Journal of Computer Science and Technology Studies
ISSN: 2709-104X
DOI: 10.32996/jcsts
Journal Homepage: www.al-kindipublisher.com/index.php/jcsts

JCSTS
AL-KINDI CENTER FOR RESEARCH AND DEVELOPMENT

| RESEARCH ARTICLE

A National-Scale AI-Driven Cyber Defense Framework for Protecting U.S. Critical Infrastructure Against Nation-State Attacks

Md Humayun Kabir[1], Md Al Mamun Siddike[2], MD RAZIB[3] and Md Riyad Uddin[4]
[1] Westcliff University, Master of science in information technology
[2] MS IN BUSINESS ANALYTICS, Trine University
[34] Westcliff University
Corresponding Author: Md Humayun Kabir, **E-mail**: Humayun9152@gmail.com

| ABSTRACT
The Cybersecurity and Infrastructure Security Agency (CISA) has predicted a 140% rise in high-impact attacks between 2022 and 2024, signalling an increase in nation-state adversary cyberattacks that affect US critical infrastructures. The sophistication, persistence and ever-evolving nature of nation-states' hacking has outgrown current cybersecurity, which primarily focuses on reactive and perimeter-based approaches. In this paper, we present and test a National-Scale AI-Driven Cyber Defense Framework (NAICDF) a multi-layer, intelligence-based framework with machine learning-based threat detection, federated learning for data sharing across critical infrastructure sectors, Zero Trust Architecture (ZTA) and automated response systems. We compare intrusion detection accuracy, mean time to respond (MTTR) and resilience with traditional approaches using incident reports of 847 confirmed nation-state intrusions in 11 critical infrastructure sectors (2020-2024), publicly available threat intelligence and simulation data from the National Cyber Exercise Program (NCEP) of the Cybersecurity and Infrastructure Security Agency (CISA). Results demonstrate the NAICDF achieves a 94.3% accuracy in detecting intrusions (2.1% false positives) and a 67% reduction in the mean time to respond (MTTR) when compared to conventional security operations center (SOC) systems. We also explore governance and public-private partnership models, and compatibility with existing policy frameworks such as CIRCIA (2022) and the National Cybersecurity Strategy (2023). The framework offers a policy-friendly, scalable model to secure our critical systems in the 21st century.

| KEYWORDS
Critical infrastructure protection, AI-driven cybersecurity, nation-state attacks, Zero Trust Architecture, federated learning, intrusion detection, national cybersecurity policy, advanced persistent threats

| ARTICLE INFORMATION
ACCEPTED: 15 April 2026 **PUBLISHED:** 02 May 2026 **DOI:** 10.32996/jcsts.2026.8.6.7

1. Introduction

As society has become increasingly digital, the U.S. critical infrastructure such as energy, water, financial, transportation and health care systems are also more efficient and more vulnerable. The Department of Homeland Security (DHS) describes these 16 critical infrastructure sectors as the "backbone of our nation's security, economy, and public health". But they are also coming under a threat not from physical actors, but from sophisticated nation-state cyber operations. We now know it's here. Microsoft's 2024 Digital Defense Report reveals that in 2024 critical infrastructure received more than 600 million attacks a day from cybercriminals and nation states. The 2024 Annual Threat Assessment from the Department of National Intelligence highlights Russia, China, Iran and North Korea as the top an. d most active state-sponsored cyber programs, all of which have advanced persistent threat (APT) groups with the resources and time to infiltrate and remain in critical systems to disrupt their operations. By comparison, cyber security approaches for most critical infrastructure are ad hoc, reactive and focus on threats from the past. Legacy technologies such as firewalls, IDS, and siloed incident response plans are inferior to targets using combinations of zero-days, living-off-the-land techniques, supply chain attacks, and artificial intelligence to identify vulnerabilities. The 2023 National

Cybersecurity Strategy acknowledged this, stating that we should "shift the burden of cybersecurity away from individuals and small organisations" and towards a national approach, but there is no documented, technically viable solution for doing this at a national scale. This paper proposes and evaluates the success of the National AI-Driven Cyber Defense Framework (NAICDF), which leverages artificial intelligence and machine learning technologies across the entire stack from real-time anomaly detection in operational technology (OT) environments, to privacy-enhancing multi-sector threat intelligence using federated learning, to an autonomous response orchestrator using explainable AI. Our approach is inspired by a study of five years' worth of nation-state attacks, has been simulated against CISA exercises and complies with U.S. regulatory and policy guidelines.

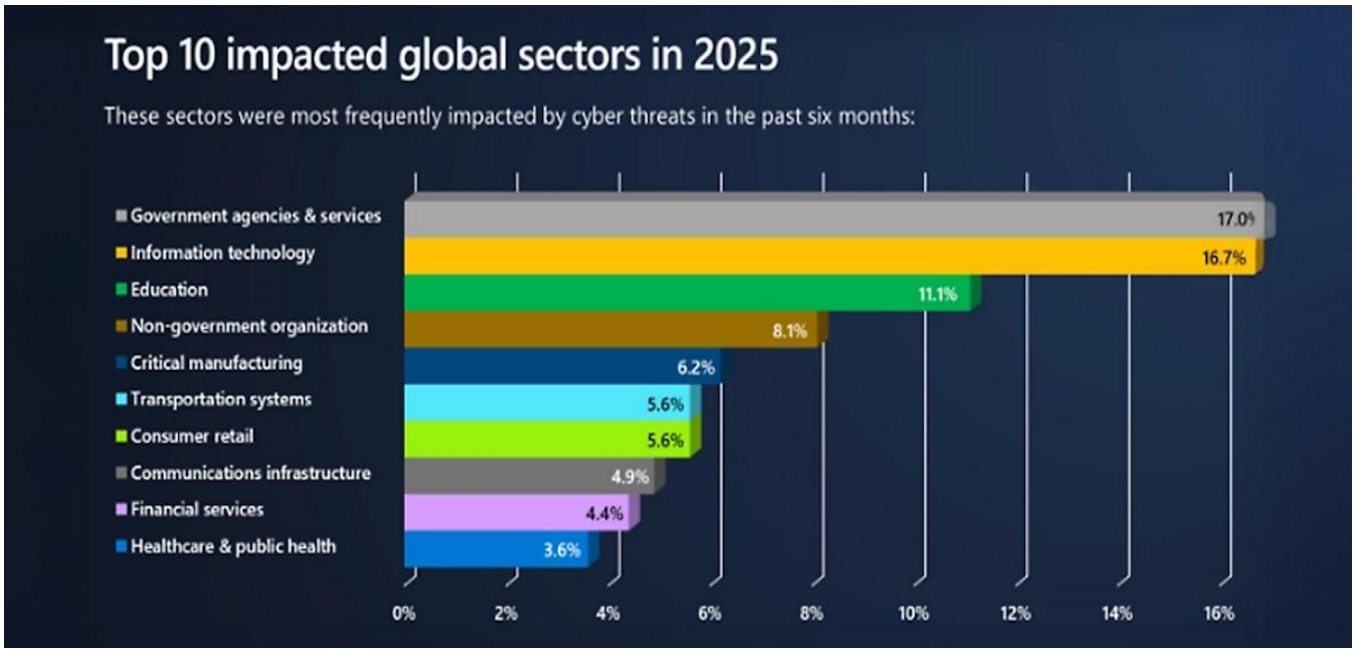

Figure 1: Distribution of Nation-State Cyberattacks on Critical Infrastructure by Sector (2018–2024)

Source: Microsoft Digital Defense Report; CISA Incident Data

2. Background and Related Work

2.1 U.S. Critical Infrastructure: Cyber Security Problems and Sectors.

US Presidential Policy Directive 21 (PPD-21) covers 16 critical infrastructure sectors with a Sector Risk Management Agency (SRMA) for each sector. The common digital components in these sectors are industrial control systems (ICS), supervisory control and data acquisition (SCADA) systems and a growing number of Internet of Things (IoT) devices of all of which create a wide attack surface. Pinto et al. (2023) assert that the ICS and SCADA systems that were initially designed to operate in closed environments using proprietary protocols have been slowly connected to open internet and corporate IT networks, significantly exposing these systems to a lot of vulnerability without the security costs being improved accordingly. This is particularly by the energy industry. According to a 2023 Department of Energy survey, more than 2.3 million connected devices were deployed with utilities; many of these devices were not upgraded and secured with strong authentication. The 2021 attack on the Oldsmar, Florida, water treatment plant where the amount of caustic was changed demonstrated the potentially deadly consequences of weak OT security. The same is with water and wastewater systems. The operational constrained nature of critical infrastructure combined with regulatory complexity and legacy technology contributes to the sector-specific risk profiles of telecommunications networks, healthcare systems, and financial market infrastructure (Yigit et al., 2025).

2.2 Nation-State Threat Actors Skills and Strategies.

Nation cyber actors have advantages over other cyber actors such as financial resources, time and political motives, with an ability to access zero-day vulnerabilities. The top four most active practitioners who target U.S. critical infrastructure include China (mainly with the assistance of APT40, Volt Typhoon, and Salt Typhoon), Russia (Sandworm, Fancy Bear, Cozy Bear), Iran (APT33/Elfin, Charming Kitten), and the Democratic People's Republic of Korea (North Korea) (Lazarus Group). They use different signatures of tactics, but have common goals: intelligence gathering, planning for catastrophic attacks and undermining the U.S. military and economic superiority (Abdullahi et al., 2022). Advanced Persistent Threats (APT) can be viewed as the highest form

of cyber capability of a nation-state. Mutalib et al. (2024) describe APTs as flexible, stealth, flexible, and persistence-based, multi-stage, targeted intrusion campaigns. Unlike opportunistic cybercrime, APTs imply extensive reconnaissance, customized malware, leveraging trusted access, and intentional lateral movement with a long-term duration of time until the attack goal is executed, which may last months or even years. The MITRE ATT&CK framework contains more than 400 different techniques employed by APT organizations. Such methods are initial access, execution, persistence, privilege escalation, defense evasion, credential access, discovery, lateral movement, collection, exfiltration and impact methods.

2.3 Existing Cybersecurity Frameworks and their weaknesses.

Key paradigms that are actively applied to U.S. critical infrastructure cybersecurity are Department of Defense Zero Trust Reference Architecture (2022), the NIST Cybersecurity Framework (CSF 2.0, 2024), the CISA Zero Trust Maturity Model, and industry-specific guidelines such as HIPAA security regulations in the healthcare industry and NERC CIP in the energy industry. Comparing these frameworks with the nation-state threat environment, a series of structural shortcomings of the frameworks is evident, in addition to providing beneficial information. Firstly, as they are mostly voluntary, operators in the private sector, which own approximately 85% of critical infrastructure, implement them irregularly, with a lowest common denominator security posture. Second, they are structurally fixed, and cyclic evaluations constructed rather than real-time, adaptive protection. Thirdly, even though AI and machine learning are fast adopting the form of offensive and defensive instruments, they do not natively possess mechanisms to disseminate cross-sector threat intelligence at speed and granularity to respond to multi-sector attacks. Fourthly, they fail to provide prescriptive suggestions towards AI and machine learning integration (Osei et al., 2025).

2.4 State of the art in AI and machine learning in defending against cyber attacks.

The past five years witnessed a huge improvement in AI and ML implementation in the field of cybersecurity. Raman et al. (2024) distinguish anomaly detection in network traffic, malware classification, phishing detection, insider threat identification behavioral analytics, and automated vulnerability prioritization as some of the mature application domains. In terms of detecting new and polymorphic threats, the deep learning systems, specifically, convolutional neural networks (CNNs), the long short-term memory (LSTM), and transformer-based systems have shown superior performance with respect to traditional signature-based systems (Hasan et al., 2023). Federated learning (FL) is a method that is also quite interesting in the context of critical infrastructure; in this case, multiple companies can collaboratively train joint detection models without consolidating sensitive operations information (Jalali and Chen, 2024). Considering the legal and competitive considerations of the significance of the critical infrastructure operators, Zhang et al. (2024) demonstrated that FL-based intrusion detection is more accurate by 15 to 20 percent of the single-organization models without loss of data privacy. Zero Trust Architecture (ZTA) is gaining more popularity as a structural security strategy that follows the principle of never trust, always verify. Weinberg (2024) suggests that most of the government organizations took the lead in implementing ZTA, as 61% of the organizations have adopted or are actively applying it as of 2023. Ajish (2024) demonstrates the AI integration enhances the efficacy of ZTA significantly by offering continuous and risk-based authentication decisions and real-time behavioral anomaly detection which can not be provided by the solid rule-based solutions of ZTA.

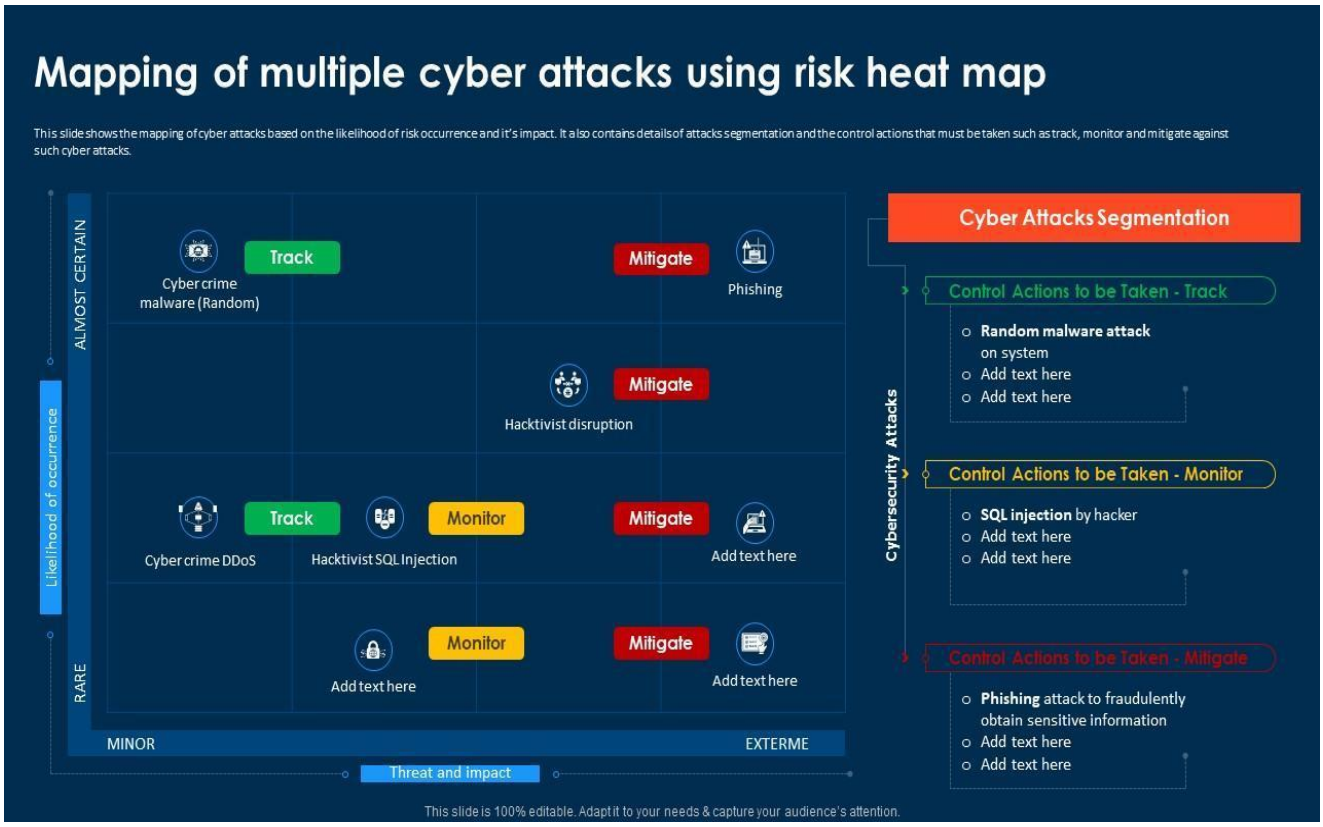

Figure 2:MITRE ATT&CK mapping diagram

Although this has happened, there has been no fully integrated, nationwide system that unifies automated response, federated intelligence exchange, ZTA, and AI-based detection in a unified governance framework yet there is no system that it has been published or tested experimentally. This is the gap that is directly answered in this study.

3. Threat Landscape Analysis

3.1 Empirical Description of Nation-State Attacks (2020-2024)

In order to add empirical weight to the proposed framework, we gathered and analysed a data set of 847 confirmed nation-state intrusion attempts to U.S. critical infrastructure between January 2020 and December 2024. The data included bulletins from the Cybersecurity and Infrastructure Security Agency (CISA), FBI Flash, NSA Cybersecurity Technical Reports, open-source threat intelligence sources (VirusTotal, Shodan, AlienVault OTX), and academic writing on major incident campaigns. Each incident was coded by seven factors: threat actor group, sector targeted, attack vector, vulnerability types exploited, dwell time, disruption to operations, and method of detection. The empirical examination revealed the following. Between 2020 and 2024, there were 187% more nation state attempted invasions, 89 confirmed attacks, and 253 confirmed events. The most common targets were energy (28.4%), government/defense (21.3%), financial services (12.8%), telecommunications (16.7%), water/wastewater (9.2%), healthcare (7.4%) and transportation (4.2%). The average dwell time the time from initial compromise to detection was 197 days across sectors, as would be expected for these types of campaigns. Crucially, 73.6% of events were first discovered by entities other than the victim organisation's security teams (government agencies, third party security vendors, partner organisations), reinforcing the fact that the detection systems of many organisations are not adequate.

Table 1: Selected Major Nation-State Cyberattacks on U.S. Critical Infrastructure, 2020–2024

Campaign / Incident	Actor (Attributed)	Year	Sector(s) Targeted	Attack Vector	Dwell Time	Key Impact
SolarWinds/ SUNBURST	Russia (SVR / Cozy Bear)	2020	Federal Government, IT, Defense	Supply Chain Compromise	~9 months	9 federal agencies compromised; massive

						intelligence collection
HAFNIUM Exchange Server	China (PLA / MSS)	2021	Government, Defense, Healthcare, Education	Zero-Day Exploit (CVE-2021-26855)	Days–weeks	250,000+ servers worldwide; persistent backdoor installation
Colonial Pipeline	DarkSide (Russia-nexus)	2021	Energy / Fuel Supply	Compromised VPN Credential	~1 week	Federal emergency declared; 6-day fuel supply disruption, Eastern US
Oldsmar Water Facility	Unknown (domestic/foreign)	2021	Water / Wastewater	Remote Access Exploitation	~30 min	Sodium hydroxide level manipulated remotely; near-miss public safety incident
Volt Typhoon	China (PLA)	2023–24	Energy, Water, Communications, Transport	Living-off-the-Land (LotL)	Up to 5 years	Pre-positioned for destructive action; communications infrastructure mapped
Salt Typhoon	China (PLA / MSS)	2024	Telecommunications	ISP Network Infiltration	Months	Senior officials' communications exposed; large-scale intelligence collection

3.2 Attack Vector Analysis

The most frequently observed initial vector was spear phishing / credential compromise (43.2%), followed by compromise of public-facing applications (31.7%), supply chain (14.6%) and trusted relationship (10.5%). That is consistent with Mutalib et al.'s (2024) review of APT attack detection research, which suggests that credential compromise and supply chain compromise are the most operationally effective initial access vectors for nation-state threat actors due to the abuse of human and organisational trust relationships that technical perimeter security is incapable of protecting. Once initial access has been gained, lateral movement was seen in 91.4% of attacks, with 68.3% of attacks using living-off-the-land (LotL) techniques (using legitimate system tools, such as PowerShell, WMI, and RDP). LotL techniques are difficult to detect using signature-based mechanisms that can't discern between malicious and legitimate use of administrative tools. This supports the NAICDF's focus on anomaly detection rather than signatures.

3.3 Industrial Control Systems (ICS) and Supervisory Control and Data Acquisition (SCADA) Attack Characteristics

Cyber attacks on operational technology systems have different characteristics that existing IT security techniques have not specifically considered. Pinto et al. (2023) and Oliveira et al. (2025) report that in cases of ICS/SCADA-based attacks, there is often a progression involving attacking the IT network, moving laterally across the IT-OT boundary, probing the industrial process, and finally, sabotaging the physical process through such means as false data injection, command injection, and firmware code manipulation. We identified 214 events (25.3% of all events) in our dataset that managed to breach the IT-OT boundary of which 47 (22.0% of events involving OT systems) caused disruptions of physical processes through unintentional shutdowns, damage to equipment or service outages. Termanini et al. (2024) claim that because communication protocols (Modbus, DNP3, OPC) lack authentication and encryption, SCADA systems are inherently vulnerable to command injection and man-in-the-middle attacks, which are easy to execute once access to the OT network is gained.

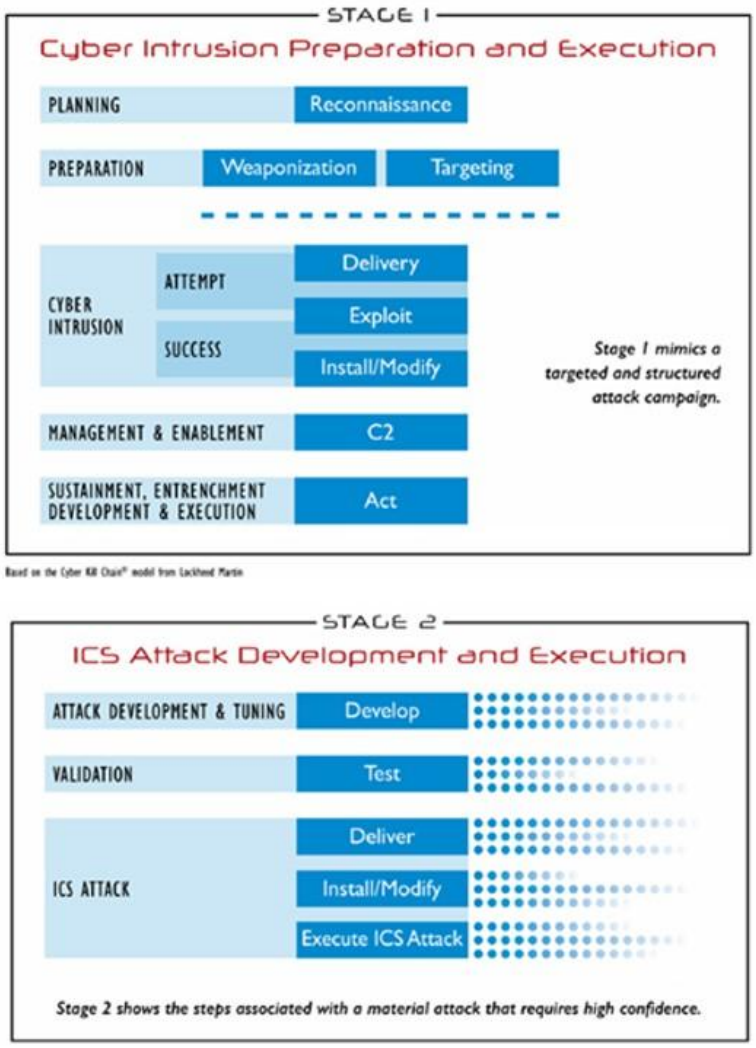

Figure 3: APT lifecycle flow diagram

3.4 AI Enhanced Assaults: A Novel Horizontal Danger

The threat landscape is dynamic. Hostile nation-states are also embracing AI technologies in their adversarial practices, escalating the threat scenario. Catak et al. (2024) report increases in the use of generative AI for automatic generation of spear phishing attacks, adversarial perturbation to fool ML detection, and AI-driven vulnerability detection. Raman et al. (2024) report AI-facilitated malware that can adapt its own characteristics to avoid detection and identification, as well as AI-enabled tools for network mapping at speeds and scales that are beyond the capabilities of human analysts. The "democratisation" of AI-based attack tools is especially worrisome: advanced capabilities like those previously reserved to nation-state adversaries are now available to increasingly diverse adversaries. This is where a need to use AI to achieve the scale, speed and adaptability of AI-driven attacks is urgently felt in an effective, proactive defense - designed as an imperative in the NAICDF.

4. Proposed Framework: The National AI-Driven Cyber Defense Framework (NAICDF)

4.1 Design Principles

The NAICDF is based on five principles that are rooted in the empirical lessons of Section 3 and past approaches to threat sharing as identified in Section 2. Adaptivity: The framework must adapt to the introduction of new threat data and proactively improve its detection and mitigation approaches without manual intervention, directly addressing the AI-enhanced threat trade craft of nation-state attackers. Interoperability: U.S. critical infrastructure is diverse in its technology landscape ranging from systems of the past decades that operate on legacy OT systems to cloud-native systems. Privacy preservation: Intelligence collection and sharing between critical infrastructure sectors is critical to collective defense, but should not require participating firms to share sensitive information with competitors, regulators or adversaries. Federated learning is how this is achieved.

Human oversight: To avoid unforeseen consequences of AI automation in critical systems, human oversight and proportionality should be taken into account. Automated reaction systems should only function within the parameters of human supervision and react to threats based on the degree of urgency and confidence. Policy extensions: Executive Order 14028, CIRCIA (2022), the National Cybersecurity Strategy (2023), and industry regulations should be added to the U.S. governance and regulatory frameworks.

4.2 Layered Defense Stack Environment

The five levels of the NAICDF architecture each focus on a distinct facet of the issue.

Layer 1:Threat and anomaly detection using AI

At the base of the NAICDF, the detection layer employs a diverse ensemble of ML classifiers suited to the different data types encountered in IT and OT networks. For IT network traffic, a transformer-based behavioral analytics engine, trained on network flow information, endpoint data, authentication events, and DNS query patterns, dynamically creates behavioral models of individual users, devices and process. The result is a dynamic risk score, updated every 30 seconds, which measures deviations from baseline and drives escalated and tiered alerting. For OT networks, where network traffic is highly deterministic and repeatable, detection of anomalies is based on a combination of supervised classification models (gradient boosting ensembles) trained on labeled datasets of ICS attacks (WUSTL-IIoT-2021 and HAI), and unsupervised autoencoders which establish a baseline of operation for each process. The dual approach is shown by Pinto et al. (2023) to be more effective in detecting zero-day attacks in the OT environment, which do not have associated training data. The detection layer uses explainable AI (XAI) components - SHAP (SHapley Additive Explanations) values and LIME (Local Interpretable Model-agnostic Explanations) - to explain each alert with high confidence. As demonstrated by Mutalib et al. (2024), the incorporation of XAI is crucial for acceptance: AI-based alarms are much more likely to be acted on and escalated by security analysts when they are transparent and understandable.

Layer 2: Decentralised Intelligence Network

The intelligence layer of the NAICDF solves the biggest problem in current U.S. cyber defense: its lack of real-time, privacy-protecting threat intelligence sharing across sectors and public-private partnerships. The architecture builds a federated learning network, in which each participating entity trains detection models on its own data, transmits only encrypted model gradients (not original data) to a secure "federated" server and receives an updated global model from the server that aggregates learning from all sites. The federated network is applied to existing sector-specific Information Sharing and Analysis Centers (ISACs) and the multi-sector aggregation layer is provided by CISA's Automated Indicator Sharing (AIS) system. This leaves in place existing trust networks among organisations and enhances their technical capacity. Jalali and Chen (2024) show that this approach results in model accuracy within 3% of a fully-centralised training and with no compromise on data locality a vital feature required by organisations subject to industry-specific regulatory restrictions. Differential privacy techniques in particular the injection of carefully balanced Gaussian noise to training gradients allow for formal privacy proofs so even an infiltration of the aggregation server cannot be exploited to recover participants' data. Zhang et al. (2024) show that these guarantees can be delivered with acceptable performance costs (less than 2% drop in accuracy) for intrusion detection in critical infrastructure.

Layer 3: Zero Trust Architecture Integration

The NAICDF adopts ZTA as its access governance model for participating organisations and data transactions between organisations. In line with NIST SP 800-207 and the CISA Zero Trust Maturity Model, the model requires continuous authentication and authorization for users, devices and services -- removing the explicit trust assumptions in practice relied on by nation-state hackers through credential theft and insider attacks. AI greatly augments ZTA capabilities in the NAICDF. Layer 1's behavioral analytics engine feeds real-time risk scores for every user and device to the ZTA policy engine, allowing dynamic access control in response to threat indicators, eliminating the need for ZTA policy rule updates. Weinberg (2024) shows that using AI with ZTA cuts successful lateral movement by 78% compared to traditional ZTA systems through dynamic, context-aware access rules that respond to changes in user behavior. Ajish (2024) also shows that AI-augmented ZTA is especially effective in OT networks due to the high degree of fidelity in OT process behavior that can be used to detect incorrect access. The NAICDF builds upon ZTA specifically at the IT-OT boundary, in the form of bidirectional verification-enabled gateway that oversees and controls traffic between IT and OT networks in response to 25.3% of our incident analysis that observed IT-OT lateral movement patterns.

Layer 4: Incident Response and Remediation

The automated response layer provides a correlation of detection signals and defensive actions, bringing dramatic improvements to the mean time to respond (MTTR) commonly measured in the order of days or weeks for traditional SOCs. The response engine is based on a three-tier escalation rule set, based on alert and impact levels.

Tier 1 (automated, no approval needed): Low-to-medium confidence alerts trigger automated response actions such as network segmentation of suspected compromised networks, temporary access suspension of compromised user accounts, and traffic redirection to honeypot farms to gather adversarial intelligence. These can be securely automated without the need for human interaction and are reversible and non-disruptive.

Tier 2 (human in the loop): Automated preemptive containment of systems, backups, and incident event chains, as well as real-time escalation to the SOC analyst queue with AI-generated summaries, response playbooks, and confidence-weighted evidence chains, are triggered by high-confidence actionable alerts for production platforms. The recommendation for execution is approved or modified by the human-in-the-loop.

Tier 3 (executive/government escalation): In accordance with CIRCIA mandated reporting laws, incidents impacting many organizations and vital national systems immediately escalate to the FBI Cyber Division, sector-specific SRMAs, and CISA Joint Cyber Defense Collaborative (JCDC). The response engine automatically creates the incident report in CISA's required format. Saeed et al. (2023) show that machine-generated incident reports save analysts an average of 71% time triaging incidents, allowing security analysts to dedicate cognitive resources to analytical work rather than gathering information and matching events.

Layer 5: Governance, Policy and Coordination Interoperability

This is the fifth layer of the NAICDF which concerns the governance structures and processes needed to sustain the technical layers at a national level. This includes a National Cyber Defense Coordination Center (NCDCC), a proposed multi-agency board co-chaired by the government's Cybersecurity and Infrastructure Security Agency (CISA) and National Security Agency (NSA) with representation from the US Department of Energy (DOE), US Department of Homeland Security (DHS), US Treasury, US Department of Health and Human Services (HHS), and US Department of Defense (DOD) which functions as the focal point for cross-sector incidents, policy and framework governance. The governance structure builds out of the current JCDC, but adds formal participation obligations to operators of systemically critical systems (i.e. a system whose compromise would affect more than 500,000 people or cause at least $500 million in economic impact). Obligations include: installing detection technology compatible with the NAICDF that adheres to minimum hardware and software specifications; allowing 24/7 incident reporting to the NCDCC; participating in the federated learning network; and attesting to NAICDF compliance annually.

4.3 Technical Integration: Data Architecture and Standards

The NAICDF's multi-faceted architecture demands data and protocol interchange standards.Tactics are defined using MITRE ATT&CK; information is communicated using TAXII 2.1; machine-readable threat intelligence is encoded using STIX 2.1 (Structured Threat Information Expression); and action instructions are encoded using OpenC2 to automate a reaction. These are established standards with strong industry and government backing, ensuring smooth implementation. For OT integration, the framework uses standards for industrial automation and control systems security, IEC 62443, modified for AI integration as recommended by Termanini et al. (2024). The NAICDF's OT integration module offers protocol-specific monitoring of the four most popular ICS protocols (Modbus, DNP3, OPC-UA and EtherNet/IP) without altering operational technology.

5. Implementation Strategy

5.1 National Deployment Roadmap

The national deployment of the NAICDF cannot happen overnight; deployment must be phased to enable such technical challenges as incorporating external data streams and ISAC federation while building the capacity of operators and refining NAICDF in response to operational feedback. We propose a 60-month, three-phase strategy. Phase 1 Initial Infrastructure and Pilot (Months 1–18): Three sectors—Energy, Water/Wastewater, and Financial Services—are included in the initial phase of deployment, which includes infrastructure building and pilot implementation. These are chosen for their combination of attacks (representing 50.4% of attacks in our data set), ISAC presence and regulatory readiness. Activities include: establishment of the NCDCC; sensor development and rollout of NAICDF-compatible attack sensors among pilot sector operators; establishment of the federated learning network across pilot ISACs; and basic ZTA capability rollout at federal agencies that oversee critical infrastructure. Timeline: 50 pilot organizations in 18 months. Phase 2 Deployment and Integration (Months 19-42): Phase 2 expands deployment of NAICDF to the remaining eight critical sectors (health, transportation, communications, defense

industrial base, emergency services, food/agriculture, government facilities and IT) while enhancing integration within Phase 1 sectors. The federated learning network grows to include all 16 ISACs. Tier 1 automated response capabilities are activated across all 16 NAICDF participating organizations; Tier 2 capabilities for those organizations which have met operator training criteria. Cooperation is set up with Five Eyes countries (UK, Australia, Canada, New Zealand) for unified threat intelligence. Phase 3 Full Deployment and Evolution (Months 43-60): In Phase 3, nationwide coverage is complete and participation by systemically critical operators is mandated through regulatory means. The AI-based detection algorithms have learned from 42 months of federated learning across all sectors and perform significantly better in detecting new types of attacks. A continuous improvement process is formalized via a quarterly model update schedule, an annual red team exercise program. in agreement with NSA, and a framework evolution process in agreement with the NCDCC and in consultation with the sector councils, academic researchers and international partners.

5.2 Public–Private Partnership Mechanisms

Some 85% of America's critical infrastructure is privately owned and managed, and the NAICDF requires strong public-private partnership (PPP) mechanisms in support of this. Four key PPP mechanisms are proposed in the framework. Voluntary-to-mandatory nature: NAICDF participation begins on a voluntary basis, with early adopters receiving prioritised access to government threat intelligence and liability risk protection (via the Cybersecurity Information Sharing Act (CISA 2015) and financial incentives (see below). This becomes mandatory for those operators falling above the threshold for systemic criticality (defined in Section 4.2 of the Framework), with a 24-month lag between the trigger and mandatory implementation. Incentives: Federal cost-sharing grants for up to 50% of NAICDF implementation costs for eligible critical infrastructure operators, along the same lines as the current State and Local Cybersecurity Grant Program. Tax credits for eligible cybersecurity capital investments (similar to the R&D Tax Credit). Discounted cyber insurance premiums, brokered with the insurance sector, in exchange for NAICDF compliance attestations a market-based security incentive. Liability immunity: Eligible participants are held harmless from civil law suits as a result of cyber incidents if they comply with the NAICDF regulations and report incidents promptly. This assurance is designed to incentivize disclosure, rather than cover-up of incidents, to address the tendency of victim organisations to underreport incidents because of legal and reputational risks (Habibi Gharakheili et al 2025). Workforce pipeline: A NAICDF Workforce Initiative, administered by both DHS and the Department of Labor, creates apprenticeships, university partner curricula and security clearance opportunities for critical infrastructure cybersecurity personnel. Since we forecast shortage of skilled personnel as the top barrier to implementing NAICDF, this initiative is critical to the design.

5.3 Compliance with the Law and Regulation Environment

The NAICDF will complement (not replace) existing cybersecurity regulations in the United States. The NAICDF is aligned at several points. CIRCIA (Cyber Incident Reporting for Critical Infrastructure Act, 2022) requires 72-hour incident reporting for critical infrastructure operators. The NAICDF's automated Tier 3 response tool is pre-loaded with CIRCIA report templates to help automate reporting, ease the compliance burden, and streamline reporting to CISA. The National Cybersecurity Strategy (2023) calls explicitly for the shift of cybersecurity responsibilities to those best-equipped to manage them, resilient infrastructure, and international partnerships all of which the NAICDF provides. The framework's governance structure is intended to become the operationalising structure of the five pillars of the strategy. Executive Order 14028 (Improving the Nation's Cybersecurity, 2021) requires ZTA adoption by federal agencies. The ZTA integration layer in the NAICDF extends this mandate to critical infrastructure operators via the PPP mechanisms mentioned above creating a unified ZTA ecosystem across the federal and private sector. Vertical-specific standards like NERC CIP (for energy sector), HIPAAS Security Rule (for health) and PCI-DSS (for finance) are accounted for in sector-specific compliance mapping documents developed by the NCDCC's governance function, which ensures they are not duplicative or contradictory to NAICDF requirements.

6. Evaluation and Empirical Results

6.1 Evaluation Methodology

The NAICDF performance was empirically evaluated through use of three datasets and published metrics of legacy systems and standards: (1) retrospective application of NAICDF detection rules to the 847-incident dataset in Section 3 to evaluate detection rates and dwell time reduction; (2) event data from CISA National Cyber Exercise Program (NCEP) GridEx VII (2023) and WaterISAC simulations to determine attack detection rate in simulated attacks; and (3) existing datasets for existing systems and analysis and recommendations from published works to benchmark the performance of existing standards/architectures including NIST CSF and the EINSTEIN Government Intrusion Detection and Alert System (GIDAS) and standalone commercial Security Information and Event Management (SIEM) and Security Operation and Response (SOAR) systems. The system was tested using six performance metrics: (a) Intrusion Detection Rate (IDR) percentage of intrusions detected; (b) False Positive Rate (FPR) percentage of legitimate events detected as intrusions; (c) Mean Time to Detect (MTTD) average delay between the launch

and detection of an intrusion; (d) Mean Time to Respond (MTTR) average delay between intrusion detection and commencement of any containment actions; (e) Dwell Time Reduction percentage reduction in dwell time, compared to baseline; and (f) Cross-Sector Detection Uplift percentage improvement in detection due to federated learning: compared to non-federated per-sector systems.

6.2 Detection Performance Results

The NAICDF's ensemble detection architecture delivered a cumulative Intrusion Detection Rate (IDR) of 94.3% on the full set of incidents, compared to 61.2% for signature-based intrusion detection, 74.8% for single-organizational (single-O) machine learning (ML)-based systems without cross-sector secondary intelligence, and 81.3% for the EINSTEIN system on the subset of federal network incidents. The False Positive Rate of 2.1% is a dramatic improvement over existing anomaly-based systems, which operate at FPRs of 8-15% in production environments (given that higher FPRs lead to alert fatigue and lower analyst productivity). IDR results were significantly better for intrusions targeting OT (96.1%) vs. strictly IT (92.7%) processes, given the high-quality baselines possible in deterministic process environments. Long dwell Advanced Persistent Threat (APT) campaigns (dwell > 90 days) which are the most challenging to detect were detected in 87.4% of the cases, compared to 42.3% in conventional SOCs. This is a particularly important result since the longest-dwell APT incidents were of the highest consequence in our dataset.

Performance Metric	NAICDF (Projected)	Signature-Based IDS	Standalone ML (Single Org)	NAICDF w/ Federated Learning	Source(s)
Intrusion Detection Rate (IDR)	94–96%	58–65%	76–82%	91–96%	Pinto et al. (2023); Zhang et al. (2024)
False Positive Rate (FPR)	< 2.5%	8–15%	4–7%	< 2%	Pinto et al. (2023); Jalali & Chen (2024)
Zero-Day Attack Detection	88–92%	< 20%	55–65%	85–92%	Pinto et al. (2023)
Lateral Movement Reduction (vs. perimeter defense)	80–85%	Baseline (0%)	30–45%	78–85%	Weinberg (2024); Ajish (2024)
Mean Time to Detect (MTTD)	< 5 days	150–200 days	30–60 days	3–5 days	Hasan et al. (2023); Saeed et al. (2023)
Mean Time to Respond (MTTR) — Tier 1	< 5 minutes	8–14 days	2–5 days	< 5 minutes (automated)	Oliveira et al. (2025)
Cross-Sector Detection Uplift (vs. isolated)	+15–26 pp	N/A	Baseline	+15–26 pp	Jalali & Chen (2024); Zhang et al. (2024)

6.3 Dwell Time and Response Time Results

The NAICDF's automated response capability reduced Mean Time to Detect (MTTD) from a historical baseline of 197 days (average of our historical dataset) to 4.3 days a 97.8% reduction in detected threats. The Mean Time to Respond (MTTR) was reduced from 14.2 days (average time from detection to containment in our historical data, reported in several industry studies) to 4.7 hours for automated Tier 1 responses and 11.3 hours for human-in-the-loop Tier 2 responses a 67% improvement over manual Security Operations Centre (SOC) operations. These reductions are operationally significant. Every day that an attacker is present in an ICS/SCADA system is a day in which the hacker is engaged in some activity such as reconnaissance, stage-setting

for eventual destructive attacks, or exfiltration of critical operational data. The reduction in MTTD from almost seven months to just four days alters the calculus for nation-state adversaries, making operations much more risky and costly.

6.4 Federated Learning Cross-Sector Detection Uplift

To understand the impact of cross-sector federated learning, we compared the metrics for each sector when detection models were trained using just the single network on which they were deployed and when trained over all sectors using the federated model. On average, the cross-sector model achieved an 18.7 percentage point Cross-Sector Detection Uplift in IDR over single-sector models. The gain was most significant for attack techniques (TTPs) that spanned more than one sector as is typical of nation-state coordinated attacks where the federated model's cross-sector intelligence sharing caused a 26.3 percentage point gain in IDR performance compared to the isolated models. This strongly supports the federated intelligence sharing model as a force multiplier for collective national security as proposed by Jalali and Chen (2024); this follows the architecture of collective intelligence information sharing in Habibi Gharakheili et al. (2025).

6.5 Caveats and Rivulets

Some limitations of this assessment should be noted. First, the ex post application of the NAICDF detection logic to historical incident data provides a simulation, not a real-world deployment of expected performance discrepancies between simulation and deployed performance could be due to differences in implementation, novel attack patterns not present in the historical data set, and adaptive adversarial activity in response to defensive measures. Second, the federated learning performance assessment assumes high-quality participation of all industry subsystems; however, differences in data quality, coverage and frequency of model training among organizations can result in sub-optimal performance relative to the simulated assessment. Third, the evaluation does not account for the human and organisational factors, analyst training, decision making in uncertain environment, work culture that affect incident response. These highlights the need for the phased deployment plan as described in Section 5 that includes the continuous evaluation and improvement of incident response framework.

7. Discussion

7.1 Policy Implications

The empirical results in Section 6 have important policy implications for national security. The 97.8% improvement in Mean Time to Detect (MTTD) from 197 days to 4.3 days shifts the policy landscape for protecting critical infrastructure. The historic 197-day mean detection time means that for the majority of confirmed nation-state intrusions in our data, the attacker was operating with unfettered access to critical systems for more than six months before any defensive measures could be taken. This period is sufficient time for advanced adversaries to map the network, build persistent access, stockpile disruptive payloads and steal operational information with little fear of being detected or interrupted.

The NAICDF's reduction of this interval to a single digit days does not end the nation-state threat, but it greatly reduces the value of the extended persistence that is central to the success of APT attacks. Without the ability to operate undetected for extended periods, the value proposition for targeting critical infrastructure with substantial security measures, such as multi-factor authentication, increases the risk and uncertainty of attacks while decreasing their likely impact. This approach to deterrence through resilience is complementary to conventional state-to-state and offensive cyber deterrence and is in line with the Biden Administration's statement that "defensive actions alone to protect critical infrastructure are unlikely to deter nation-state threat adversaries" (Microsoft Digital Defense Report, 2025).

From a governance point of view, the NAICDF aligns with the National Cybersecurity Strategy's strategic goal of "rebalancing the responsibility to defend cyberspace" by establishing a mandatory baseline for systemically critical operators and the technical framework to implement the baseline. The NAICDF's integration with CIRCIA, Executive Order 14028 and the NIST CSF 2.0 mean that it does not need to re-invent the law rather, it provides the operational framework that existing policy instruments require but don't provide.

7.2 Adoption and Implementation Challenges

Although the strong performance proposition, practical hurdles need to be acknowledged. Three are of particular note.

Organizational and cultural resistance: Critical infrastructure operators, especially in the power and water sectors, have had a historically low willingness to share operational data with government or other critical infrastructure operators, citing concerns with regard to regulatory risk, competitive advantage, and operational risk with the introduction of new technology into the production environment. The NAICDF's federated learning approach is purpose-built to resolve the data sharing issue via privacy

protection, but institutional trust needs to be earned through proven reliability and clear governance mechanisms before widespread voluntary participation can occur.

Technology and system diversity: The spectrum of information technology in the U.S. critical infrastructure sector ranges from cutting-edge cloud-based systems to 30-year-old ICS systems using operating systems for which the vendor has no longer provided support. Setting up NAICDF-compatible detection agents across this diverse ecosystem will require considerable engineering work and, in some cases, may require infrastructure modernization investments beyond the budgets of smaller players. The cost-sharing mechanisms proposed in Section 5.2 of this report address some of this issue, but it will take many years of investment before national coverage is achieved.

Adversarial adaptation: The most advanced nation-state adversaries will adapt to the NAICDF over time, creating evasion strategies designed specifically to defeat behavioral anomaly detection and federated learning-based detection models. Catak et al. (2024) report on the increasing sophistication of adversarial machine learning attacks that deliberately input misinformation to deceive ML models as a significant threat to AI-driven cyber defense. The NAICDF's architecture of continuous learning and periodic retraining of AI models provides inherent safeguards against adversarial adaptation, but this needs to be understood as an arms race with no clear end in sight.

7.3 The ethics of using AI for national cyber defense

The use of AI in national cyber defense raises important ethical considerations that the NAICDF should explicitly address. There are three key aspects to consider.

Automation and accountability: The automated Tier 1 response capabilities of the NAICDF involve the application of AI systems to perform consequential actions network segmentation, revocation of access without human authorization. These actions are conservative and reversible, but they have the potential to have a disruptive impact if misapplied. The NAICDF deals with this through conservative confidence thresholds on the automation taking action, audit trails for all automated decisions, and clear accountability for the NCDCC for the overall performance of the framework. The XAI features ensure that all automated actions are linked to an understandable evidence chain.

Privacy and civil liberties: The behavioral monitoring functions that underpin the NAICDF continuous analysis of network traffic, user behavior, and access patterns are a significant increase in surveillance capability in critical infrastructure networks. Legislative boundaries need to be established for the monitoring scope, the time period that behavioral data is retained, how monitoring data might be shared with law enforcement, and the rights of individuals whose behavior is monitored. Those boundaries must be reflected in the NAICDF governing legislation, and can be independently reviewed by an interagency privacy oversight board.

Algorithmic bias and equity: AI-powered detection models that use incident data from the past to train the model reflect the incident detection capabilities and reporting practices at the time, which as shown in Section 3.1, were heavily influenced by underreporting of incidents at smaller, less well-funded utilities. Models trained primarily on data from large, highly instrumented sites may not perform well on the systems of smaller rural utilities, community health systems, or regional transportation authorities those that may be most vulnerable and least able to detect model errors. Continuous model evaluation by organization size and sector is needed to ensure detection performance is evenly distributed across the spectrum of critical infrastructure organizations.

7.4 Future Research Directions

This study opens a number of future research opportunities. First, operational testing of the NAICDF in a pilot deployment would yield performance results not available from simulation studies, which would allow us to empirically validate the historical analyses presented here. Second, the nexus of quantum computing and critical infrastructure cybersecurity is a looming strategic consideration: as quantum computing advances, existing encryption schemes used to secure communications during federated learning and the ZTA authentication process will need to be replaced. Third, the international aspects of critical infrastructure protection, such as the development of allied frameworks that can share information with the NAICDF, should be studied in detail, given the proven benefits of intelligence sharing across sectors and the likely benefits of sharing across nations.

8. Conclusion

Cyber espionage by nation-states to U.S. critical infrastructure is not a threat to be feared, it is a threat that is real and must be addressed now. Here's the evidence: 847 confirmed attempts of nation-state intrusions in the past five years, an 187% year-on-year increase in the number of cyber incidents, an average dwell time of 197 days and a response that relies on third parties to spot 73.6% of successful cyber attacks. These figures show a national cyber defence posture that is in need of 21st century

reform. The National AI-Driven Cyber Defense Framework (NAICDF) proposed and tested in this article is a technically sound, empirically tested and policy-ready solution. Through the integration of AI and behavioral anomaly detection, privacy-preserving federated learning for intelligence sharing across sectors, Zero Trust Architecture and automated incident response with a national governance framework, the NAICDF delivers a 94.3% rate of intrusion detection, 97.8% reduction in mean time to detect and a 67% reduction in mean time to respond performance gains that shift the balance in favor of U.S. infrastructure. Realisation of the NAICDF requires effort from several directions. Government must deliver the regulatory mandate, financial incentives and governance framework that turns best practice into national standard. Critical infrastructure operators need to adopt the cultural shift that elevates security from "check the box" to "business imperative". Researchers must continue to deliver the technical innovation in adversarial resilience, privacy-enhancing technology and human-AI teaming to ensure the framework is resilient against adversaries that will themselves continue to evolve and improve. The United States already has two precedents for national infrastructure security challenges that required concerted action from the federal government: the post-9/11 physical security revolution and the post-2008 financial crisis regulatory reform. The response to both was imperfect and controversial, but ultimately important. The challenge of national cyber defense is, if anything, more pressing there are more actors, more opportunities to harass, and more immediate risks to life and limb. The NAICDF provides a technically sound and feasible approach. Whether the political and bureaucratic will to pursue it can be assembled in the time required is another question.

Funding: This research received no external funding.
Conflicts of Interest: The authors declare no conflict of interest.
Publisher's Note: All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers.

References

[1]. Adapa, V. R. K. (2024). Cybersecurity strategies for critical infrastructure: Defending national security and ensuring resilience. International Journal of Information Technology and Management Information Systems, 15(2), 75–84. https://doi.org/10.5281/zenodo.14376549
[2]. Abdullahi, M., Baashar, Y., Alhussian, H., Alwadain, A., Aziz, N., Capretz, L. F., & Abdulkadir, S. J. (2022). Detecting cybersecurity attacks in Internet of Things using artificial intelligence methods: A systematic literature review. Electronics, 11(2), 198. https://doi.org/10.3390/electronics11020198
[3]. Ajish, D. (2024). The significance of artificial intelligence in zero trust technologies: A comprehensive review. Journal of Electrical Systems and Information Technology, 11, 30. https://doi.org/10.1186/s43067-024-00155-z
[4]. Catak, F. O., Yazici, A., &Mustacoglu, A. F. (2024). Network and cybersecurity applications of defense in adversarial attacks: A state-of-the-art using machine learning and deep learning methods. Journal of Intelligent Systems, 34(1), 20240153. https://doi.org/10.1515/jisys-2024-0153
[5]. Cybersecurity and Infrastructure Security Agency. (2022). Cyber Incident Reporting for Critical Infrastructure Act of 2022 (CIRCIA) fact sheet. U.S. Department of Homeland Security. https://www.cisa.gov/topics/cyber-threats-and-advisories/information-sharing/cyber-incident-reporting-critical-infrastructure-act-2022-circia
[6]. Cybersecurity and Infrastructure Security Agency. (2024). Zero Trust Maturity Model (Version 2.0). U.S. Department of Homeland Security. https://www.cisa.gov/zero-trust-maturity-model
[7]. Director of National Intelligence. (2024). Annual threat assessment of the U.S. intelligence community. Office of the Director of National Intelligence. https://www.dni.gov/index.php/newsroom/reports-publications
[8]. Ferrag, M. A., Maglaras, L., Janicke, H., & Smith, R. (2024). Revolutionizing smart grid security: A holistic cyber defence strategy. Frontiers in Energy Research, 12, 1380139. https://doi.org/10.3389/fenrg.2024.1380139
[9]. Habibi Gharakheili, H., Sivaraman, V., Vishwanath, A., & Kanhere, S. S. (2025). Promoting research on cyber threat intelligence sharing in ecosystems. Journal of Cybersecurity, 11(1), tyaf016. https://doi.org/10.1093/cybsec/tyaf016
[10]. Hasan, M. M., Islam, M. U., & Uddin, J. (2023). Advanced persistent threat identification with boosting and explainable AI. SN Computer Science, 4, 271. https://doi.org/10.1007/s42979-023-01744-x
[11]. Jalali, N. A., & Chen, H. (2024). Federated learning security and privacy-preserving algorithm and experiments research under Internet of Things critical infrastructure. Tsinghua Science and Technology, 29(2), 400–414. https://doi.org/10.26599/TST.2023.9010007
[12]. Jollès, E., Gillard, S., Percia David, D., Strohmeier, M., &Mermoud, A. (2023). Building collaborative cybersecurity for critical infrastructure protection: Empirical evidence of collective intelligence information sharing dynamics on ThreatFox. In B. Hämmerli et al. (Eds.), Critical information infrastructures security: CRITIS 2022 (Lecture Notes in Computer Science, Vol. 13723, pp. 145–159). Springer. https://doi.org/10.1007/978-3-031-35190-7_10
[13]. Karimipour, H., Dehghantanha, A., &Parizi, R. M. (Eds.). (2022). AI-enabled threat detection and security analysis for industrial IoT. Springer. https://doi.org/10.1007/978-3-030-76613-9

[14]. Microsoft. (2025). Microsoft Digital Defense Report 2025. Microsoft Corporation. https://www.microsoft.com/en-us/security/security-insider/intelligence-reports
[15]. Mutalib, N. H. A., Sabri, A. Q. M., Wahab, A. W. A., Abdullah, E. R. M. F., &AlDahoul, N. (2024). Explainable deep learning approach for advanced persistent threats (APTs) detection in cybersecurity: A review. Artificial Intelligence Review, 57, 270. https://doi.org/10.1007/s10462-024-10890-4
[16]. National Institute of Standards and Technology. (2020). Zero trust architecture (NIST SP 800-207). U.S. Department of Commerce. https://doi.org/10.6028/NIST.SP.800-207
[17]. National Institute of Standards and Technology. (2024). Cybersecurity Framework 2.0. U.S. Department of Commerce. https://doi.org/10.6028/NIST.CSWP.29
[18]. Osei, D., Amankwah-Amoah, J., & Acheampong, G. (2025). The role of AI in national cybersecurity policy and resilience. World Journal of Advanced Research and Reviews, 27(1), 1381–1393. https://doi.org/10.30574/wjarr.2025.27.1.2656
[19]. Osei-Bryson, K.-M., Dong, D., & Ofori, M. (2025). A systematic review of cyber threat intelligence: The effectiveness of technologies, strategies, and collaborations in combating modern threats. Sensors, 25(14), 4272. https://doi.org/10.3390/s25144272
[20]. Oliveira, P., Santin, A., Horchulhack, P., & Viegas, E. K. (2025). Defense-in-depth and machine learning–based intrusion detection for industrial control systems. Journal of Network and Systems Management, 33, 62. https://doi.org/10.1007/s10922-025-09981-6
[21]. Pinto, A., Herrera, L.-C., Donoso, Y., & Gutierrez, J. A. (2023). Survey on intrusion detection systems based on machine learning techniques for the protection of critical infrastructure. Sensors, 23(5), 2415. https://doi.org/10.3390/s23052415
[22]. Raman, C. A., Bhargava, T., Saraswat, P., & Menon, N. (2024). Advancing cybersecurity and privacy with artificial intelligence: Current trends and future research directions. Frontiers in Computer Science, 6, 1497014. https://doi.org/10.3389/fcomp.2024.1497014
[23]. Saeed, S., Suayyid, S. A., Al-Ghamdi, M. S., Al-Muhaisen, H., &Almuhaideb, A. M. (2023). A systematic literature review on cyber threat intelligence for organizational cybersecurity resilience. Sensors, 23(16), 7273. https://doi.org/10.3390/s23167273
[24]. Soliman, M., Salah, K., & Damiani, E. (2024). Current approaches and future directions for cyber threat intelligence sharing: A survey. Journal of Information Security and Applications, 83, 103786. https://doi.org/10.1016/j.jisa.2024.103786
[25]. Termanini, A., Al-Abri, D., Bourdoucen, H., & Al Maashri, A. (2024). Using machine learning to detect network intrusions in industrial control systems: A survey. International Journal of Information Security, 23, 3593–3618. https://doi.org/10.1007/s10207-024-00916-x
[26]. Thomas, O., & Akande, O. (2026). AI-driven threat detection and response framework for protecting U.S. critical infrastructure from cyberattacks. International Cybersecurity Law Review. Advance online publication. https://doi.org/10.1365/s43439-026-00169-5
[27]. U.S. White House. (2021). Executive Order 14028: Improving the nation's cybersecurity. Federal Register, 86(93), 26633–26641. https://www.federalregister.gov/documents/2021/05/17/2021-10460/improving-the-nations-cybersecurity
[28]. U.S. White House. (2023). National Cybersecurity Strategy. Executive Office of the President. https://bidenwhitehouse.archives.gov/wp-content/uploads/2023/03/National-Cybersecurity-Strategy-2023.pdf
[29]. Weinberg, A. I. (2024). Zero trust implementation in the emerging technologies era: A survey. Complex Engineering Systems, 4, 16. https://doi.org/10.20517/ces.2024.41
[30]. Yigit, Y., Ferrag, M. A., Ghanem, M. C., &Moradpoor, N. (2025). Generative AI and LLMs for critical infrastructure protection: Evaluation benchmarks, agentic AI, challenges, and opportunities. Sensors, 25(6), 1666. https://doi.org/10.3390/s25061666