those operating in the critical infrastructure sector, have already invested heavily in their security infrastructure, and this approach ensures they do not need to incur new costs to replace a system with one that does the exact same thing. By using APIs to ingest data, the system is able to focus on a less complex, more focused approach of measuring uncertainty and automating responses (Chandramouli & Butcher, 2025). As such, systems already integrated into critical infrastructure environments (such as Qualys and Tenable) are used in developing this framework.

5.2.2 Uncertainty-Driven Automation

The second core design principle of the GUARDIAN framework is uncertainty-driven automation, because traditional security automation processes redirect high-uncertainty cases to human cybersecurity analysts, and this framework intends to circumvent this by applying an automated graduated protective response instead. This is aimed at preventing the bottlenecks that come in the conventional cybersecurity system, where, as the system remains on hold as people review and organize a response to the attack, the more advanced attackers use the period to exploit the vulnerability already published within minutes (Qualys, 2024a; Charrier & Weiner, 2024). In the GUARDIAN framework, however, an uncertainty event is considered a signal of risk necessitating a graduated response based on the severity it poses, with high uncertainty events triggering quick isolation, while low uncertainty events trigger a slower response in terms of duration before protection sets in. This ensures there are no opportunities left open in the system for the attackers to exploit.

5.2.3 Temporal Self-Correction

The third core design principle is temporal self-correction, implemented with a fail-open logic, which is the risk mitigation strategy used to ensure that no operational disruption occurs while the uncertainty is being resolved. This is integral because sometimes an automated response means the system will shut down and potentially become unrecoverable. The GUARDIAN framework avoids this by implementing the fail-open logic, where the isolation practices follow a risk decay pattern over time unless new confirmed exploitations are identified in the system, in which case the system renews the heavy isolation procedures. If, over time, evidence shows that there are no new exploits, then the GUARDIAN framework mandates that the system should be designed to relax its restrictions.

5.2.4 Graduated Protection

The next core design principle is graduated protection responses, as a replacement for the binary responses often found in industrial control systems, either blocking or allowing access. Industrial control systems use programmable logic controllers (PLCs), and these are programmed to have binary responses in most cases as a protection mechanism, especially because availability is paramount in these systems as well (Stouffer et al., 2023; Humayed et al., 2017). In the GUARDIAN framework, however, the response approach is graduated, such that the level of attack severity determines whether the response is a simple monitoring or full system quarantine. This graduated scale approach ensures that the system can respond appropriately to the severity level when a severe attack arises, while also recognizing the criticality of the operations in the critical infrastructure sector and ensuring their availability where possible.

5.2.5 Evidence-Based Decision Making

The GUARDIAN framework also prioritizes evidence-based decision making, where a multi-signal approach is used to predict the threat level (based on data published about other known CVEs by CISA and the NVD catalog) using an ensemble machine learning model. This approach ensures that the framework makes decisions based on an algorithmic model that is both authoritative in terms of prioritizing accuracy and direct in using external and publicly available data, such that the decisions made by the framework are informed by trends based on past real-world threats and attacks.

5.2.6 Platform Agnostic Enforcement

Another core design principle for the GUARDIAN framework is platform-agnostic enforcement, achieved by using an adapter pattern that can facilitate independent deployment of infrastructure without being controlled by firewall vendors. This is a key aspect because if the framework is dependent on the workings of various other firewalls or vendors, or if the framework cannot adapt to those vendors, then using it as part of a security architecture will introduce new vulnerabilities in the system (Opara-Martins et al., 2016). There are common firewalls used by critical infrastructures such as VMWare, Cisco, and Illumio, and the framework has to be able to convert its high-level safety policies into these platforms for the organizations; otherwise, vendor-specific firewalls will create lock-in.

5.2.7 Comprehensive Observability with Override Capability

Last, the GUARDIAN framework has to embrace oversight. Automation does not mean eliminating oversight, especially not human oversight (Langer et al., 2021). While the GUARDIAN framework ensures that the delays often arising from human coordination are eliminated, it does not eliminate the need for oversight. Instead, this framework applies the glass-box approach to ethical AI implementation as described by Aler Tubella et al. (2019), ensuring that real-time data dashboards that illustrate the JSD calculations and logic path are always evident, and the audit logs capturing how decisions were made are stored and available for review. This also includes a strategy to manually override the capabilities and response decisions when necessary, ensuring that operators can override the automated decisions when situations call for it, based on their professional and contextual judgement.

5.3 System Component Architecture

The GUARDIAN framework needs an architecture that allows for scalability, fault isolation, and horizontal expansion when needed, while remaining interactive, responsive, and reliable. As a result, it is implemented on a microservices architecture because of the benefits such as fault isolation and independent scalability associated with it (Zhou et al., 2019). There are four different components: an ingestion component for data scanning from external sources, an assessment component to analyze and calculate the severity and uncertainty, an isolation component to facilitate the implementation of protective actions in the system, and an enforcement layer to ensure the rules of the network infrastructure are applied. These four components need to communicate despite each operating independently; thus, an event-driven messaging backbone is used (Apache Kafka). The components interact following the publish-subscribe pattern where the ingestion component normalizes the CVE events once they are published, then the assessment engine consumes, analyzes and publishes the results of the CVE event, the isolation component then publishes the policies to address the CVE event based on the findings from the assessment component, and lastly the enforcement component follows through and deploys the policies shared by the isolation component. To ensure the approaches can be audited, the system also deploys PostgreSQL to ensure persistent storage of the ingestion outputs, assessment outputs, isolation specifications, and enforcement outputs. The system also implements Redis to provide high-performance caching, which ensures that frequently accessed data is cached with sub-millisecond processing latencies, supporting audit trail requirements (Redis, 2025). Additionally, this architecture is decoupled to ensure that in case of a failure in one of the components, the system remains resilient while isolating the fault for resolution. This section describes in detail the four components of the GUARDIAN architecture.

5.3.1 The Vulnerability Ingestion Service

The first component in the GUARDIAN framework is the vulnerability ingestion service. As stated above, its role is to address the data heterogeneity challenge in the data received through APIs from the CVE catalogs, through normalization, turning it into standardized representations that are suitable for ensemble assessment. Data heterogeneity is a key challenge in this process because the different data types, formats, and structures can lead to challenges in integration and analysis, as well as model performance, meaning the insights from the ensemble would have errors (Gawlikowski et al., 2023). Commercial platforms focused on vulnerability management use proprietary schemas, data formats, and identification systems when recording and reporting vulnerabilities, and these do not align with those used in OT systems; thus, there is a need to translate in order to make processing possible using consistent raw formats.

As stated above, a core principle of the GUARDIAN framework is the use of platform adapters to obtain information from various platforms through integration instead of re-implementation. Examples of this are the use of a Qualys adapter to integrate the Qualys Vulnerability management API using the OAUTH 2.0 authentication to aid in retrieving vulnerabilities detected using REST endpoints and the Tenable adapter for tenable.io cloud integration using API keys, then implementing webhook reception for real-time updating of the data (Qualys, 2024b; Condon et al., 2023). Similarly, the Rapid7 adapter integrates with InsightVM through the Insight Platform API, providing a pathway for vulnerability and assets data correlation automatically (Condon et al., 2023). The results received from those platforms are then normalized into a single schema system that not only retains the data but also makes it easy to process uniformly, with the vital attributes being those of the CVE identifiers, the CVSS metrics and the textual descriptions necessary for NLP and all information related to the disclosure to make it easy to trace.

After data normalization, data deduplication is carried out, which aims at eliminating all redundant CVEs that may have been reported on different platforms, and this would affect the outcomes of the assessment stage. Duplication in this process would result in repetitive analysis in the asset processing, where vulnerabilities are being double-counted. Through deduplication, this is eliminated, ensuring that the context-specific variables are retained instead (Croft et al., 2023). After deduplication, this component also prepares the system for resilience by implementing state management synchronization tracking. NIST (2021) posits that this is an integral process in the ingestion process as it enables resilience against failures and ensures that even when there are breakdowns in the system, the resumption of services is progressive instead of starting afresh. This is integral in this framework, where duplication should be avoided.

5.3.2 The Ensemble Assessment Engine

The second component is the ensemble assessment engine, whose role in the GUARDIAN framework is to analyze the data, following the processes defined as part of the methodology when documenting this study. This component is responsible for the generation of severity predictions that are calibrated to measure uncertainty, through a series of processes as previously documented.

The first step in this process is extracting features and then executing parallel model inference. As previously described, the feature extraction process is conducted on the raw data forms to transform them into dimensional vectors that are suitable for the model inference (Corizzo et al., 2020). After this, the data undergoes text vectorization using the TF-IDF with a fixed list of vocabularies to capture key aspects such as vulnerability descriptions, while also identifying the critical severity indicators in the text using the term weighting process (Silvestri et al., 2023). After that, the CVSS metric decomposition is used to identify the attack vectors, the

complexities, the privilege requirements, and the impact scores using binary and continuous features, resulting in a structured feature set that reflects vulnerability characteristics. After this, the vectors are then simultaneously processed using the six models covered (Random forest, Support vector machine, Logistic regression, Neural network, LightGBM, and XGBoost), ensuring minimal latency in the system while also ensuring that the severity is collectively determined without introducing processing bottlenecks in the process (Verbraeken et al., 2020).

Once the severity of the vulnerability has been ascertained, then this component fulfills the most important element of the GUARDIAN framework, which is how it handles disagreement among these six models. This is the innovation of the framework architecture with the component computing the JSD to quantify disagreement among the models' predictions. The two conditions used in this process are that when all six models are in agreement, the JSD will be low, which indicates high confidence, and when they all contradict each other, the JSD will be high, which indicates a high uncertainty level. The JSD score obtained is then calibrated using temperature scaling due to its low complexity and efficiency, requiring simply an adjustment of the logits using a learned scalar, T, before the raw prediction values are converted to final probabilities using the softmax function (Wang, 2023). The value of T for this is 1.47. This ensures that the calculated uncertainty does statistically match the error rates identified, providing a trustworthy numerical signal indicating the system's confidence in a given risk assessment (Guo et al., 2017).

The component then uses threat intelligence after the model assessment to incorporate real-world exploitation evidence in order to identify exploitation levels. It queries the CISA KEV catalog, alongside other commercial threat feeds, to determine whether there are any active campaigns or indicators of exploitation and how vulnerabilities are being exploited, to determine

the real level of protection to trigger. The outputs of such an evaluation are shared as message queues with metadata attached for downstream processing and auditing, with the addition of the severity prediction scores, the individual model predictions, JSD values, and the threat intelligence scores.

5.3.3 The Adaptive Isolation Engine

The next element in the GUARDIAN framework is the adaptive isolation engine, which is the decision maker that implements the validated uncertainty-parameter relationships once the data has been evaluated to determine vulnerability severity. This element processes the results from the assessments and reviews the recommended levels and duration of isolation and afterward determines what components of the system infrastructure will be impacted by the protection and generates a platform-independent policy with these specifications. The automated parameter mapping was performed depending on the JSD scores as follows: If JSD score is less than 0.35, this is level 0-1, a low uncertainty level, and the system does nothing except monitor and possibly restricted communications. If JSD score is between 0.35 and 0.45, this is level 2, and the system has to balance between protection and operational continuity in its reaction and isolation. If JSD score is between 0.45 and 0.60, this is level 3, which is an escalation and necessitates either implementing a zero-trust micro-segmentation or partial quarantine of nonessentials. If JSD score is greater than or equal to 0.60, this is level 4, a high uncertainty level, which shows extreme exploitation risk and necessitates complete quarantine.

This is followed by an inverse application of isolation duration, whereby higher uncertainty results in shorter isolation duration, using this formula: Isolation Duration equals Base multiplied by (1.5 minus JSD score) multiplied by Severity modifier, where the base equals 48 hours, and the severity modifier ranges from 0.5 to 2.0. The low uncertainty allows for a

maximum period of 168 hours in order to balance operational stability and prioritization of security.

Asset correlation is the next step after determining the duration of isolation, which assists in ensuring the protection being applied aligns with the context and structure of the organization (NIST, 2024). The protection scope is established by matching vulnerability signatures with infrastructure inventory to identify the type of response that is needed and adjust the isolation parameters based on the criticality of that particular setting. After that, the component then conducts a temporal configuration process, which determines the self-correction and relaxation policies of the decay logic based on how the vulnerabilities are exploited. This involves determining the timeline for how long the protection should be relaxed, and the specified duration that should pass without a new exploitation of the vulnerability being recorded, or a patch being available for the system to begin rollback. The decay rate recommended based on the simulation is 0.02 per hour. This is a key step in this process to ensure that the isolations do not accumulate in the system, such that the network is clogged due to the restrictive rules forgotten in the system. All this data is then generated as a policy, as was done with the previous engine.

5.3.4 The Enforcement Orchestration System

The final element of the GUARDIAN framework architecture is the enforcement orchestration system, whose role is to translate the policies that are published by the previous engine into platform-specific configurations that can be technically executed in the system. It serves as the universal technical translator of the entire network infrastructure. This includes integrating the platform adapters and deploying the safety logic. The integration of platform adapters is based on standardized interfaces recognized in the industry, in order to ensure the system can also accommodate the vendor-specific needs as much as possible. Thus, three

adapters are selected: The Illumio adapter serves as the translator for policies defined above into label-based enforcement groups using REST APIs with OAuth authentication (Illumio, 2025a). The VMware NSX adapter translates the policy into distributed firewall rules, defining the security policies and groups using the NSX-T policy API (Broadcom, 2025b). The Cisco adapter translates the application policies into application policy contracts using Google Remote Procedure Call (gRPC) interfaces (Cisco, 2025). This ensures that the core of the GUARDIAN framework is able to remain platform agnostic, allowing organizations to easily switch hardware vendors without needing a complete rewrite of their automated defense logic.

Once this is done, the next step for the component is to perform a safe deployment of the logic, which demands a two-phase commitment protocol to prioritize safety. The first step is to ensure that the firewall currently used by the organization's security system can accept the new rule based on the assessment decision and determination of adaptive isolation. This is integral because it ensures there is not going to be a deployment challenge where one system locks the other out. Once this confirmation is done, the system is then able to commit to implementing the recommended adaptive isolation response from the framework. With that capability set up, if the platform reports an error, either because the device is offline or due to similar issues, the system is set to roll back in order to avoid deploying in a broken or partial security state. Otherwise, the capability proceeds as defined, with the system monitoring for any changes to determine the health of the connections and be ready to implement isolation using circuit breakers if a tool, such as a firewall controller, is unresponsive, preventing a cascade of errors. When circuit breakers succeed in health checks, automatic recovery can be used. With this, the automated response system is ready to deploy if necessary. Figure 18 illustrates this four-component

architecture and the publish-subscribe data flow that maintains sub-10-second processing latency.

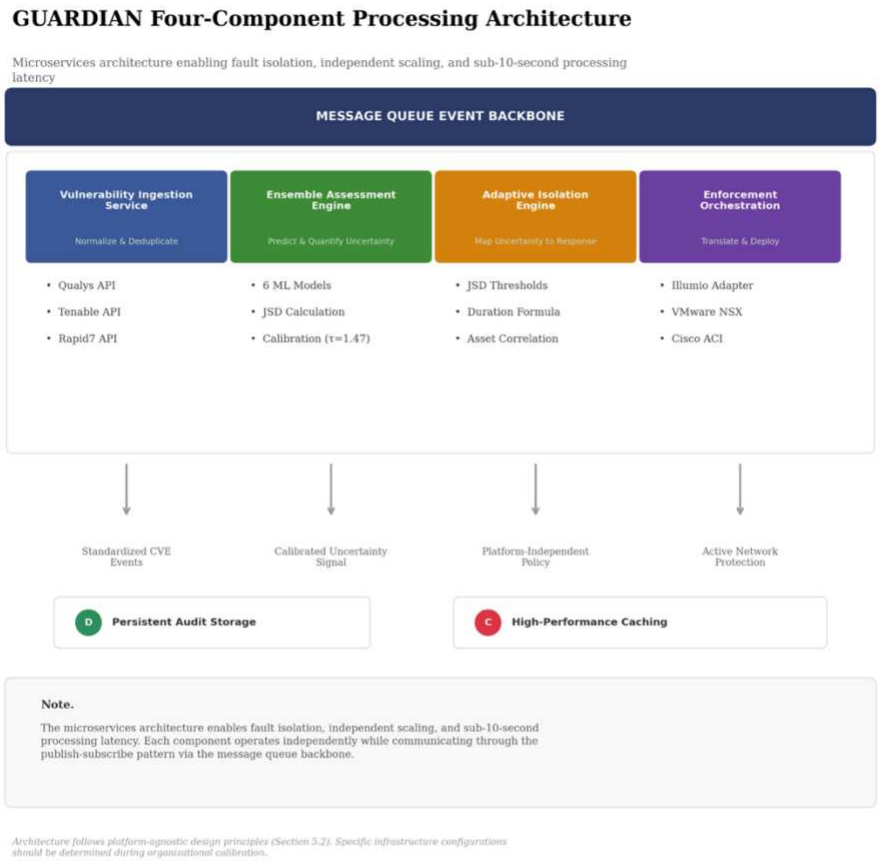

Figure 18. GUARDIAN Four-Component Processing Architecture

5.4 Operational Deployment Considerations

Beyond the configuration of the system components, the other crucial step in the GUARDIAN framework is the deployment phase. This is because critical infrastructure companies require a generally cautious approach when implementing any new systems, but this

is especially the case for an automated response system, because errors could cascade through the system and result in more problems than the potential solutions it offers. As such, the strategy here is to tailor the deployment to the specific organization, ensuring that the infrastructure requirements, organizational integrations, and the technical capabilities assumed by the framework all align.

The GUARDIAN framework has a number of infrastructure requirements that will make this system computationally efficient, yet robust. These requirements include sufficient computational resources for parallel model inference, message queue infrastructure for fault tolerance, and persistent storage with adequate retention capacity. These requirements ensure system resilience through fault tolerance and data persistence even when there is a network glitch or a system failure.

Beyond the technical requirements, the organization also needs to be prepared in terms of process, personnel considerations, and roles in implementing the GUARDIAN framework. While the specified requirements, as well as the defined system component architecture, ensure that the system will function as intended, the human and process elements are still key factors in its successful deployment. The organization has to prepare the personnel through training on uncertainty interpretation, policy overriding approaches, and incident response integration strategy. Elements such as change isolation, automated isolation impacts, rollback criteria, and network operations have to be clear and understood by these teams for the graduated protection levels to work successfully. Additionally, the operational trade-offs that come with these risk-reduction benefits documented in the study also have to be clear to key stakeholders, to ensure expectations are clear before deployment.

To implement the deployment process, a staged approach is recommended for validation before full automation with oversight is embraced. A staged approach is a process where the next project step is launched based on the outcomes and lessons from the previous one (Buttrick, 2009). In this case, the staged validation approach is recommended to take at least three months in each phase, to track outcomes before proceeding to the next phase. Phase 1, which is monitoring only, should last 3 to 6 months, whereby the system is deployed, the system connects to the network and generates assessments and policies, but instead of proceeding to the enforcement process, this component is switched off. This stage is aimed at observing the recommended responses and also reviewing the logs to ascertain whether the decisions make sense in relation to the vulnerabilities being handled. Based on the analysts' findings, the project can then proceed to the next phase. The second phase after monitoring-only is enforcement with human approval, and this phase should span 3-6 months, where the system is permitted to carry out enforcement, but analysts must approve protective responses before they deploy. This step is meant to build the analysts' trust in the system while also validating that it will not cause disruptions in organizational operations.

The next phase is graduated automation, which should last between 6 and 12 months, whereby the automation is applied selectively across low-risk assets such as workstations, with the high-risk assets remaining under manual operation and approval (phase 2). This further provides the analysts with an opportunity to experience the impact of the GUARDIAN framework in the system while also underscoring its ability to address the vulnerabilities and protect the system, without creating new challenges for them. The longer duration is recommended to cement the trust in the model, as well as to see how the model would operate over longer durations. The last phase is full automation, where the system is now fully set up to

run autonomously but with human oversight. In this phase, the analysts and experts in the security department still have the ability to override the decisions in the environment if the need arises. Figure 19 presents the four-phase timeline, with durations calibrated to the trust gap identified in field validation.

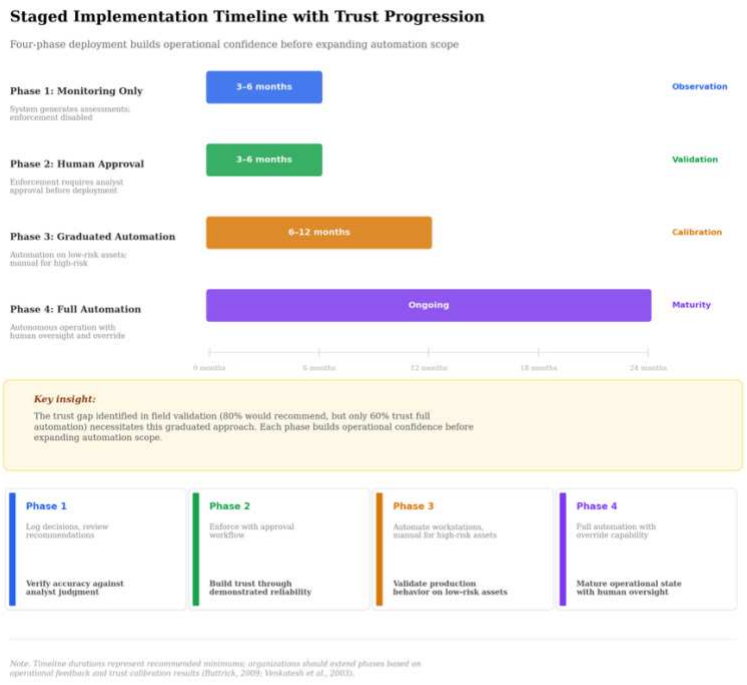

Figure 19. Staged Implementation Timeline with Trust Progression

5.5 Validation Metrics and Success Criteria

The ability to translate the theoretical vision and performance into practical and operational contexts has been one of the most important aspects of this study since it is these that will define the ultimate effectiveness of the GUARDIAN framework. In this respect, several metrics can be taken as target thresholds which, when not met, would decrease the chances of attaining the simulation results or adhering to the industry benchmarks. These are covered below.

The design of the system component architecture mentioned a sub-10-second latency as a requirement, and the system can handle assessments without imposing delays that create exposure windows. This is a design factor that is a major determinant to other performance measures. To begin with, assessment processing time must be under 10 seconds, at the 95th percentile, and this is the time between the ingestion of the CVE and the completion of the assessment process. The delay between production of the policy and the platform accepting it must also not exceed 5 seconds, and the processing capacity should be able to handle 500 CVEs per minute, as occurs in Log4Shell-magnitude events. These temporal performance constraints guarantee that the framework has the capability to respond quicker to a vulnerability than the break-out time of contemporary attackers.

Such situations also demand accuracy measures so that the automated decisions generated by the framework can be in agreement with expert judgment. These include a 95 percent correctness rate on the policies following human audit, and a false positive rate of 15 percent or less to prevent premature rollback and focus on operational continuity and system security. Moreover, a lower than 20% override rate would help, though not a requirement, since it would mean that the automated response system is more reliable and that the analysts do not have to continually interfere with it through manual means.

Lastly, operational stability metrics also exist that demand that the API integration should be operational 99% of the time. The system is also to be configured to automatically go into safe mode in the event the automated defense system is offline so as not to cause any impact to the rest of the system traffic or become a single point of failure. Figure 20 summarizes these validation thresholds across temporal performance, decision accuracy, and operational stability.

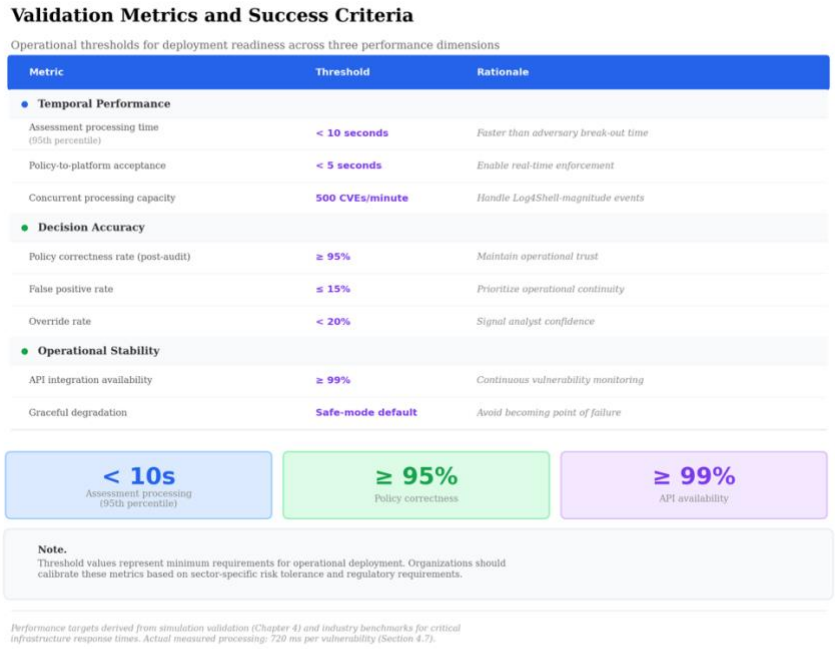

Figure 20. Validation Metrics and Success Criteria Thresholds

5.6 Chapter Summary

This chapter describes the system architecture of the GUARDIAN framework. It has discussed the seven major principles of the framework, which are integration instead of re-implementation, uncertainty-driven automation, temporal self-correction, graduated protection, evidence-based decision making, platform-agnostic enforcement, and comprehensive observability with override capability. All these principles help to transform the theoretical approach to the methodology of research, testing, and validation of the GUARDIAN framework into practical and robust software specifications. These have been implemented in four main system components to ingest data from CVE databases, perform assessment on the data after

normalization, develop the graduated response (adaptive isolation) parameters once the vulnerabilities are evaluated, and prepare for enforcement of the policies published across a microservices architecture that is both easy to scale and fault-tolerant. The described architecture can resolve the temporal impossibility facing critical infrastructure systems with the help of ensemble disagreements defining the isolation levels and subsequently applying graduated protection to those isolation decisions. In the process, it also recognizes the necessity of human oversight, leaving audit trails and monitoring capabilities to ensure that the decisions made within the framework can always be reviewed by human experts. This chapter has provided a detailed strategy of how the GUARDIAN framework can be designed to address the challenge faced by critical infrastructures where it is necessary to react fast, while also taking into consideration the industrial timelines, but avoiding exploitation of vulnerabilities immediately after they are published.

CHAPTER 6: DISCUSSION, IMPLICATIONS, AND CONCLUSION

6.1 Introduction

The aim of this study was to develop an adaptive-isolation framework that could address the irreducible uncertainty common in machine learning algorithms used to manage vulnerabilities in cybersecurity. This framework also required a mechanism for bridging the gap between the need for time to coordinate with organizational leads following a vulnerability occurrence, the industrial guidelines on when a patch is to be available and implemented, and the speed with which attackers exploit the zero-day window following a vulnerability being published. To fulfill this objective, the study had several research objectives and research questions. The first research question was to determine how machine learning models perform in terms of accuracy levels during vulnerability prediction, across different algorithmic compositions. This question was answered by conducting empirical analysis to illustrate that despite the algorithmic differences across six different models, the accuracy levels converged between 47.5% and 69.3% when the models were trained, validated, and tested using the same CVEs. Tweaking classification thresholds to favor recall is the preferred optimization in critical infrastructure environments where missing a single exploitable vulnerability can be catastrophic. However, even with recall-optimized thresholds, the accuracy ceiling imposed by the insufficiency of information in standardized vulnerability data remains. The GUARDIAN framework accounts for this by using ensemble disagreement to trigger graduated protective action independent of any individual model's threshold configuration. The second research question sought to determine whether ensemble disagreement provides an uncertainty measurement that allows the use of an automatically adjusted response without needing human

verification. The third research question was to determine what the suitable threshold parameters are for an automated graduated protection framework that can temporally adjust based on the identified uncertainty levels without the need for human monitoring, while maintaining system safety. The fourth research question sought to determine whether the graduated protection framework developed was feasible under simulated conditions of cyber-physical water treatment with set signal quality parameters. Lastly, the fifth research question sought to determine whether critical infrastructure practitioners found the framework operationally feasible to balance the security requirements while managing the operational constraints.

The previous chapters have answered the research questions through empirical study, simulations, and statistical as well as qualitative field validations. Additionally, a system architecture for the GUARDIAN framework was developed in Chapter 5 based on these findings. This chapter presents a discussion of the findings, analyzes the theoretical and practical contributions and implications, and discusses the deployment method for the GUARDIAN framework. Additionally, the chapter covers limitations and boundaries of the study in terms of validity, recommends future research directions, and presents a conclusion for this research.

6.2 Discussion of Empirical Findings

This study has comprehensively answered the five research questions and validated the development of the GUARDIAN framework through a simulation and a field study across a group of experts in a regional healthcare facility. This section discusses the findings in relation to the research questions and study objectives.

6.2.1 Validation of a Structural Accuracy Ceiling Due to Information Constraints

The initial research question asked how machine learning models perform in terms of accuracy levels during vulnerability prediction across different algorithm compositions, and this research established that information constraint is an important factor that leads to the convergent accuracy of the models. The six models used in the study represent different algorithmic approaches, which include: Random forest, SVM, Logistic regression, Neural network, LightGBM, and XGBoost. Despite these algorithmic differences, the accuracy levels of the models converged within a narrow range. The random forest model reached an accuracy level of 58.3% and a precision rate of 61.2%. The balanced accuracy and precision rate of SVM were 47.5% and 51.2%, respectively. The balanced accuracy of logistic regression was 54.2% with precision of 57.8%. The neural network had an accuracy of 62.7% and precision rate of 65.4%. LightGBM had the best-balanced accuracy of 69.3% and precision of 71.8%, and the last is the XGBoost model, which had a 67.1% balanced accuracy and precision rate of 69.4%. In critical infrastructure environments, recall is the more operationally significant metric because a single missed exploitable vulnerability can produce catastrophic failure. The recall rates across the six models ranged from 47.5% to 69.3%, meaning even the best-performing model missed roughly 30% of exploitable vulnerabilities. This concern is precisely why the GUARDIAN framework uses ensemble disagreement as the protective signal rather than relying on any single model to catch every threat. Beyond individual performances, the study also estimated the differences in model performance, and observed that between SVM and LightGBM the difference was 21.8 percentage points (the best performing and worst performing models), and between the two high-performing models LightGBM and XGBoost the difference was 2.2 percentage points, which demonstrates the ceiling in performance (approximately 70%) was necessitated by the nature of the available data rather than the models used.

Such performances are important in the sense that they show that although there are variations in the models, the accuracy does not rise above the 70% threshold. This finding aligns with Croft et al. (2023), who conducted systematic analysis of data quality attributes across major vulnerability datasets, quantifying how noise and sparsity in ground truth data prevent models from learning more complex patterns regardless of algorithm sophistication. The finding is further supported by Risse and Böhme (2024), who demonstrated through rigorous benchmarking that state-of-the-art models severely overfit to unrelated features rather than learning vulnerability characteristics, with performance dropping significantly when realistic constraints are applied. Earlier foundational work by Jacobs et al. (2021) and Sabottke et al. (2015) similarly identified structural information gaps that cannot be resolved through algorithmic tuning; however, these more recent empirical studies provide direct quantification of the data quality ceiling that constrains model performance.

Information insufficiency as the cause of the accuracy plateau between the models was further emphasized by the temporal stability analysis, which found that across the 23 years, there was less than 1% variance in model performance from the NVD data. This finding validates the views by Croft et al. (2023) that deep learning models hide the real issue of information scarcity through the illusion of complexity that they exhibit when processing inherently under-signaled data. Additionally, that level of temporal stability across over two decades of data illustrates the presence of persistent irreducible characteristics across the whole ecosystem, where the missing organizational contexts impact the quality and sufficiency of the data when conducting a vulnerability scoring. This also aligns with findings by Anwar et al. (2021), who conducted comprehensive quality assessment of the National Vulnerability Database and documented significant inconsistencies in severity rankings and weakness classifications, confirming that

such deficiency in contextual metadata is a crucial constraint on the analytical utility of CVEs as the foundation of prioritization, especially in applications with diverse technology stacks. As a result, the focus in enhancing the outcome of vulnerability prediction moves from the performance of the models in this study to finding ways of operationalizing the irreducible uncertainty that currently exists.

6.2.2 Illustrated Use of Ensemble Disagreement as a Measure of Calibrated Uncertainty

The second research question focused on determining whether ensemble disagreement can provide an uncertainty measurement that allows the use of an automatically adjusted response without needing human verification. Following the question on whether there was a way to operationalize the irreducible uncertainty that exists between the models, the goal of this research question was to determine whether the disagreement between the models could be used as a measurement metric, which, when parameterized, would serve as a measurable calibrated uncertainty value. This question was important because it focused on the ability of the ensemble across the six models to serve as a tool for developing a parameterized system, where the uncertainty was a valuable tool. It was central to the study because Bassi and Singh (2023) had previously opined that uncertainty measures have been underused in the wider cybersecurity community due to the historical application of deterministic scoring frameworks and binary alerting frameworks, which both tend to promote the belief that threats can be detected in a categorical manner instead of being probabilistically derived. This limitation is further elaborated by Geng et al. (2024), who provide a comprehensive taxonomy of uncertainty measures and explicitly state that quantifying uncertainty is essential because generating responses with appropriately calibrated confidence helps determine when the answer is trustworthy. The result of deterministic approaches is a gap where the legacy of determinism has impeded the use of

uncertainty-aware models, although the importance of confidence in prediction is increasingly being recognized as being more valuable than the prediction itself, especially in operations where the outcomes of operational choices are highly valued.

In order to fill this gap, this study conducted three key explorations: first, the use of the JSD distribution to measure the similarity between the probability distributions, and determine the level of disagreement across the models on the severity of a vulnerability (Lin, 1991; Nielsen, 2019); second, conduct a calibration quality assessment to ensure that when the uncertainty is used in an automated system, it has the ability to successfully provide reliable decision making where the identified probability of error predicted matches the actual frequency of error; and lastly, that there is a correlation between the disagreement and the accuracy level identified. The first step, on JSD distribution, found that when the 68,711 test CVEs were analyzed for ensemble disagreement, there was room to reliably conduct thresholding across this data. The JSD distribution among the CVEs gave it a mean of 0.400, standard deviation of 0.047 and coefficient of variation of 11.75%. The median JSD was 0.398 and this indicated that the distribution was symmetrical with no significant skew and that even when a very large number of CVEs are considered, all the models were in agreement as far as their analysis was concerned, hence the low value of JSD. However, another thing that became clear was that in the 95th percentile, many of the vulnerabilities were ambiguous and resulted in divergent conclusions amongst the six models. The temporal analysis had indicated no likelihood of temporal drift over the two decades, with variation over the test data (68,711 CVEs) showing less than 1% variation over the period of January 2023 to December 2024 and a mean JSD ranging from 0.394 to 0.406. This indicated that when the thresholds were set using these figures, there was potential stability across the thresholds calibrated.

The third step was to make sure that when the thresholds were calibrated, the decisions to be made relying on these calibrations possessed some probability of error that was as close to the true frequency of error as possible. This was significant since in machine learning models it is observed that there are instances where the models may be too confident in their level of accuracy thereby giving high values that are not accurate (Wang, 2023; Guo et al., 2017). In this study, the temperature scaling approach was applied where parameter $T = 1.47$ was learned on validation data and this gave an $ECE = 0.023$, which means the predicted error rate of the models is within 2.3 percentage points of the observed error rate. The effect of the temperature scaling was that it decreased the ECE from 0.147 (14.7%) to this value. When the accuracy between the ensemble-calibrated and the actual values was compared, it was also established that the accuracy predicted by the ensemble and that which was observed was off by a margin of less than 3 percentage points.

After that, the final step in answering this question was determining the relationship between the disagreement and the accuracy, and this was identified using a statistical analysis. The Spearman's rank correlation showed that the correlation between the two was $\rho = -0.92$ ($p < 0.001$), which meant that the ensemble had strong ability to identify unreliable assessments. This means that the calibration quality of the disagreement does facilitate automated decision making, because the predicted error rate deviated from observed error by only 2.3 percentage points, enabling the system to trigger protection responses at thresholds where the probability of premature action falls within acceptable operational bounds. This way, the uncertainty could be used to determine the graduated responses when applied in a framework.

Beyond answering the research question, this strong negative correlation between ensemble disagreement and prediction accuracy also filled a gap in the literature. According to

Lakshminarayanan et al. (2017), in the theoretical landscape, there is a need for a new framework that can fill a few gaps that exist in the existing systems, and which are tackled only partially and individually. This includes the lack of a model in the literature that is shown to incorporate machine learning uncertainty into automated or semi-automated protection measures, even though it is demonstrated that uncertainty measures can be useful to give meaningful information on the credibility of vulnerability forecasts (Lakshminarayanan et al., 2017). By treating disagreement not as noise but as a signal of unreliable assessment, this approach changes the concept of uncertainty from a nuisance to a tool that can serve for developing a protection framework for the system, thereby establishing the foundational concepts of the GUARDIAN framework.

6.2.3 Operational Feasibility of the Graduated Protection Response Due to Monotonic Parameter Relationships

Having determined that the ensemble disagreement can provide an uncertainty measurement that allows use of an automatically adjusted response without needing human verification, the third research question sought to determine the suitable threshold parameters for an automated graduated protection framework based on the identified uncertainty levels that can be automated to temporally adjust without the need for human monitoring, while maintaining system safety. This question required that the relationship between the parameters be determined and then be assessed in terms of their operational usability in a graduated protection framework as the determinants of the response. This was important to address as it answers the concerns brought up by Rose et al. (2020) and Basta et al. (2021) on zero-trust architecture. Rose et al. (2020) posit that ZTA, regardless of the conceptual attractiveness, provides minimal advice on how systems need to vary in case of uncertainty or other unclear risk situations, assuming correct

confidence scores based on contextual analytics but not providing a structure to deal with cases of incomplete, contradictory, or missing telemetry. Basta et al. (2021), on the other hand, noted that systems that use advanced solutions such as microsegmentation technologies to circumvent the need for graduated protection strategies often work in binary states in which workloads are either completely permitted or completely denied access, thus offering only a few intermediate states representing different levels of risk or confidence. According to the same authors, these systems had a deficiency in their graduated levels of enforcement, which resulted in minimized power of the segmentation controls to adjust protection proportions to prediction unpredictability, which is worsened in a setting where the cost of a false positive is expensive. As such, it was imperative to determine whether the use of disagreements based on JSD distributions provided a pathway to address the binary limitations of the current ZTAs.

For this research question, the study first sought to determine how the disagreements based on the JSD could be used to predict the exploitation rate while ensuring that system protection was implemented in time before the attackers could exploit the vulnerabilities once published. The first step was to determine the exploitation rate and how that related to uncertainty across different levels. This included using real-world data in order to be sure that the performance of the ensemble was in line with what would occur in the real world. The study verified the exploitation status of CVEs within each uncertainty band using VulnCheck KEV, CISA's Known Exploited Vulnerability Catalog, and EPSS to determine whether the exploitation rate correlated with the uncertainty. The analysis found that there was a positive correlation between exploitation rate and the uncertainty (with a Spearman's rank correlation of $\rho = 0.89$ ($p < 0.001$)), and that as the uncertainty value rises, the likelihood of exploitation also increases, rising from 8.2% in the low-uncertainty band to 31.2% in the high-uncertainty band. This finding

was foundational for this research because it served as the empirical basis for why the graduated protection approach is needed, especially because as the level of uncertainty rises, the probability of exploitation also increases, nearly quadrupling in the high-uncertainty band. It also illustrates that the higher the complexity of the vulnerability, the more likely it is to be targeted for exploitation.

In addition, this research question, beyond illustrating that uncertainty levels were indeed suitable to use in determining the graduated response of a critical infrastructure system, sought to validate this in the development of the framework. As such, the next step to cement the findings for these research questions was determining how the protection stringency and the protection duration were related. The need for protection had been proven, as seen in the previous section, and it had become apparent that as the level of uncertainty rose, so did the risk of exploitation, meaning that as the level of uncertainty rose, the need for protection in the system rose as well. The question, however, remained on the duration during which that protection would be implemented. As stated, other systems that are currently in practice conduct a binary response strategy where a risk is either allowed or blocked, which, while effective sometimes, limits the ability of the organization to continue operation (Xiaojian et al., 2021; Rose et al., 2020). This is consequential, however, because, as previously highlighted, critical infrastructure systems constitute essential infrastructure for national security operations, and these represent primary targets for adversaries seeking to destabilize national operations, as they define the state of key national pillars (Hemme, 2015; IBM X-Force, 2025). So, it was imperative to develop a way that these systems could stay in operation even during an attack when a vulnerability is detected, without that causing a national risk.

To address this, the relationship between JSD and isolation levels, which had shown a correlation of $\rho = 0.89$ ($p < 0.001$), was similarly correlated with the protection duration based on what has been described by industries as the right time to react to such vulnerabilities. The findings of such correlation were a Spearman rank value of $\rho = -0.87$ with $p < 0.001$ indicating that a similar though inverted relationship was present between the two. This implied that high-uncertainty vulnerabilities exhibit exploitation patterns concentrated in the first 24-48 hours because attackers prioritize targets where defensive ambiguity maximizes success probability, hence requiring shorter initial timelines to spend on isolation and continuous assessments, and low uncertainty cases are on the opposite end, in most cases aligned to longer exploitation patterns (Qualys, 2023, 2024a; Garrity, 2025). This justified the use of inverse duration in the GUARDIAN framework, where the logic for the architecture was simply that high uncertainty requires a short duration of isolation in order to protect the system and force a rapid assessment of the situation, while low uncertainty levels require a longer duration before isolation is implemented, as the level of exploitation associated with those is lower. Through this research question, the need for the GUARDIAN framework, and the recommended approach where a graduated response to the protection implemented had been identified as directly proportional to the severity of the vulnerability, which is determined by the JSD levels, while the duration for when the isolation is to be implemented had been determined as inversely proportional, as the higher levels of severity need quicker protection responses compared to lower ones.

6.2.4 Feasibility of the GUARDIAN Framework While Using Temporal Self-Correction

The fourth research question sought to determine whether the graduated protection framework developed was feasible under simulated conditions, considering a temporal self-correction approach is to be used. Answering this question was imperative to the study because it

determined whether the use of an automated graduated response would result in more challenges in the system, or had the ability to adapt to the lack of exploitation attempts, for instance, or increased exploitation attempts, such that a shift accommodating those needs and instead implementing rollback or a new immediate response could be done. It simply sought to determine whether the GUARDIAN framework was capable of adjusting its response as the evidence changed. This was a gap to address because, as research by Sommer and Paxson (2010) and Jacobs et al. (2021) highlight, traditional models assume risk as being stagnant or monotonic, but time-sensitive models have recognized the fact that the lack of exploitation, threat intelligence, or an anomalous action may be evidence of a depreciating risk condition, thus justifying the reduction of defenses.

The first step in this process was determining whether the rollback timing design can be applied in this context as a tool for ensuring the feasibility of temporal self-correction. To prove this, the study synthesized industrial research on the time taken to react to exploitation, and the ways in which attackers exploited a vulnerability once it was published, using reports on key vulnerabilities published: Mandiant, VulnCheck, Rapid7, and Qualys (Qualys, 2024b; Condon et al., 2023; Charrier & Weiner, 2024; Garrity, 2025). The four reports illustrated a bimodal distribution in the strategy applied by attackers, where there were exploits that were immediately launched upon the publishing of vulnerabilities, and then there were sophisticated campaigns that were launched after a long period following the vulnerability being published. This meant that a robust threshold design where both the immediate attacks and those that took longer to get launched were addressed, was necessary.

Once it was proved that thresholds were necessary to address the attack strategies applied by the attackers, the study then distributed the graduated thresholds across the uncertainty bands

previously determined through JSD distribution, such that in line with findings from the previous research question, the highest uncertainty band gained the lowest duration of protection/isolation response (24 to 48 hours protection window) while the lowest band received the longest duration (between 7 and 30 days). These were then tested for validation using a simulation environment, which showed that while there were different strategies that could be applied to facilitate the necessary rollback, the level of exploitation that was able to circumvent those pathways did not favor all the recommended strategies. The Time-only strategy was deemed unsuitable because, by applying a fixed 48-hour rollback period with no conditions, premature rollback occurred in 18.4% of scenarios, meaning at least 18% of exploits would be potentially successful. The time plus exploitation detection baseline strategy, which applied a 48-hour rollback period if no alerts were detected, showed a higher level of performance compared to the time-only approach, but the room for error was still allowing premature rollback in 7.6% of scenarios, with an appropriate rollback of 89.2% which was still a significant number of premature rollbacks to let through. The third strategy, which demonstrated robustness, was the multi-signal integration strategy, which included the use of graduated thresholds using uncertainty, as shown in previous sections and this research question, and this included several conditions such as ensuring that threat intelligence was absent, a perfect patch status, and exploitation detection had been prioritized. This resulted in a premature rollback rate of only 1.6%, which was a significant decay from the initial time-only strategy, and had an accuracy of 92%, increasing the efficiency of the framework and its ability to ensure that as few exploitations as possible were able to penetrate through the system following the identified rollback time. This finding was important in addressing the concerns raised in literature by Nguyen and Reddi (2021), noting that temporal models also have to have strategies for modelling the changing risk levels, and noting that some

recommended approaches include Bayesian or other reinforcement learning models. It also proves that automated response systems can find ways of implementing restrictive policies that ensure the implemented crisis response strategies do not become permanent limitations to organizational/system operations.

6.2.5 Practitioner Acceptance and the Trust Gap

The fifth and final research question sought to determine whether critical infrastructure practitioners would accept uncertainty-parameterized isolation recommendations for operational deployment. To address this, the study assessed whether experts and analysts in the field would be willing to implement this system physically, by conducting a field validation study, using the advisory mode. While different insights were shared across several departments, it was also apparent, based on these participants, that they were largely open to the implementation of the GUARDIAN framework, as long as the automation aspect would be limited to less critical aspects of the organization, such as client servers instead of operational systems that, when impacted, could limit their ability to operate. One of the main points of interest in this case was the existence of a trust gap, as people were not eager to accept an automated system, because there was some doubt concerning the abilities of the model. This finding aligns with research by Pollini et al. (2022), who documented human factors barriers in cybersecurity automation and the trust calibration required for operators to accept automated decision-making systems. The staged implementation strategy recommended in this dissertation directly addresses this trust gap by allowing practitioners to observe system performance before progressing to higher levels of automation.

6.3 Theoretical Contributions and Implications

This dissertation makes several important contributions that affect the approach to studies of machine learning models and their application to vulnerability prediction as well as system protection. Here, the contributions are discussed.

6.3.1 Resolving the Temporal Impossibility

The first contribution is identifying the temporal impossibility as a persistent issue within critical infrastructure cybersecurity and demonstrating why traditional approaches fail to address it while the GUARDIAN framework succeeds. The temporal impossibility, as previously highlighted, exists because, while agencies governing critical infrastructure systems such as NERC and TSA have timelines for coordination, patch validation, and implementation (often 15 to 35 days), attackers start exploiting a vulnerability within hours of being published (Garrity, 2025; Charrier & Weiner, 2024). This results in a scenario where the organizations either implement a speedy solution to ensure that the attackers do not exploit those vulnerabilities, or shut down the systems to avoid potential attacks as they seek a solution and develop a patch, all of which present potential risks since there is no guarantee the solution implemented is reliable, or the users end up lacking necessary services as the systems are shut down to resolve the challenges. Shan and Myeong (2024) underscore the need for a proactive approach to protecting these systems, which has been heavily emphasized, leading to the exploration of machine learning adoption in vulnerability management.

Manual coordination has not been able to address this gap, and this is the challenge that this research sought to address by examining the mathematical aspect of the crisis response strategy compared to documented exploitation patterns. The first aspect that stood out was the limitations of manual systems. According to several research studies, the duration between the detection of an alert, the configuration of the assets across the management databases, and the

stakeholder coordination across several organizational levels is 60 minutes for each vulnerability (Dissanayake et al., 2022). Based on how many systems an organization has, addressing vulnerabilities quickly can become a significant challenge, as only 50 systems would require about 50 hours to address, and if the facility is a regional utility with about 150 systems, the whole process would take around 150 hours to complete.

Despite these timelines, however, bodies such as CISA (2021) have also documented that exploitation of a vulnerability after it has been published often begins within the first 24 hours, and sometimes even in minutes of being published. This means that by the time a decision has been made by the stakeholders on the way forward after following the hierarchy required, and the feedback has been provided, a large fraction of attackers will have potentially launched an attack on the system. Reports show that the first 24 hours saw a high number of exploitation attempts. According to VulnCheck, for instance, 28% of over 768 confirmed CVEs were exploited within 24 hours of being published (Garrity, 2025), while Qualys notes that of the 206 high-severity vulnerabilities it had examined, 23% were exploited on the day of publication (Qualys, 2024a). CISA further emphasizes this, noting that forty-two percent of exploits occur within the first 24 hours of disclosure, and fully half materialize within 48 hours (CISA, 2021). Figure 21 quantifies this structural mismatch, showing the 26+ hour gap that no process optimization can eliminate.

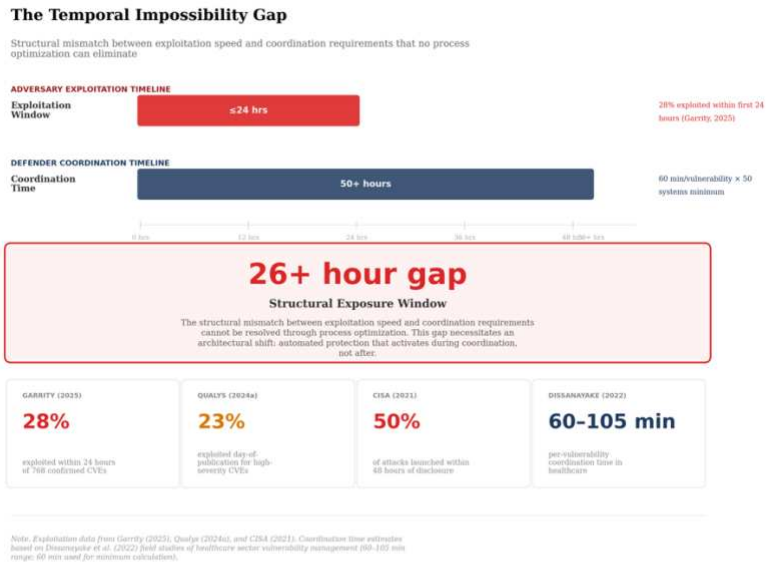

Figure 21. The Temporal Impossibility Gap

The result here is that the coordination time of 50 or more hours, and the attack period of 24 hours or less, leads to a case where organizations have no control other than to accept the substantial exploitation as they work towards assessing the situation and identifying a solution, or finding a graduated strategy that can offer a protective response before assessment is complete. This was the foundation of the GUARDIAN framework, which, instead of implementing the better processes and faster patching recommendations present in traditional theories and frameworks, introduces the concept of an uncertainty-parameterized automation framework to address the challenges. This framework focuses on implementing a protective response while assessment and coordination are being performed, instead of waiting for assessment and coordination as current industrial models require. Due to the implementation of a graduated response based on the nature of the uncertainty level, this framework eradicates the

exploitation window that attackers have in current systems for zero-day attacks. Instead, the system is in protection mode based on the nature and severity of the vulnerability, as the human coordination practices are ongoing, and then a decision is made. The approach proposed and validated by the GUARDIAN framework in this research ensures that where a response would otherwise take 50 or more hours, there is an automated protective response instead, without sacrificing any operational capability, as the decision on assessment and way forward is reviewed and then implemented. This introduces the much-needed proactive approach to these system vulnerabilities, where instead of security analysts serving as gatekeepers of the response action and causing bottlenecks which could extend the duration between an attack and when decisions are made, they are instead reviewers of protective action.

6.3.2 The Recognition and Use of Uncertainty as a Security Primitive

This study has also shifted the perception of uncertainty from a deficiency to a security primitive. In traditional systems, uncertainty is a limitation that should be eliminated by the use of human routing, improved models, or delayed responses (Lakshminarayanan et al., 2017). However, in the development of the GUARDIAN framework, this study establishes that instead of considering uncertainty a limitation, it should be perceived as a tool that, when well implemented, can be central to facilitating a parameterized automated response addressing a persistent security challenge. Instead of relying solely on 'identify,' 'protect,' and 'detect,' the GUARDIAN framework argues that the level of risk in a system can be determined by the degree of belief, which is determined by the JSD distribution, and based on the relationship between uncertainty and response parameters for protection (exploitation rate and time), the statistical correlations provide the operational foundations. As the uncertainty increases, exploitation risk increases, but the duration of response decreases, illustrating the need for

having a shorter duration for higher uncertainty cases to reduce the exploitation risk while implementing more stringent responses for protection. On the other hand, those with lower exploitation risks have a longer response duration, but with less stringent protection strategies, such as 'monitoring.' Treating uncertainty as a determiner of system security level instead of as a limitation that needs addressing makes this approach possible, and by extension, the GUARDIAN framework.

6.3.3 The Implementation of Self-Correction Based on Evidence

This study has also addressed another aspect of current automated security systems, where the responses are often 'blocked' or 'allowed.' In such automated systems, the implemented response persists until an analyst (human) reviews the decision and changes it. If human intervention is absent, then the selected choice persists. However, the GUARDIAN framework, by introducing the multi-signal integration strategy, presents a self-correction strategy where this automated graduated response to system vulnerabilities has a rollback capability based on preprogrammed settings. This contribution was significant for the framework because it demonstrates how analysts serve a monitoring, review, and oversight position, instead of being the direct determinants of system operations. The GUARDIAN framework will automatically correct itself through evidence: it reviews the number of alerts and ensures there are no threats detected on the SIEM monitoring systems or by IDS, verifies that no threat intelligence indicates active targeting, confirms that these signals represent a declining threat level and sufficient time has elapsed, and then verifies the existence of patches before deciding to roll back. When experimented on the 1000 vulnerability scenarios, the premature exploitation rate of this self-correction mechanism was only 1.6%, which is a significant achievement for the framework. This potential of the framework provides a long-awaited remedy for removing

bottlenecks in the system that would be brought about by human coordination, while maintaining audit trails that can be consulted and tracked when the need arises to ascertain whether the decisions made by the system are consistent with the organizational preferences. Ultimately, the implementation of a self-correction mechanism presents an evidence-based solution that organizations can explore in implementing much-needed crisis responses while they conduct assessments, especially in critical infrastructure environments. Figure 22 synthesizes these three contributions, and the paradigm shifts they represent.

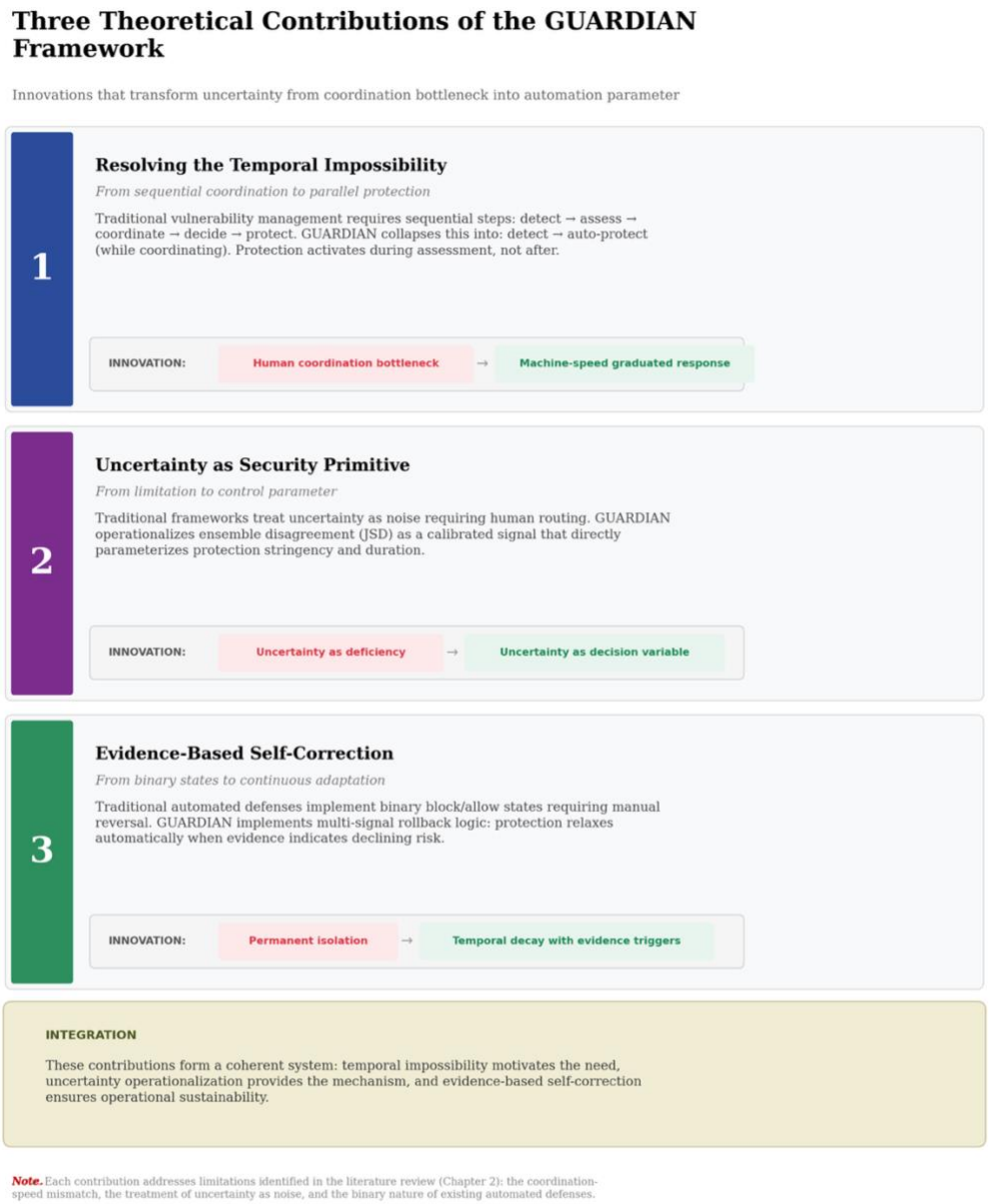

Figure 22. Three Theoretical Contributions of the GUARDIAN Framework

6.4 Practical Deployment Methodology

Throughout the process of describing the research design, empirical results, and proposed system architecture for the GUARDIAN framework, two aspects continuously arose regarding the deployment methodology: first was the need for a staged approach to implementation, and second was the need for parameter customization before implementation. This section emphasizes those aspects.

6.4.1 Staged Implementation Strategy

Organizations need to implement new systems using systematic strategies that allow for risk management while the users build confidence in the system and review its performance in practice (Venkatesh et al., 2003). In line with that, the recommended implementation for the GUARDIAN framework is a phased implementation strategy that begins with monitoring, then enforcement with approval, then graduated automation, and then full automation. This approach provides a step-by-step implementation of the framework, where in each phase there is enough time for review and insights, providing users with an opportunity to build trust in the system (Buttrick, 2009; Venkatesh et al., 2003). The first phase is the monitoring-only phase, which should last 3 to 6 months, where the system is deployed, and it connects to the network, generating assessments and policies, but the enforcement component is switched off. This stage will aim at monitoring the recommended responses and reviewing the logs to determine whether the decisions make sense in relation to the vulnerabilities under consideration. Progression to the next phase depends on the findings of the analysts.

The second stage, after the monitoring-only stage, is enforcement with human approval, and this should span a 3 to 6 month period where the system is permitted to carry out enforcement; however, analysts must approve each action before the deployment of those

protective measures. This step is aimed at earning the analysts' trust in the system while also proving that it will not disrupt the operations of the organization.

Graduated automation would be the third phase, spanning 6 to 12 months, when automation would be applied selectively to low-risk assets (e.g., workstations) while high-risk assets continue to require manual approval. This further provides the analysts with an opportunity to experience the impact of the GUARDIAN framework in the system while also underscoring its ability to address the vulnerabilities and protect the system, without creating new challenges for them. The 6 to 12 month duration provides sufficient observation time to cement the trust in the model, as well as to see how the model would operate over longer durations. The fourth and last phase is full automation, where the system is now fully set up to run autonomously but with human oversight. This is a mature operational state of the system, but the analysts and experts in the security department still have the ability to override the decisions in the environment if the need arises.

6.4.2 Parameter Customization

One of the main points that should be highlighted is that the parameters across organizations vary depending on the governing bodies and the operations of the company. The parameters applied to the creation of the GUARDIAN framework are industry-based, whereby the baseline values that guide the framework are based on openly available data. But certain organizations like the nuclear plants must have more conservative and strict parameters than others because of the level of isolation or protection that the organization prefers to employ when dealing with any sort of attacks. Where operational requirements differ, as they do in these situations, analysts within the organization must modify the framework parameters accordingly, such as adjusting the JSD distributions that define the severity bands and the isolation levels.

Additionally, regarding the protection durations, some organizations might have shorter patching cycles compared to others, which means that the protection window of 7 days to 30 days may be unnecessary given their faster response capabilities. Similarly, in those cases, the organizations have to adjust the protection durations and isolation timelines to mirror their specific operational needs. The GUARDIAN framework provides a working solution that has been empirically validated; however, parameters must be configured by each organization to fit their specific context, as is the case with any other software. Otherwise, this framework will not adequately address the specific organizational needs.

6.5 Study Limitations

Although this was a detailed study, there were a few areas that experienced limitations. These limitations are recognized in this section, which provides context for future testing and implementation of the GUARDIAN framework improvements and enhancements.

6.5.1 Aspects Requiring Future Operational Confirmation

The first limitation is that a number of elements of the GUARDIAN framework will be in need of constant revision and validation of operations in the future. Although this research was extensive and employed empirical analysis to confirm the majority of the findings and results that were introduced, certain areas are either changeable or will certainly require revision and research before they can be applied to various situations. The training, testing, and empirical validation of various components based on the dataset of 279,056 documented vulnerabilities included the performance patterns of individual models (with accuracy ranging from 47.5% to 69.3%), the distribution of ensemble disagreement and the quality of uncertainty calibration, and the monotonic relationship between disagreement and accuracy. Other design specifications were

based on industrial research, namely the rollback timing thresholds, multi-signal integration design specifications, and parameter mapping functions that convert uncertainty to isolation levels and durations. However, other factors such as cross-sector generalizability and long-term sustainability of the GUARDIAN framework remain dynamic and require review checkpoints, especially given that attackers are ever-changing and getting more sophisticated. This is a limitation because it means that the findings of the current research might need revision as the level of threat sophistication rises, or that findings in other critical infrastructure environments might differ from those described in this study.

6.5.2 Use of Simulated Environments Versus Reality

A major part of the research and subsequent findings is premised on a simulated environment. Simulations, although useful in learning, testing, and experimentation, cannot completely replicate production conditions as human factors alone introduce significant variability (Pollini et al., 2022). The field validation strategy was used to partially address this limitation. Nevertheless, that approach also has its own limitation since the system was deployed in advisory mode using a prototype, leaving practitioners without the opportunity to experience the actual impact of automated enforcement. This could shift their perspective and acceptance rate in either direction. This leaves a gap where the evidence on human perspectives of the methodology is largely complementary, and the need for primary evidence from production deployment remains. Nonetheless, as this is a framework suggested for critical infrastructure environments, the appropriate approach is to apply the staged implementation strategy that has already been recommended.

6.5.3 The Trust Gap

Although the field validation study acknowledges that more research might be needed to make the implementation of those findings sustainable, it nonetheless provided themes that reflect the views of stakeholders in the regional healthcare facility. A notable aspect is the trust gap, where 80 percent of participants were willing to recommend the GUARDIAN framework for implementation based on the advisory mode experience, yet indicated that they did not fully trust it to perform the protection response in an automated state without human oversight. This limitation persists since trust calibration requires operators to observe the accuracy of decisions the system makes across multiple vulnerability events, building confidence through demonstrated reliability rather than theoretical assurance. The trust gap observed in this study suggests that the path toward full automation (stage 4) will require extended periods in earlier stages where practitioners can verify framework performance against their own operational judgment.

6.5.4 Limits in the Study's Generalizability

Finally, all findings in the current research are primarily based on the quality of data published by the NVD on CVEs. Although the NVD dataset is extensive, it has known limitations that can affect the generalizability of the current study. These include reporting delays between vulnerability discovery and database publication, inconsistent descriptions across different vendors and vulnerability types, potential underrepresentation of vulnerabilities in proprietary or less-monitored systems, and the absence of organizational context that this dissertation identifies as a fundamental constraint. If reporting bias or systematic gaps exist in the NVD data, such limitations would propagate to the trained models and uncertainty quantification mechanisms built upon them. Future research should examine whether the

accuracy ceiling and disagreement patterns observed in this study hold across alternative vulnerability databases or proprietary organizational datasets.

6.6 Future Research Directions

Further researchers ought to pay attention to field trials that are not restricted to validation simulations and controlled experiments. Despite the technical viability and efficiency of the framework tested under simulated OT conditions, longitudinal field experiments in critical infrastructure sectors of energy, water, transport and healthcare would give more holistic results on operational resilience, operator confidence, and unintended effects. These experiments would also allow exploration of how GUARDIAN operates in the presence of noisy, incomplete or degraded telemetry, or in cases where a human operator is in conflict with automated advice in real-time. The uncertainty thresholds, rollback logic and the protection tier mappings could also be optimized using empirical deployment research using the real adversarial behavior and less simulated attack models.

Uncertainty modeling methods should also be improved and diversified in the future research. The existing structure is based upon ensemble disagreement and Jensen-Shannon divergence in the quantification of epistemic uncertainty, which is appropriate to suit heterogeneous signals and not complete. Other uncertainty estimation techniques such as Bayesian neural networks, conformal prediction, or probabilistic graphical models might be compared in future studies to determine whether those techniques are associated with improved calibration or interpretability in OT settings. Secondly, researchers may also think of adaptive solutions, which vary the weighting of signals with time based upon their past reliability, environmental conditions, or separate asset importance. This would enhance the theoretical

framework of uncertainty-based defense and enhance capability to deal with concept drift, and alteration in attacker behavior.

Human-automation interaction and organizational behavior is another field of research that has not been researched. Although the dissertation recognizes practitioner acceptance of staged automation, future studies should systematically analyze practitioner perceptions, trust, and learning of automated decision making based on uncertainty on such decisions over time. This can be achieved using mixed-method studies such as usability tests, ethnographic research, and survey instruments to demonstrate how GUARDIAN fits into the current security processes, accountability framework, and team coordination mechanisms. Specifically, the impact of automated isolation decisions on safety culture in high-reliability organizations and the long-term consequences of using automation in terms of skills retention as opposed to enhancing strategic orientation among security analysts and engineers should be studied.

Economic analysis must be also expanded to reflect larger system-wide value than coordination-cost savings. This includes the modeling of prevented downtime, reduction of incident escalation, impact on insurance, and reputational risk aversion based on different threat conditions. Special app-specific economic models need to be developed to represent the much different cost structure of utilities, manufacturing plants, and operators of public infrastructure. It might be possible to simplify the complexity of the societal value of uncertainty-conscious automated defense, relative to cyber events, by integrating macroeconomic risk modeling with a micro-level decision logic of GUARDIAN.

Further works should also look at the regulatory, legal and ethical impacts of uncertainty-based automated protection. Though GUARDIAN is intended to be auditable and reversible, issues of liability division, permissible amounts of automation, and the regulatory compliance

procedure of machine-initiating network controls have not been resolved yet. The process of comparing policy across jurisdictions would help to determine the determinants of successful adoption which would guide the development of standards and improve certification programs of automated cyber defense in critical infrastructure. These governance challenges should be resolved so that the technical solutions can be converted to mass implementable and renewable security solutions.

6.7 Conclusion

Overall, this dissertation has illustrated in detail that vulnerability assessment uncertainty can help in the creation of an actionable adaptive-isolation framework (GUARDIAN framework) to tackle the challenges that the critical infrastructure systems face in realizing cybersecurity infrastructural security. This is the existence of a temporal impossibility in which the mechanisms established to address any vulnerability in the system come into conflict with the actions of modern-day cybersecurity attackers who are eager to use the vulnerabilities as soon as they are posted. This time-related impossibility is not solvable by harder work and more manpower, as the constraints are based on the natural necessity of the coordinated processes of responding by the multiple parties and sequential validation. An architectural shift is the most effective way of dealing with this challenge.

The dissertation came up with a framework GUARDIAN, trained six machine learning models; Random Forest, Support Vector Machine, Logistic Regression, Neural Network, LightGBM, and XGBoost, and proceeded with the process of validating and testing them against a dataset of 279,056 vulnerabilities retrieved by NVD. Having proved that the six models had accuracy levels between 47.5% to 69.3% with convergence in this range happening because of the information constraint, the study determined that irreducible uncertainty of the ensemble was

one of the inherent attributes that might be utilized in the management of vulnerability. Ensemble disagreement was transformed into a calibrated sensor and thus employed to introduce graded uncertainty-conscious isolation, to establish a protection system which responds to vulnerabilities rapidly and automatically rather than human coordination and approval. This led to the GUARDIAN framework, which has been evaluated in terms of technical and operational viability and this demonstrates that uncertainty is not always a weakness, a shortcoming, or a failure of a security system. It is an alternative that can be a security primitive. The research mentioned above demonstrates that the automated responses can be operational in a critical infrastructure environment, though it does not rule out the necessity of further research.

In addition to the patterns of model performance reported in Chapter 4, the discrepancies within models and across the ensemble offer significant interpretive cues to the source of the accuracy ceiling. Within models, the results departed from architectural expectations to suggest that the data rather than the algorithms are the culprit. Most significantly, SVM, described by its architecture as suitable for high-dimensional text classification, had the lowest balanced accuracy of 47.5%, whereas the architecturally simpler logistic regression (a linear classifier) outperformed it by 6.7 percentage points. This suggests that the performance of complex models is not superior when the signal is not present in the data. In the ensemble, the Friedman test confirmed that the differences in performance are greater than would be expected by chance, $\chi^2(5) = 12{,}847.3$ and $p < 0.001$. The 21.8 percentage point range between the worst-performing SVM (47.5%) and the best-performing LightGBM (69.3%) was accompanied by a statistically insignificant 2.2 percentage point difference between the two best-performing models (Cliff's $\delta = 0.048$), demonstrating convergence at a ceiling the ensemble could not surpass. The ensemble accuracy of 70.8% was only marginally higher than the best individual model, and confirms the

models make correlated errors on the same structurally ambiguous CVEs. These intra- and inter-model discrepancies are not the product of design flaws in the individual algorithms; they are empirical confirmation that the ceiling is a function of the CVE description data, and the disagreement in the ensemble that produces the JSD signal is its byproduct.

The 47.5% to 69.3% range of accuracy observed in this study agrees with the most rigorously benchmarked external estimates, and the gap between this range and other studies that report higher accuracy is in the feature regime, not model performance. Risse and Böhme (2024) showed that state-of-the-art models perform significantly worse on real-world data and attributed their previously reported success to overfitting on features that are not related to vulnerability content. Croft et al. (2023) also confirmed that the application of deep learning models to NVD data does not improve accuracy over simpler models when data quality is accounted for, attributing accuracy gains to temporal data leakage. Studies reporting accuracy greater than 70% invariably include proprietary organizational telemetry, database metadata, or EPSS scores that are not available to most critical infrastructure deployments, and as a consequence most likely represent overtraining. As a result, the difference between the GUARDIAN framework and other models is not a performance failure. It simply reflects the best possible performance under the information constraints that apply to most critical infrastructure vulnerability management systems, and the GUARDIAN framework is engineered to operate within that constraint by translating residual uncertainty into a graduated protective response.

To understand which signals drive threat detection in the GUARDIAN framework, it is necessary to consider two distinct mechanisms: the feature-driven predictions that each model makes, based on its CVSS metric decompositions and TF-IDF text features, and the model disagreement that defines the graduated protective response. The vulnerability category analysis

directly confirms which features drive classification accuracy. SQL injection vulnerabilities, described using distinctive keywords whose semantics are predictive of high-severity classifications, achieved the highest balanced accuracy of 71.4%. Information exposure vulnerabilities, whose descriptions contain lexically vague language, had the lowest accuracy. This confirms that classification accuracy correlates with the lexical distinctiveness of the description, rather than model complexity. The ensemble disagreement signal works differently. The six architecturally distinct models, when assessing the same CVE, produce different probability distributions, and disagreement among them is the point at which feature signals do not constrain their predictions. The Spearman rank correlation of $\rho = -0.92$ ($p < 0.001$) for JSD and prediction accuracy indicates that the disagreement signal identifies this point. JSD identifies those CVEs whose textual and structural features fail to resolve into stable severity classifications across the ensemble, and that failure to resolve is the operational criterion for graduated protection to be applied before the human assessment is finished. The GUARDIAN framework is therefore explainable in terms of the structure of the disagreement signal, where the same feature uncertainty that prevents prediction is the cue for protection.
GUARDIAN framework provides machine speed protection, which closes the exploitation window that attackers use during human coordination delays.

REFERENCES

Agresti, A., & Kateri, M. (2025). Categorical data analysis. In International Encyclopedia of Statistical Science (pp. 408-411). Springer Berlin Heidelberg.

Ahmad, M., & Wilkins, S. (2025). Purposive sampling in qualitative research: A framework for the entire journey. Quality & Quantity, 59(2), 1461-1479.

Ahmetoglu, H., & Das, R. (2022). A comprehensive review on detection of DDoS attacks using ML techniques. Journal of Computer Networks and Communications, 2022, Article 7620125.

Akano, T. T., & James, C. C. (2022). An assessment of ensemble learning approaches and single-based machine learning algorithms for the characterization of undersaturated oil viscosity. Beni-Suef University Journal of Basic and Applied Sciences, 11(1), 149.

AlAhmadi, B., Martinovic, I., & Axon, L. (2022). 99% false positives: A qualitative study of SOC analysts' perspectives on security alarms. 31st USENIX Security Symposium (USENIX Security 22), 2783–2800.

Alcaraz, C., & Zeadally, S. (2015). Critical infrastructure protection: Requirements and challenges for the 21st century. International Journal of Critical Infrastructure Protection, 8, 53-66.

Alexander, O., Belisle, M., & Steele, J. (2020). MITRE ATT&CK for industrial control systems: Design and philosophy. MITRE Corporation.

Allodi, L. (2015a). Risk-based vulnerability management: Exploiting the economic nature of the attacker to build sound and measurable vulnerability mitigation strategies [Doctoral dissertation, University of Trento].

Allodi, L. (2015b). The heavy tails of vulnerability exploitation. In Proceedings of the 7th International Symposium on Engineering Secure Software and Systems (pp. 1-17). Springer.

Allodi, L. (2017). Economic factors of vulnerability trade and exploitation. In Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security (pp. 1483-1499). ACM.

Allodi, L., & Massacci, F. (2014). Comparing vulnerability severity and exploits using case-control studies. ACM Transactions on Information and System Security, 17(1), 1-20.

Allodi, L., & Massacci, F. (2017). Security events and vulnerability data for cybersecurity risk estimation. Risk Analysis, 37(8), 1606-1627.

Amomo, C. G. (2025). Implementing Zero Trust security models: Challenges, best practices, and future directions in enterprise networks. Iconic Research and Engineering Journals, 8(10), 244–261.

Anwar, A., Abusnaina, A., Chen, S., Li, F., & Mohaisen, D. (2021). Cleaning the NVD: Comprehensive quality assessment, improvements, and analyses. IEEE Transactions on Dependable and Secure Computing, 19(6), 4255-4269.

Apruzzese, G., Colajanni, M., Ferretti, L., & Marchetti, M. (2019). Addressing adversarial attacks against security systems based on machine learning. In 2019 11th International Conference on Cyber Conflict (CyCon) (Vol. 900, pp. 1-18). IEEE.

Barrett, M. P. (2018). Framework for improving critical infrastructure cybersecurity (NIST Cybersecurity Framework v1.1). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.CSWP.04162018

Bassi, D., & Singh, H. (2023). A systematic literature review on software vulnerability prediction models. IEEE Access, 11, 67234-67255.

Basta, N., Ikram, M., Kaafar, M. A., & Walker, A. (2021). Towards a zero-trust micro-segmentation network security strategy: An evaluation framework. arXiv preprint arXiv:2111.10967.

Beerman, J., Berent, D., Falter, Z., & Bhunia, S. (2023). A review of Colonial Pipeline ransomware attack. In 2023 IEEE/ACM 23rd International Symposium on Cluster, Cloud and Internet Computing Workshops (CCGridW) (pp. 8-15). IEEE.

Bilge, L., & Dumitras, T. (2012). Before we knew it: An empirical study of zero-day attacks in the real world. In Proceedings of the 2012 ACM Conference on Computer and Communications Security (pp. 833-844). ACM.

Bisong, E. (2019). Building machine learning and deep learning models on Google cloud platform (pp. 59-64). Apress.

Bochman, A. A., & Freeman, S. (2021). Countering cyber sabotage: Introducing consequence-driven, cyber-informed engineering (CCE). CRC Press.

Boyes, H. (2015). Cybersecurity and cyber-resilient supply chains. Technology Innovation Management Review, 5(4), 28-34.

Bozorgi, M., Saul, L. K., Savage, S., & Voelker, G. M. (2010). Beyond heuristics: Learning to classify vulnerabilities and predict exploits. In Proceedings of the 16th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 105-114). ACM.

Breiman, L. (2001). Random forests. Machine Learning, 45(1), 5–32. https://doi.org/10.1023/A:1010933404324

Bridges, R. A., Jones, C. L., Iannacone, M. D., Testa, K. M., & Goodall, J. R. (2013). Automatic labeling for entity extraction in cyber security. arXiv preprint arXiv:1308.4941.

Broadcom. (2025a). NSX distributed firewall. Broadcom. https://techdocs.broadcom.com/us/en/vmware-cis/nsx/vmware-nsx/9-0/administration-guide.html

Broadcom. (2025b). NSX-T Data Center Rest API guide. Broadcom. https://developer.broadcom.com/xapis/nsx-t-data-center-rest-api/9.0.0/

Brodersen, K. H., Ong, C. S., Stephan, K. E., & Buhmann, J. M. (2010). The balanced accuracy and its posterior distribution. In *2010 20th international conference on pattern recognition* (pp. 3121-3124). IEEE.

Buczak, A. L., & Guven, E. (2015). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153-1176.

Bullough, B. L., Yanchenko, A. K., Smith, C. L., & Zipkin, J. R. (2017). Predicting exploitation of disclosed software vulnerabilities using open-source data. In Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics (pp. 45-53). ACM.

Buttrick, R. (2009). The project workout (4th ed.). Financial Times/Prentice Hall.

Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 1721-1730). ACM.

Case, D. U. (2016). Analysis of the cyber attack on the Ukrainian power grid. Electricity Information Sharing and Analysis Center (E-ISAC), 388(1-29), 3.

Cervini, J., Rubin, A., & Watkins, L. (2022). Don't drink the cyber: Extrapolating the possibilities of Oldsmar's water treatment cyberattack. In International Conference on Cyber Warfare and Security (Vol. 17, No. 1, pp. 19-25). Academic Conferences International Limited.

Chakraborty, S., Krishna, R., Ding, Y., & Ray, B. (2021). Deep learning based vulnerability detection: Are we there yet? IEEE Transactions on Software Engineering, 48(9), 3280-3296.

Chamkar, S. A., Maleh, Y., & Gherabi, N. (2022). The human factor capabilities in security operation center (SOC). Edpacs, 66(1), 1-14.

Chandramouli, R., & Butcher, Z. (2025). Building secure microservices-based applications using service-mesh architecture (NIST SP 800-228). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-228

Charrier, C., & Weiner, S. (2024). How low can you go? An analysis of 2023 time-to-exploit trends. Mandiant. https://cloud.google.com/blog/topics/threat-intelligence/time-to-exploit-trends-2023

Chen, T., & Guestrin, C. (2016). XGBoost: A scalable tree boosting system. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (pp. 785–794). ACM.

Cheng, X., Wang, H., Hua, J., Xu, G., & Sui, Y. (2021). DeepWukong: Statically detecting software vulnerabilities using deep graph neural network. ACM Transactions on Software Engineering and Methodology, 30(3), 1-33.

CISA. (2019). Binding Operational Directive 19-02: Vulnerability remediation requirements for internet-accessible systems. U.S. Department of Homeland Security. https://www.cisa.gov/news-events/directives/bod-19-02-vulnerability-remediation-requirements-internet-accessible-systems

CISA. (2021). *Binding Operational Directive 22-01: Reducing the significant risk of known exploited vulnerabilities.* U.S. Department of Homeland Security. https://www.cisa.gov/news-events/directives/bod-22-01-reducing-significant-risk-known-exploited-vulnerabilities

CISA. (2022). Known exploited vulnerabilities catalog: Analysis and trends. Cybersecurity and Infrastructure Security Agency. https://www.cisa.gov/known-exploited-vulnerabilities-catalog

CISA. (n.d.). Chemical sector profile. Cybersecurity and Infrastructure Security Agency. https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/critical-infrastructure-sectors/chemical-sector

Cisco. (2025). Cisco ACI programmability with object-oriented data model and REST APIs. https://developer.cisco.com/docs/aci/

Cloudflare. (2021). Exploitation of Log4j CVE-2021-44228 before public disclosure and evolution of evasion and exfiltration. Cloudflare Blog. https://blog.cloudflare.com/exploitation-of-cve-2021-44228-before-public-disclosure-and-evolution-of-waf-evasion-patterns/

Coglianese, C., & Nash, J. (2020). Compliance management systems: Do they make a difference? In B. van Rooij & D. D. Sokol (Eds.), The Cambridge Handbook of Compliance (pp. 571-593). Cambridge University Press.

Condon, C., Bowes, R., Galinkin, E., & Caiazza, C. (2023). 2022 vulnerability intelligence report. Rapid7. https://www.rapid7.com/globalassets/_pdfs/2022-vulnerability-intelligence-report.pdf

Corizzo, R., Ceci, M., & Japkowicz, N. (2020). Feature extraction for intrusion detection using word embedding-based neural networks. Intelligent Data Analysis, 24(4), 829-850.

Corona, I., Giacinto, G., & Roli, F. (2013). Adversarial attacks against intrusion detection systems: Taxonomy, solutions and open issues. Information Sciences, 239, 201-225.

Cortes, C., & Vapnik, V. (1995). Support-vector networks. Machine Learning, 20(3), 273-297.

Croft, R., Babar, M. A., & Kholoosi, M. M. (2023). Data quality for software vulnerability datasets. In 2023 IEEE/ACM 45th International Conference on Software Engineering (ICSE) (pp. 121-133). IEEE.

Cyber Safety Review Board. (2022). Review of the December 2021 Log4j event. U.S. Department of Homeland Security.

Davidavičienė, V. (2018). Research methodology: An introduction. In J. Marx Gómez & S. Mouselli (Eds.), Modernizing the academic teaching and research environment. Progress in IS. Springer. https://doi.org/10.1007/978-3-319-74173-4_1

Demšar, J. (2006). Statistical comparisons of classifiers over multiple data sets. Journal of Machine Learning Research, 7, 1–30.

DiMolfetta, D. (2024, May 20). More than 70% of surveyed water systems failed to meet EPA cyber standards. Nextgov/FCW. https://www.nextgov.com/cybersecurity/2024/05/more-70-surveyed-water-systems-failed-meet-epa-cyber-standards/396727/

Dissanayake, N., Jayatilaka, A., Zahedi, M., & Babar, M. A. (2022). Software security patch management: A systematic literature review of challenges, approaches, tools and practices. Information and Software Technology, 144, 106771.

Dragos. (2025). 2025 OT cybersecurity action guide. Dragos, Inc. https://hub.dragos.com/year-in-review-action-guide-2025

Edemekong, P. F., Annamaraju, P., & Haydel, M. J. (2018). Health insurance portability and accountability act. In StatPearls. StatPearls Publishing.

Edgescan. (2024). 2024 vulnerability statistics report. Edgescan. https://www.edgescan.com/wp-content/uploads/2025/04/2024-Vulnerability-Statistics-Report.pdf

Edkrantz, M., & Said, A. (2015). Predicting cyber vulnerability exploits with machine learning. In Proceedings of the 13th Scandinavian Conference on Artificial Intelligence (pp. 48-57). IOS Press.

EPA. (2025). EPA cybersecurity for the water sector. Environmental Protection Agency. https://www.epa.gov/cyberwater/epa-cybersecurity-water-sector

Erdemir, E. (2022). Privacy and security in cyber-physical systems [Doctoral dissertation, Imperial College London].

Falco, G. J., & Rosenbach, E. (2022). Confronting cyber risk: An embedded endurance strategy for cybersecurity. Oxford University Press.

FDA. (2025). Cybersecurity in medical devices: Quality system considerations and content of premarket submissions. U.S. Food and Drug Administration. https://www.fda.gov/regulatory-information/search-fda-guidance-documents/cybersecurity-medical-devices-quality-system-considerations-and-content-premarket-submissions

Fitzgerald, T. (2018). CISO compass: Navigating cybersecurity leadership challenges with insights from pioneers. Auerbach Publications.

Forescout Technologies. (2024). 2023 global threat roundup: Trends in cyberattacks, exploits, and malware. Forescout Research. https://www.forescout.com/blog/2023-threat-roundup/

Frühwirth, C., & Männistö, T. (2009). Improving CVSS-based vulnerability prioritization and response with context information. In Proceedings of the 2009 3rd International Symposium on Empirical Software Engineering and Measurement (pp. 535–544). IEEE. https://doi.org/10.1109/ESEM.2009.5314230

Fu, M., & Tantithamthavorn, C. (2022). Linevul: A transformer-based line-level vulnerability prediction. In Proceedings of the 19th International Conference on Mining Software Repositories (pp. 608-620).

Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of the 33rd International Conference on Machine Learning (pp. 1050-1059). PMLR.

Ganin, A. A., Massaro, E., Gutfraind, A., Steen, N., Keisler, J. M., Kott, A., ... & Linkov, I. (2016). Operational resilience: Concepts, design and analysis. Scientific Reports, 6(1), 1-12.

Garbis, J., & Chapman, J. W. (2021). Zero trust security: An enterprise guide. Apress.

Gardner, M. W., & Dorling, S. R. (1998). Artificial neural networks (the multilayer perceptron)—A review of applications in the atmospheric sciences. Atmospheric Environment, 32(14-15), 2627-2636.

Garrity, P. (2025). VulnCheck exploited vulnerabilities report - May 2024. VulnCheck. https://www.vulncheck.com/blog/kev-report-may-2024

Gawlikowski, J., Tassi, C. R. N., Ali, M., Lee, J., Humt, M., Feng, J., ... & Zhu, X. X. (2023). A survey of uncertainty in deep neural networks. Artificial Intelligence Review, 56(Suppl 1), 1513-1589.

Geng, J., Cai, F., Wang, Y., Koeppl, H., Nakov, P., & Gurevych, I. (2024). A survey of confidence estimation and calibration in large language models. In Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers) (pp. 6577–6595). Association for Computational Linguistics. https://doi.org/10.18653/v1/2024.naacl-long.366

Ginter, A. (2019). SCADA security: What's broken and how to fix it. Apress.

Gonzalez-Granadillo, G., Gonzalez-Zarzosa, S., & Diaz, R. (2021). Security information and event management (SIEM): Analysis, trends, and usage in critical infrastructures. Sensors, 21(14), 4759.

Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. Proceedings of the 34th International Conference on Machine Learning, 1321–1330. https://proceedings.mlr.press/v70/guo17a.html

Halfawy, M. M. R. (2008). Integration of municipal infrastructure asset management processes: Challenges and solutions. Journal of Computing in Civil Engineering, 22(3), 216–229. https://doi.org/10.1061/(ASCE)0887-3801(2008)22:3(216)

Hassan, W. U., Guo, S., Li, D., Chen, Z., Jee, K., Li, Z., & Bates, A. (2019). NoDoze: Combatting threat alert fatigue with automated provenance triage. In Proceedings of the 26th Network and Distributed System Security Symposium. ISOC.

He, H., & Yan, J. (2016). Cyber-physical attacks and defences in the smart grid: A survey. IET Cyber-Physical Systems: Theory & Applications, 1(1), 13–27. https://doi.org/10.1049/iet-cps.2016.0019

Hemme, K. (2015). Critical infrastructure protection: Maintenance is national security. Journal of Strategic Security, 8(3), 25-39.

Hiesgen, R., Nawrocki, M., Schmidt, T. C., & Wählisch, M. (2024). The Log4j incident: A comprehensive measurement study of a critical vulnerability. IEEE Transactions on Network and Service Management, 21(6), 5921–5934.

Hinton, G., Srivastava, N., Krizhevsky, A., Sutskever, I., Rachmad, Y., & Salakhutdinov, R. (2012). Improving neural networks by preventing co-adaptation of feature detectors. arXiv preprint arXiv:1207.0580.

Hollander, M., Wolfe, D. A., & Chicken, E. (2013). Nonparametric statistical methods (3rd ed.). Wiley.

Holm, H., & Afridi, K. K. (2015). An expert-based investigation of the Common Vulnerability Scoring System. Computers & Security, 53, 18–30. https://doi.org/10.1016/j.cose.2015.04.012

Holm, H., Karresand, M., Vidström, A., & Westring, E. (2015). A survey of industrial control system testbeds. In Lecture Notes in Computer Science (Vol. 9417, pp. 11–26). Springer.

Hosmer, D. W., & Lemeshow, S. (1980). Goodness of fit tests for the multiple logistic regression model. Communications in Statistics - Theory and Methods, 9(10), 1043-1069.

Householder, A. D., Wassermann, G., Manion, A., & King, C. (2017). The CERT guide to coordinated vulnerability disclosure (No. CMUSEI2017SR022).

Hu, Y., Yang, A., Li, H., Sun, Y., & Sun, L. (2018). A survey of intrusion detection on industrial control systems. International Journal of Distributed Sensor Networks, 14(8), 1550147718794615.

Humayed, A., Lin, J., Li, F., & Luo, B. (2017). Cyber-physical systems security—A survey. IEEE Internet of Things Journal, 4(6), 1802-1831.

IBM X-Force. (2025). X-Force 2025 threat intelligence index. IBM Security. https://www.ibm.com/reports/threat-intelligence

Iivari, J. (2015). Distinguishing and contrasting two strategies for design science research. European Journal of Information Systems, 24(1), 107–115.

Illumio. (2025a). Illumio Core REST API guide. Illumio, Inc. https://product-docs-repo.illumio.com/Tech-Docs/Core/25.2/REST-APIs/Quick_Reference/index.html#Illumio-Core

Illumio. (2025b). The Illumio Zero Trust segmentation platform. Illumio, Inc. https://www.illumio.com/resource-center/the-illumio-zero-trust-segmentation-platform

Jacobs, J., Romanosky, S., Adjerid, I., & Baker, W. (2021). Improving vulnerability remediation through better exploit prediction. Journal of Cybersecurity, 7(1), tyab019. https://doi.org/10.1093/cybsec/tyab019

Jané, M., Xiao, Q., Yeung, S., Ben-Shachar, M. S., Caldwell, A., Cousineau, D., ... & Feldman, G. (2024). Guide to effect sizes and confidence intervals. https://doi.org/10.17605/OSF.IO/D8C4G

Jordaney, R., Sharad, K., Dash, S. K., Wang, Z., Papini, D., Nouretdinov, I., & Cavallaro, L. (2017). Transcend: Detecting concept drift in malware classification models. In 26th USENIX Security Symposium (USENIX Security 17) (pp. 625-642).

Jurafsky, D., & Martin, J. H. (2013). Speech and language processing: Pearson new international edition. Pearson Higher Ed.

Kavak, H., Padilla, J. J., Lynch, C. J., & Diallo, S. Y. (2018). Big data, agents, and machine learning: Towards a data-driven agent-based modeling approach. In Proceedings of the Annual Simulation Symposium (pp. 1-12).

Ke, G., Meng, Q., Finley, T., Wang, T., Chen, W., Ma, W., Ye, Q., & Liu, T.-Y. (2017). LightGBM: A highly efficient gradient boosting decision tree. Advances in Neural Information Processing Systems, 30, 3146–3154.

Kendall, A., & Gal, Y. (2017). What uncertainties do we need in Bayesian deep learning for computer vision? In Advances in Neural Information Processing Systems 30 (pp. 5574-5584). Curran Associates.

Kindervag, J. (2010). Build security into your network's DNA: The Zero Trust network architecture. Forrester Research.

Kingma, D. P. (2014). Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980.

Knowles, W., Prince, D., Hutchison, D., Disso, J. F. P., & Jones, K. (2015). A survey of cyber security management in industrial control systems. International Journal of Critical Infrastructure Protection, 9, 52-80.

Kruegel, C., & Vigna, G. (2003). Anomaly detection of web-based attacks. In Proceedings of the 10th ACM Conference on Computer and Communications Security (pp. 251-261). ACM.

Kudrati, A., & Pillai, B. A. (2022). Zero Trust journey across the digital estate (1st ed.). CRC Press. https://doi.org/10.1201/9781003225096

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. Annals of Mathematical Statistics, 22(1), 79-86.

Kuncheva, L. I., & Whitaker, C. J. (2003). Measures of diversity in classifier ensembles and their relationship with the ensemble accuracy. Machine Learning, 51(2), 181-207.

Kutscher, J. (2023). M-Trends 2023: Cybersecurity insights from the frontlines. Mandiant. https://cloud.google.com/blog/topics/threat-intelligence/m-trends-2023

Kutscher, J. (2024). M-Trends 2024: Our view from the frontlines. Mandiant. https://cloud.google.com/blog/topics/threat-intelligence/m-trends-2024

Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. Advances in Neural Information Processing Systems, 30.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., ... & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)?—A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence, 296, 103473.

Le, T. H., Chen, H., & Babar, M. A. (2020). Deep learning for source code modeling and generation: Models, applications, and challenges. ACM Computing Surveys (CSUR), 53(3), 1-38.

Le, T. H., Chen, H., & Babar, M. A. (2022). A survey on data-driven software vulnerability assessment and prioritization. ACM Computing Surveys, 55(5), 1-39.

Li, Z., Zou, D., Tang, J., Zhang, Z., Sun, M., & Jin, H. (2019). A comparative study of deep learning-based vulnerability detection system. IEEE Access, 7, 103184-103197.

Lin, J. (1991). Divergence measures based on the Shannon entropy. IEEE Transactions on Information Theory, 37(1), 145-151.

Lin, J., & Mohaisen, D. (2025). From large to mammoth: A comparative evaluation of large language models in vulnerability detection. In Proceedings of the 2025 Network and Distributed System Security Symposium (NDSS).

MacDonald, N., & Firstbrook, P. (2014). Designing an adaptive security architecture for protection from advanced attacks. Gartner.

Madsen, T. (2024). Zero-trust: An introduction. River Publishers.

Maharana, K., Mondal, S., & Nemade, B. (2022). A review: Data pre-processing and data augmentation techniques. Global Transitions Proceedings, 3(1), 91-99.

Makrakis, G. M., Kolias, C., Kambourakis, G., Rieger, C., & Benjamin, J. (2021). Industrial and critical infrastructure security: Technical analysis of real-life security incidents. IEEE Access, 9, 165295-165325.

Maksimovic, J., & Evtimov, J. (2023). Positivism and post-positivism as the basis of quantitative research in pedagogy. Research in Pedagogy, 13(1), 208-218.

Manning, C. D., Raghavan, P., & Schütze, H. (2008). Introduction to information retrieval. Cambridge University Press.

Masys, A. J. (2014). Critical infrastructure and vulnerability: A relational analysis through actor network theory. In Networks and Network Analysis for Defence and Security (pp. 265-280). Springer.

Mauthner, N. S. (2020). Research philosophies and why they matter. In A. P. A. Thomson & J. Harris (Eds.), How to keep your doctorate on track (pp. 18-23). Edward Elgar Publishing.

May, T. (1997). Issues, methods and process. Open University Press.

Mell, P. (2005). The National Vulnerability Database. https://csrc.nist.rip/ispab/2005-12/P_Mell-Dec2005-ISPAB.pdf

MITRE. (2024). Common Weakness Enumeration (CWE). https://cwe.mitre.org/

MITRE. (2025). Common Vulnerabilities and Exposures (CVE) program. MITRE. https://www.cve.org

Morris, T., Srivastava, A., Reaves, B., Gao, W., Pavurapu, K., & Reddi, R. (2011). A control system testbed to validate critical infrastructure protection concepts. International Journal of Critical Infrastructure Protection, 4(2), 88–103. https://doi.org/10.1016/j.ijcip.2011.06.005

Murphy, E., Foster, E., Diaz, K., Rivera, W., & Lopez, A. (2025). Comparative analysis of entropy based divergence metrics for ensemble diversity control.

NERC. (2016). CIP-007-6: Cyber security - Systems security management. North American Electric Reliability Corporation. https://www.nerc.com/globalassets/standards/reliability-standards/cip/cip-007-6.pdf

NERC. (2019). 2019 state of reliability report. North American Electric Reliability Corporation. https://www.nrc.gov/reading-rm/doc-collections/commission/slides/2019/20190925/lauby-20190925.pdf

NERC. (2025). 2025 state of reliability report. North American Electric Reliability Corporation. https://www.nerc.com/globalassets/programs/rapa/pa/nerc_sor_2025_technical_assessment.pdf

Nguyen, T. T., & Reddi, V. J. (2021). Deep reinforcement learning for cyber security. IEEE Transactions on Neural Networks and Learning Systems, 34(8), 3779-3795.

Nielsen, F. (2019). On the Jensen-Shannon symmetrization of distances relying on abstract means. Entropy, 21(5), Article 485.

NIST. (2004). Standards for security categorization of federal information and information systems (FIPS Publication 199). U.S. Department of Commerce.

NIST. (2021). Developing cyber-resilient systems: A systems security engineering approach (NIST SP 800-160 Vol. 2 Rev. 1). https://doi.org/10.6028/NIST.SP.800-160v2r1

NIST. (2024). Cybersecurity framework (CSF) 2.0. https://nvlpubs.nist.gov/nistpubs/CSWP/NIST.CSWP.29.pdf

NIST. (2025). National Vulnerability Database. National Institute of Standards and Technology. https://nvd.nist.gov/

NIST. (n.d.). National Vulnerability Database: Data feeds. National Institute of Standards and Technology. https://nvd.nist.gov/vuln/data-feeds

NRC. (2021). Title 10, Part 73, Section 54: Protection of digital computer and communications systems and networks. https://www.nrc.gov/reading-rm/doc-collections/cfr/part073/part073-0054

Nuclear Sector Coordinating Council. (n.d.). Nuclear sector overview. https://www.cisa.gov/topics/critical-infrastructure-security-and-resilience/critical-infrastructure-sectors/nuclear-reactors-materials-and-waste-sector

Opara-Martins, J., Sahandi, R., & Tian, F. (2016). Critical analysis of vendor lock-in and its impact on cloud computing migration: A business perspective. Journal of Cloud Computing, 5, Article 4. https://doi.org/10.1186/s13677-016-0054-z

Pendlebury, F., Pierazzi, F., Jordaney, R., Kinder, J., & Cavallaro, L. (2019). TESSERACT: Eliminating experimental bias in malware classification across space and time. In Proceedings of the 28th USENIX Security Symposium (pp. 729–746). USENIX Association.

Pierazzi, F., Pendlebury, F., Cortellazzi, J., & Cavallaro, L. (2020). Intriguing properties of adversarial ML attacks in the problem space. In Proceedings of the 2020 IEEE Symposium on Security and Privacy (pp. 1332-1349). IEEE. https://doi.org/10.1109/SP40000.2020.00073

Pisner, D. A., & Schnyer, D. M. (2020). Support vector machine. In A. Mechelli & S. Vieira (Eds.), Machine learning: Methods and applications to brain disorders (pp. 101-121). Academic Press. https://doi.org/10.1016/B978-0-12-815739-8.00006-7

Pollini, A., Callari, T. C., Tedeschi, A., Ruscio, D., Save, L., Chiarugi, F., & Guerri, D. (2022). Leveraging human factors in cybersecurity: An integrated methodological approach. Cognition, Technology & Work, 24, 371–390. https://doi.org/10.1007/s10111-021-00683-y

Popescu, M. C., Balas, V. E., Perescu-Popescu, L., & Mastorakis, N. (2009). Multilayer perceptron and neural networks. WSEAS Transactions on Circuits and Systems, 8(7), 579–588.

Qualys. (2023). 2023 Qualys TruRisk research report. Qualys, Inc. https://cdn2.qualys.com/docs/mktg/qualys-2023-trurisk-threat-research-report.pdf

Qualys. (2024a, August 6). Cybersecurity threat landscape 2024 midyear review. Qualys Blog. https://blog.qualys.com/vulnerabilities-threat-research/2024/08/06/2024-midyear-threat-landscape-review

Qualys. (2024b). Qualys API. https://cdn2.qualys.com/docs/qualys-api-vmpc-user-guide.pdf

Quinn, S., Ivy, N., Chua, J., Barrett, M., Witte, G., Feldman, L., Topper, D., & Gardner, R. (2022, November 17). Using business impact analysis to inform risk prioritization and response. National Institute of Standards and Technology. https://doi.org/10.6028/NIST.IR.8286D-upd1

Redis. (2025). The definitive guide to caching at scale with Redis. https://redis.io/resources/caching-at-scale-with-redis/

Reeves, A., & Ashenden, D. (2023). Understanding decision making in security operations centres: Building the case for cyber deception technology. Frontiers in Psychology, 14, 1165705.

Ribeiro, A. (2022, November 6). Chemical companies set to outline cybersecurity posture, as OT and ICS threats continue to prevail. *Industrial Cyber*. https://industrialcyber.co/features/chemical-companies-set-to-outline-cybersecurity-posture-as-ot-and-ics-threats-continue-to-prevail/

Ribeiro, A. (2024, May 21). EPA alerts drinking water systems on cybersecurity vulnerabilities, increasing enforcement actions. *Industrial Cyber*. https://industrialcyber.co/utilities-energy-power-water-waste/epa-alerts-drinking-water-systems-on-cybersecurity-vulnerabilities-increasing-enforcement-actions/

Ribeiro, A. (2025a, November 4). Black & Veatch's 2025 Electric Report finds utilities prioritizing cybersecurity training over tools to tackle digital grid threats. *Industrial Cyber*. https://industrialcyber.co/reports/black-veatchs-2025-electric-report-finds-utilities-prioritizing-cybersecurity-training-over-tools-to-tackle-digital-grid-threats/

Ribeiro, A. (2025b, April 18). Resecurity warns of increased cyber threats to energy and nuclear facilities from hacktivists and nation-states. *Industrial Cyber*. https://industrialcyber.co/utilities-energy-power-water-waste/resecurity-warns-of-increased-cyber-threats-to-energy-and-nuclear-facilities-from-hacktivists-and-nation-states/

Riggs, H., Tufail, S., Parvez, I., Tariq, M., Khan, M. A., Amir, A., Vuda, K. V., & Sarwat, A. I. (2023). Impact, vulnerabilities, and mitigation strategies for cyber-secure critical infrastructure. Sensors, 23(8), 4060. https://doi.org/10.3390/s23084060

Rinaldi, S. M., Peerenboom, J. P., & Kelly, T. K. (2001). Identifying, understanding, and analyzing critical infrastructure interdependencies. IEEE Control Systems Magazine, 21(6), 11-25.

Risse, N., & Böhme, M. (2024). Uncovering the limits of machine learning for automatic vulnerability detection. In Proceedings of the 33rd USENIX Security Symposium (pp. 4247-4264). USENIX Association.

Rose, S., Borchert, O., Mitchell, S., & Connelly, S. (2020). Zero Trust architecture (NIST Special Publication 800-207). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-207

Rossman, L. A. (2000). EPANET 2 users manual. U.S. Environmental Protection Agency.

Sabottke, C., Suciu, O., & Dumitras, T. (2015). Vulnerability disclosure in the age of social media: Exploiting Twitter for predicting real-world exploits. Proceedings of the 24th USENIX Security Symposium, 1041-1056.

SANS. (2023). 2023 SOC survey. SANS Institute. https://www.sans.org/white-papers/2023-sans-soc-survey

Scalco, A., & Simske, S. (2025). *Systems engineering for critical infrastructure in a cyber world*. Springer.

Schonlau, M., & Zou, R. Y. (2020). The random forest algorithm for statistical learning. The Stata Journal, 20(1), 3-29. https://doi.org/10.1177/1536867X20909688

Shan, A., & Myeong, S. (2024). Proactive threat hunting in critical infrastructure protection through hybrid machine learning algorithm application. Sensors, 24(15), 4888. https://doi.org/10.3390/s24154888

Sharadin, G. (2022, December 6). Log4j: One year later. Imperva. https://www.imperva.com/blog/log4j-one-year-later/

Shiri Harzevili, N., Boaye Belle, A., Wang, J., Wang, S., Jiang, Z. M., & Nagappan, N. (2024). A systematic literature review on automated software vulnerability detection using machine learning. ACM Computing Surveys, 57(3), 1-36.

Shmilovici, A. (2005). Support vector machines. In O. Maimon & L. Rokach (Eds.), Data mining and knowledge discovery handbook. Springer.

SIGA. (2025). Rethinking water industry OT cybersecurity strategy. SIGA OT Solutions. https://sigasec.com/rethinking-water-industry-ot-cybersecurity-strategy/

Silvestri, S., Gargiulo, F., Ciampi, M., & De Pietro, G. (2023). A machine learning approach for the NLP-based analysis of cyber threats and vulnerabilities of the healthcare ecosystem. Sensors, 23(2), Article 651. https://doi.org/10.3390/s23020651

Simske, S. J. (2013). Meta-algorithmics: Patterns for robust, low cost, high quality systems. John Wiley & Sons.

Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In Proceedings of the 2010 IEEE Symposium on Security and Privacy (pp. 305–316). IEEE.

Souppaya, M., & Scarfone, K. (2022). Guide to enterprise patch management planning: Preventive maintenance for technology (NIST Special Publication 800-40 Revision 4). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-40r4

Spring, J. M., Hatleback, E., Householder, A., Manion, A., & Shick, D. (2021). Time to change the CVSS? IEEE Security & Privacy, 19(2), 74–78. https://doi.org/10.1109/MSEC.2020.3044475

Stallings, W. (2019). Effective cybersecurity: A guide to using best practices and standards. Addison-Wesley Professional.

Steenhoek, B., Rahman, M. M., Roy, M. K., Alam, M. S., Tong, H., Das, S., Barr, E. T., & Le, W. (2024). A comprehensive study of the capabilities of large language models for vulnerability detection. arXiv preprint arXiv:2403.17218.

Stouffer, K., Pease, M., Tang, C., Zimmerman, T., Pillitteri, V., Lightman, S., Hahn, A., Saravia, S., Sherule, A., & Thompson, M. (2023). Guide to operational technology (OT) security (NIST SP 800-82r3). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-82r3

Stouffer, K., Pillitteri, V., Lightman, S., Abrams, M., & Hahn, A. (2015). Guide to industrial control systems (ICS) security (NIST SP 800-82r2). National Institute of Standards and Technology. https://doi.org/10.6028/NIST.SP.800-82r2

Sundaramurthy, S. C., Case, J., Truong, T., Zomlot, L., & Hoffmann, M. (2014a). A tale of three security operation centers. In Proceedings of the 2014 ACM Workshop on Security Information Workers (pp. 43-50).

Sundaramurthy, S. C., McHugh, J., Ou, X. S., Rajagopalan, S. R., & Wesch, M. (2014b). An anthropological approach to studying CSIRTs. IEEE Security & Privacy, 12(5), 52-60.

Sundaramurthy, S. C., McHugh, J., Ou, X., Wesch, M., Bardas, A. G., & Rajagopalan, S. R. (2016). Turning contradictions into innovations or: How we learned to stop whining and improve security operations. In Twelfth Symposium on Usable Privacy and Security (SOUPS 2016) (pp. 237-251).

Sundjaja, J. H., Shrestha, R., & Krishan, K. (2023). McNemar and Mann-Whitney U tests. In StatPearls. StatPearls Publishing.

Tang, E. K., Suganthan, P. N., & Yao, X. (2006). An analysis of diversity measures. Machine Learning, 65(1), 247-271. https://doi.org/10.1007/s10994-006-9449-2

Torang, A., Gupta, P., & Klinke, D. J. (2019). An elastic-net logistic regression approach to generate classifiers and gene signatures for types of immune cells and T helper cell subsets. BMC Bioinformatics, 20, Article 433. https://doi.org/10.1186/s12859-019-2994-z

Tounsi, W., & Rais, H. (2018). A survey on technical threat intelligence in the age of sophisticated cyber attacks. Computers & Security, 72, 212-233. https://doi.org/10.1016/j.cose.2017.09.001

US DOJ. (2021). Four Chinese nationals working with the Ministry of State Security charged with global computer intrusion campaign targeting intellectual property and confidential business information, including infectious disease research. U.S. Department of Justice.

Vargas, V. M., Guijo-Rubio, D., Gutiérrez, P. A., & Hervás-Martínez, C. (2021). ReLU-based activations: Analysis and experimental study for deep learning. In E. Alba et al. (Eds.), Advances in Artificial Intelligence. CAEPIA 2021. Lecture Notes in Computer Science (Vol. 12882). Springer. https://doi.org/10.1007/978-3-030-85713-4_4

Vectra AI. (2023). 2023 state of threat detection. Vectra AI. https://www.vectra.ai/resources/2023-state-of-threat-detection

Venable, J., Pries-Heje, J., & Baskerville, R. (2016). FEDS: A framework for evaluation in design science research. European Journal of Information Systems, 25(1), 77–89. https://doi.org/10.1057/ejis.2014.36

Venkatesh, V., Morris, M. G., Davis, G. B., & Davis, F. D. (2003). User acceptance of information technology: Toward a unified view. MIS Quarterly, 27(3), 425–478. https://doi.org/10.2307/30036540

Verbraeken, J., Wolting, M., Katzy, J., Kloppenburg, J., Verbelen, T., & Rellermeyer, J. S. (2020). A survey on distributed machine learning. ACM Computing Surveys (CSUR), 53(2), 1-33.

VulnCheck. (2024). Known exploited vulnerabilities database. VulnCheck, Inc. https://research.vulncheck.com/2024-dashboard/

Wang, C. (2023). Calibration in deep learning: A survey of the state-of-the-art. arXiv preprint arXiv:2308.01222.

Water Sector Coordinating Council. (2021). Cybersecurity: 2021 state of the sector. WaterISAC. https://www.waterisac.org/2021survey

Weintraub, E. (2016). Evaluating damage potential in security risk scoring models. International Journal of Advanced Computer Science and Applications, 7(5), 345–353.

Williams, T. J. (1994). The Purdue Enterprise Reference Architecture. Computers in Industry, 24(2-3), 141–158. https://doi.org/10.1016/0166-3615(94)90017-5

XcelPros. (n.d.). Rising data security risks in chemical plants. Retrieved October 4, 2025, from https://xcelpros.com/rising-data-security-risks-in-chemical-plants/

Xiaojian, Z., Liandong, C., Jie, F., Xiangqun, W., & Qi, W. (2021). Power IoT security protection architecture based on zero trust framework. In 2021 IEEE 5th International Conference on Cryptography, Security and Privacy (CSP) (pp. 166-170). IEEE.

Yusta, J. M., Correa, G. J., & Lacal-Arántegui, R. (2011). Methodologies and applications for critical infrastructure protection: State-of-the-art. Energy Policy, 39(10), 6100–6119.

Zar, J. H. (2010). Biostatistical analysis (5th ed.). Prentice Hall.

Zhang, S., Caragea, D., & Ou, X. (2011). An empirical study on using the National Vulnerability Database to predict software vulnerabilities. In Proceedings of the 22nd International Conference on Database and Expert Systems Applications (pp. 217-231). Springer.

Zhou, Y., Liu, S., Siow, J., Du, X., & Liu, Y. (2019). Devign: Effective vulnerability identification by learning comprehensive program semantics via graph neural networks. In Advances in Neural Information Processing Systems 32 (pp. 10197-10207). Curran Associates.

Zio, E. (2016). Challenges in the vulnerability and risk analysis of critical infrastructures. Reliability Engineering & System Safety, 152, 137-150.

APPENDIX A: FIELD VALIDATION SURVEY INSTRUMENT AND RESULTS

This appendix presents the complete survey instrument administered during the GUARDIAN framework field validation study, followed by response data and statistical analysis from 15 practitioners at a regional healthcare facility.

A.1 Survey Instrument

GUARDIAN Framework Field Validation Instrument

Practitioner Assessment of Uncertainty-Parameterized Adaptive Isolation

Colorado State University Doctoral Research

Section A: Respondent Context

A1. Primary operational role (Select one)

- ☐ Security operations (SOC analyst, security engineer)
- ☐ Network operations (network administrator, infrastructure)
- ☐ Clinical/biomedical device management
- ☐ Security management or compliance
- ☐ Other: _____

A2. Approximate number of CVEs assessed using GUARDIAN during evaluation

- ☐ 1–10
- ☐ 11–25
- ☐ 26–50
- ☐ More than 50

Section B: Uncertainty Quantification

The framework computes Jensen-Shannon Divergence (JSD) from ensemble model disagreement to quantify assessment uncertainty. These items assess whether this uncertainty signal provided actionable information.

B1. The JSD uncertainty score distinguished cases where the automated severity assessment was reliable from cases where it was unreliable.

Strongly Disagree (1) Disagree (2) Neither (3) Agree (4) Strongly Agree (5)

B2. When JSD indicated high uncertainty, I increased scrutiny of the automated recommendation before approving.

Strongly Disagree (1) Disagree (2) Neither (3) Agree (4) Strongly Agree (5)

Section C: Graduated Protection Correspondence

The framework maps uncertainty and severity to four protection levels: Enhanced Monitoring, Restricted Communication, Zero Trust Microsegmentation, and Complete Quarantine.

C1. The four isolation levels represented meaningfully distinct protection postures for our operational environment.

Strongly Disagree (1) Disagree (2) Neither (3) Agree (4) Strongly Agree (5)

C2. The recommended isolation level matched the actual risk the vulnerability posed to affected systems.

Strongly Disagree (1) Disagree (2) Neither (3) Agree (4) Strongly Agree (5)

Section D: Temporal Self-Correction

The framework implements graduated temporal thresholds for protection rollback based on time-without-exploitation, threat intelligence, and patch status.

D1. The recommended protection durations aligned with the time required to validate and deploy patches in our environment.

Strongly Disagree (1) Disagree (2) Neither (3) Agree (4) Strongly Agree (5)

D2. The rollback conditions (time elapsed, no exploitation detected, threat intelligence stable) represented appropriate criteria for reducing protection.

Strongly Disagree (1) Disagree (2) Neither (3) Agree (4) Strongly Agree (5)

Section E: Coordination and Decision Latency

These items assess observable effects on decision latency from automated uncertainty-parameterized assessment.

E1. Using GUARDIAN, I reached an initial protection decision faster than I would have through our standard vulnerability triage process.

Strongly Disagree (1) Disagree (2) Neither (3) Agree (4) Strongly Agree (5)

E2. The framework's recommendations reduced the number of stakeholders I needed to consult before implementing initial protection.

Strongly Disagree (1) Disagree (2) Neither (3) Agree (4) Strongly Agree (5)

Section F: Decision Alignment and Override Patterns

F1. During the evaluation, how frequently did you reject or substantially modify GUARDIAN's recommended isolation level?

- ☐ Never (0% of recommendations)
- ☐ Rarely (1–10% of recommendations)
- ☐ Sometimes (11–25% of recommendations)
- ☐ Frequently (26–50% of recommendations)
- ☐ Usually (>50% of recommendations)

F2. When you modified a recommendation, what was the primary reason? (Select most common)

- ☐ Framework recommended more aggressive isolation than warranted
- ☐ Framework recommended less aggressive isolation than warranted
- ☐ Recommended duration was too short

☐ Recommended duration was too long

☐ Affected device required different handling (e.g., FDA-regulated, critical care)

☐ Organizational policy required different approach

☐ Other: _____

Section G: Open Response

G1. Identify one specific instance where GUARDIAN's uncertainty quantification or graduated protection informed a decision you would have made differently without the framework. (If none, write "None observed")

_____

_____

_____

— END OF INSTRUMENT —

A.2 Sample Demographics

Fifteen practitioners ($N = 15$) participated in the field validation study. Table A.1 presents the sample distribution by department and role.

Table A.1. Sample Distribution by Department

Department	**n (%)**	**Roles Represented**
Infrastructure	6 (40%)	Network Architect, SysAdmin, Database Admin, Virtualization Eng., Backup Admin, Service Desk Mgr
OT/Clinical	4 (27%)	Biomed Lead, PACS Admin, Clinical Analyst, Facilities/IoT
Security Ops	3 (20%)	Sr. Security Engineer, SOC Analyst, Incident Responder
Executive	2 (13%)	CISO, Director IT Operations

A.3 Descriptive Statistics

Table A.2 presents descriptive statistics for all Likert-scale items. Agreement percentage represents respondents rating 4 (Agree) or 5 (Strongly Agree).

Table A.2. Descriptive Statistics for Survey Items (N = 15)

Item	M	SD	n Agree	% Agree	
Q3: Requirements	4.13	0.83	13	86.7%	
Q4: Ease of Use	4.20	0.77	12	80.0%	
Q5: Uncertainty/JSD (B1)	3.87	0.92	10	66.7%	
Q6: Isolation Levels (C2)	3.93	1.03	11	73.3%	
Q7: Speed (E1)	4.20	0.68	13	86.7%	
Q8: Topology	4.27	0.70	13	86.7%	
Q9: Automation Trust	3.60	1.24	9	60.0%	
Q10: Recommend	4.13	0.74	12	80.0%	

Note. Comparative Assessment (Q11): 33.3% rated GUARDIAN Much Better, 46.7% Somewhat Better, 13.3% About the Same, 6.7% Somewhat Worse. Overall, 80.0% rated GUARDIAN better than current tools.

A.4 Trust Gap Analysis

A central finding was the trust gap between recommendation intent (Q10: $M = 4.13$) and automation trust (Q9: $M = 3.60$), yielding an overall gap of +0.53. Table A.3 presents department-level analysis.

Table A.3. Trust Gap by Department

Department	n	Q10 Recommend	Q9 Trust	Gap
Executive	2	4.50	3.50	+1.00
Infrastructure	6	3.67	3.33	+0.34
Security Ops	3	5.00	4.67	+0.33
OT/Clinical	4	4.00	3.25	+0.75

Security Operations demonstrated highest automation acceptance (Q9: $M = 4.67$), while OT/Clinical showed lowest ($M = 3.25$). Executive leadership exhibited the largest gap (+1.00), endorsing the framework while preferring human approval.

A.5 Qualitative Themes

Six themes emerged from open-ended responses. Table A.4 presents representative quotations.

Table A.4. Qualitative Themes from Open-Ended Responses

Theme	Representative Quotation
Compliance Value	"Auditors always ask why we haven't patched immediately; now I can show them the High Uncertainty score to justify waiting for a maintenance window." (CISO)
Continuity Concerns	"If GUARDIAN triggers Level 2 isolation, I need to know if it will interrupt the Cerner PharmNet transaction logs." (Director IT)
Automation Anxiety	"I cannot have GUARDIAN pushing ACLs to the Core automatically. It needs to open a ServiceNow ticket instead." (Network Architect)
Clinical Safety	"You cannot run automated isolation on IMG-MRI-01. If GUARDIAN isolates during a patient scan, we have a safety incident." (Biomed Lead)
Endpoint Value	"I love Level 3 isolation for endpoints. If PHARM-PC-04 gets infected, I can isolate it with one click." (Incident Responder)
Asset Visibility	"Most tools fail to identify Zebra printers correctly. GUARDIAN fingerprinted them as Android 13." (Facilities/IoT)

A.6 Complete Response Matrix

Table A.5 presents the complete anonymized response data for all participants.

Table A.5. Anonymized Response Data

ID	Dept	Q3	Q4	Q5	Q6	Q7	Q8	Q9	Q10	Q11
1	Executive	5	5	5	4	5	5	4	5	Much Better
2	Executive	4	4	4	4	5	4	3	4	Somewhat Better

3	Infra.	4	3	4	2	4	5	2	3	About Same
4	Infra.	4	4	3	3	4	4	3	4	Somewhat Better
5	Infra.	3	3	3	3	3	4	2	3	About Same
6	Infra.	4	5	4	5	4	4	4	4	Somewhat Better
7	Infra.	4	4	4	4	4	4	4	4	Somewhat Better
8	Infra.	4	5	3	4	4	3	5	4	Somewhat Better
9	Sec. Ops	5	4	5	5	5	5	4	5	Much Better
10	Sec. Ops	5	5	5	5	5	5	5	5	Much Better
11	Sec. Ops	5	5	5	5	5	5	5	5	Much Better
12	OT/Clin.	2	3	4	2	3	5	1	3	Somewhat Worse
13	OT/Clin.	4	4	4	4	4	4	3	4	Somewhat Better
14	OT/Clin.	5	5	2	5	4	3	5	5	Much Better
15	OT/Clin.	4	4	3	4	4	4	4	4	Somewhat Better

Note. Q3–Q10 scale: 1 = Strongly Disagree to 5 = Strongly Agree. Q11: Comparative assessment versus current organizational vulnerability management tools.