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

Strengthening U.S. Critical Infrastructure Resilience Through NIST-Aligned Cybersecurity Governance and AI-Driven Threat Detection

Md Humayun Kabir[1]✉, Md Razib[2], Yasin Arafat[3], Ruhul Amin Md Rashed[4] and Zakarya Jesan[5]
[1] *Westcliff University, Irvine, United States*
[2] *MBA (Digital and Strategic Marketing), Westcliff University, Irvine, United States*
[3] *Doctor of Management, International American University, Los Angeles, United States*
[4] *MBA in Management Information Systems, International American University, Los Angeles, United States*
[5] *University of Northern Iowa, Iowa, United States*
Corresponding Author: Md Humayun Kabir, **E-mail**: Humayun9152@gmail.com

| ABSTRACT

U.S. critical infrastructure operators face a persistent gap between high-level cybersecurity frameworks and day-to-day measurable execution, especially under ransomware-driven threat progression. This paper presents an applied, program-to-analytics approach that operationalizes NIST-aligned resilience into auditable actions and metrics while providing a transparent baseline for AI/ML-based threat detection. First, we map five intrusion stages—Initial Access, Privilege Escalation, Lateral Movement, Exfiltration, and Impact—to NIST CSF 2.0 functions and NIST SP 800-53 control-family domains, then define a minimal set of operational metrics (e.g., MFA coverage, patch compliance, MTTD, MTTR, backup restore success, and RTO/RPO achievement) that can be sourced from enterprise systems of record. Second, we implement a sparse-friendly preprocessing and modeling pipeline and evaluate two baseline classifiers on the UNSW-NB15 benchmark dataset (UNSW_NB15_training-set.csv; 175,341 rows; 45 columns) using an 80/20 stratified split (seed=42) and a fixed decision threshold of 0.5. XGBoost achieves ROC-AUC 0.993 and average precision 0.997, with F1 0.969 (TN=10,279; FP=921; FN=575; TP=23,294). Logistic regression (saga) achieves ROC-AUC 0.984 and average precision 0.992, with F1 0.954 (TN=9,230; FP=1,970; FN=281; TP=23,588). The results illustrate baseline tradeoffs under a fixed policy and show how model outputs can be governed through CSF-aligned resilience metrics rather than unsupported deployment claims.

| KEYWORDS

Critical infrastructure resilience; ransomware; nation-state threats; NIST Cybersecurity Framework (CSF) 2.0; NIST SP 800-53 Rev. 5; intrusion detection; UNSW-NB15; XGBoost

| ARTICLE INFORMATION

ACCEPTED: 5 Jun 2025 **PUBLISHED:** 20 Jun 2025 **DOI:** 10.32996/jcsts.2025.7.6.134

1. Introduction

1.1 Background

U.S. critical infrastructure environments face a cybersecurity problem that is less about "awareness" and more about execution under constraints. High availability requirements, legacy dependencies, heterogeneous asset ownership, and tight operational tolerances mean that security controls must be implemented in ways that are measurable, sustainable, and defensible during incidents. Ransomware intensifies this challenge because it targets operational continuity: even short-lived loss of visibility, slow containment, or untested recovery procedures can translate into prolonged service disruption, cascading business impact, and regulatory exposure. These realities motivate resilience programs that are structured, evidence-driven, and continuously monitored rather than built around one-time assessments or tool deployments.

NIST CSF 2.0 is widely adopted because it frames cybersecurity as outcomes across Govern, Identify, Protect, Detect, Respond, and Recover, enabling organizations to organize risk management without locking into a prescriptive architecture or vendor stack [1]. However, CSF-based programs often degrade into labeling exercises when outcomes are not translated into concrete operational measures and decision routines. NIST SP 800-53 Rev. 5 complements CSF by providing an implementation-oriented control catalog organized into families (e.g., AC, IA, CM, AU, SI, SC, IR, CP), which helps assign ownership and establish auditable expectations [2]. In applied resilience settings, the core issue is not the absence of frameworks; it is the persistent gap between framework language and the daily mechanisms that determine whether the organization actually prevents common entry paths, detects intrusions early, contains them quickly, and restores services predictably.

Practical ransomware guidance reinforces that the most valuable activities are those that can be verified. The CISA StopRansomware guidance emphasizes actionable practices—hardening remote access, strengthening identity controls, improving monitoring and response readiness, and validating recovery through backups and restore testing—that align naturally with Protect, Detect, Respond, and Recover outcomes [3]. The applied questions that matter are therefore concrete: What proportion of privileged and remote access is protected with strong authentication? Are patching and configuration baselines enforced within policy windows? Is telemetry sufficient to detect lateral movement and data staging with acceptable time-to-detect? Are containment playbooks rehearsed so time-to-contain is measurable and improving? Can the organization demonstrate recovery objectives (RTO/RPO) through validated testing? These are measurable claims, not narrative assurances.

NIST's incident response guidance further supports this operational framing by emphasizing incident response as an integrated part of cybersecurity risk management rather than an isolated activity that begins after compromise [4]. When incident response is treated as a lifecycle and linked to program outcomes, the quality of Respond and Recover becomes observable through operational timelines and verified recovery results, not through statements of intent [4]. This matters for ransomware because disruption outcomes frequently hinge on time: delayed detection and slow containment increase the probability that events progress from localized compromise to broad operational impact.

Measurement-driven governance is also consistent with established NIST risk management guidance. NIST SP 800-30 Rev. 1 clarifies that risk assessment should support decision-making by informing which actions are necessary and why, while NIST SP 800-37 Rev. 2 (RMF) reinforces that control selection, assessment, and monitoring are continuous activities tied to mission and business objectives [5], [6]. In practice, resilience programs fail when they cannot show that controls are operating effectively over time, especially as environments change (new assets, new access paths, new third-party dependencies). This is precisely the problem continuous monitoring guidance was designed to address. NIST SP 800-137 describes how organizations can build a continuous monitoring strategy that maintains visibility into assets, vulnerabilities, threats, and control effectiveness, while NIST SP 800-137A provides an approach for assessing whether monitoring programs are actually effective rather than merely present [7], [8].

Protect outcomes require the same evidence orientation. Patch management guidance frames patching as preventive maintenance and supports building an enterprise strategy that makes patch compliance measurable and repeatable rather than ad hoc [9]. Security-focused configuration management guidance similarly emphasizes baseline control and disciplined change management to reduce misconfiguration drift that undermines both prevention and detection [10]. These sources converge on a practical point: resilience is not achieved by stating controls, but by tracking measurable execution that demonstrates reduced exposure and improved operational readiness.

Critical infrastructure adds additional complexity when industrial control systems are present. ICS security guidance emphasizes availability and safety constraints, which can limit conventional IT interventions and increase the importance of planning, segmentation, monitoring, and recovery readiness tailored to operational realities [11]. At the enterprise governance layer, cybersecurity risk also must be communicated in a form leaders can govern. NIST IR 8286 formalizes the integration of cybersecurity risk into enterprise risk management (ERM), supporting structured decision-making, prioritization, and accountability beyond purely technical reporting [12]. Threat behavior mapping frameworks can support this translation by connecting observed adversary actions to defensive objectives and detection priorities, making stage-informed planning and telemetry decisions more coherent across stakeholders [13].

Taken together, the background establishes a practical motivation for the approach in this paper: frameworks are necessary but insufficient; resilience requires a measurement plan that can be audited and improved, and any AI/ML detection component must be governed through operational metrics (e.g., time-to-detect, time-to-contain, restore success) rather than treated as a standalone performance claim [1]–[4], [7], [12].

1.2 Problem Statement

Despite broad adoption of NIST-aligned frameworks, many U.S. critical infrastructure organizations still struggle to convert framework language into operationally verifiable resilience. The recurring failure mode is not a lack of "controls" on paper; it is the absence of a disciplined, measurement-driven linkage between (i) threat progression that drives real incidents (e.g., initial access through impact), (ii) framework outcomes that leadership claims to pursue (Govern/Identify/Protect/Detect/Respond/Recover), and (iii) evidence that execution is working under time pressure. CSF 2.0 is explicitly outcome-oriented, but without operational metrics and ownership, it is easy for programs to report alignment while remaining unable to demonstrate reduced exposure, improved detection timeliness, faster containment, or validated recovery capability [1]. SP 800-53 can provide implementation structure, yet organizations commonly treat it as a catalog to "map to" rather than as a basis for measurable control effectiveness and continuous verification [2]. Practical ransomware guidance similarly emphasizes implementable steps, but in practice those steps are often not tied to auditable metrics that can be trended, tested, and used to drive decisions [3], [4].

A second, closely related gap is how AI/ML threat-detection work is presented and consumed. Many applied studies report high-level model metrics without clearly defining the decision policy (e.g., the threshold rule), without presenting operationally meaningful error tradeoffs (false positives vs. false negatives), and without connecting the analytics to a resilience measurement layer that a critical infrastructure operator can govern. Continuous monitoring guidance is explicit that effectiveness depends on sustained visibility and measurable outcomes, not on the presence of tools or point-in-time assessments [7], [8]. Incident response guidance is explicit that response capability must be integrated into risk management and validated through lifecycle execution, not assumed from plans [4]. In other words, even a strong classifier is not a resilience solution unless it is integrated into an evidence-based program that measures what matters operationally (coverage, timeliness, containment, restore success) and can withstand audit and post-incident scrutiny [1]–[4], [7], [12].

This paper addresses these two gaps by posing the following applied problem: How can a critical infrastructure resilience program (a) translate NIST CSF 2.0 outcomes and SP 800-53 implementation domains into a small set of auditable, system-of-record metrics tied to realistic threat stages, and (b) evaluate a baseline AI/ML detector in a transparent, conservative manner—fixed dataset, fixed split, fixed threshold—so that model performance is reported as a bounded baseline rather than overstated as deployment-ready capability?

To ground the analytic component in a reproducible benchmark setting, the study uses UNSW-NB15, a widely referenced intrusion dataset designed to reflect contemporary benign and attack behaviors, with documented dataset construction and evaluation analyses in the literature [14], [15]. The goal is not to claim operational generalization from a benchmark; it is to provide a rigorous baseline, reported with decision-rule clarity and error tradeoffs, that can be governed through the same CSF-aligned resilience metrics used in practice.

1.3 Research Objectives and Contributions

This study is motivated by a practical constraint faced by U.S. critical infrastructure operators: resilience programs are expected to be "NIST-aligned," but alignment is only defensible when it produces auditable evidence of execution and measurable improvement over time. NIST CSF 2.0 provides an outcome-driven structure, and NIST SP 800-53 Rev. 5 provides a control-family implementation vocabulary, but neither automatically yields an operational measurement plan that is tied to realistic threat progression and that can be reviewed in governance cycles with systems-of-record evidence [1], [2]. Meanwhile, ransomware readiness guidance emphasizes implementable actions and validated recovery, yet organizations often lack a compact way to connect those actions to measurable indicators that directly reflect exposure reduction, detection timeliness, containment timeliness, and recovery success [3], [4]. The first objective of this paper is therefore to operationalize NIST-aligned resilience in a way that is measurable, reviewable, and stage-specific rather than abstract.

Specifically, the paper establishes an applied mapping that connects five intrusion stages—Initial Access, Privilege Escalation, Lateral Movement, Exfiltration, and Impact—to NIST CSF 2.0 functions and to selected SP 800-53 family domains to support practical ownership and evidence collection [1], [2]. This mapping is intentionally designed for program use: it is not a comprehensive compliance crosswalk. It is a decision aid that links "what the adversary is trying to do" to "which outcome must improve" and to "which implementation domain must execute," consistent with ransomware readiness priorities [3]. The second objective is to define a minimal, auditable metric set that can be sourced from standard enterprise systems (identity, vulnerability management, SIEM/EDR, incident ticketing/SOAR, and backup/DR testing), enabling continuous tracking and accountability in line with continuous monitoring and incident lifecycle expectations [4], [7], [8].

The third objective is to provide a transparent baseline AI/ML threat-detection evaluation that is conservative by design and compatible with resilience governance. Rather than claiming deployment readiness, the study evaluates baseline classifiers under fixed conditions—single benchmark dataset file, fixed stratified split, fixed preprocessing, and a fixed decision threshold—so that performance is interpretable as bounded evidence rather than a generalized guarantee. UNSW-NB15 is used as the benchmark data source because its construction and evaluation have been documented and it is widely used in intrusion detection research, enabling a reproducible baseline discussion when experimental conditions are explicit [14], [15]. The contribution is therefore two-layered: a program layer (framework-to-metrics mapping that supports audit and governance) and an analytic layer (baseline ML evaluation reported with explicit decision policy and error tradeoffs), designed to be integrated rather than treated as separate "framework" and "model" narratives [1]–[4], [7], [14], [15].

1.4 Paper Organization

The remainder of this paper is organized to separate the resilience program layer from the baseline analytic layer while maintaining a clear linkage between them. Section 2 reviews related work on NIST-aligned critical infrastructure resilience, ransomware preparedness, and AI/ML-based intrusion detection, emphasizing the need for measurable execution and conservative reporting practices rather than framework-only or metric-only narratives [1]–[4], [7], [14], [15]. Section 3 describes the study materials and experimental setup, including the confirmed dataset source (UNSW-NB15 training-set file), labeling, split policy, preprocessing pipeline, baseline models, and evaluation metrics, with dataset usage grounded in prior UNSW-NB15 publications [14], [15]. Section 4 presents the proposed NIST-aligned methodology: a threat-stage mapping to CSF functions and SP 800-53 family domains, and the associated operational metrics that enable auditability and continuous improvement [1]–[4], [7], [8]. Section 5 reports the results, including the framework alignment tables, the threat-stage versus CSF function visualization, and the baseline ML performance under the fixed decision policy, with interpretation explicitly bounded to the dataset and pipeline conditions [14], [15]. Finally, Section 6 discusses implications, limitations, and program integration considerations for U.S. critical infrastructure contexts, and Section 7 concludes the paper with a concise summary and practical recommendations aligned to measurable resilience outcomes [1]–[4], [12].

2. Literature Review

2.1 NIST-aligned resilience and measurable execution

NIST CSF 2.0 is widely adopted because it frames cybersecurity as outcomes across Govern, Identify, Protect, Detect, Respond, and Recover, allowing organizations to structure risk management without prescribing a specific architecture or vendor stack [1]. Its flexibility, however, also enables superficial adoption: organizations can "map" policies and tools to CSF categories while lacking evidence that outcomes are achieved under operational stress. NIST SP 800-53 Rev. 5 addresses part of this implementation gap by providing a detailed control catalog organized into families (e.g., AC, IA, CM, AU, SI, SC, IR, CP), which supports clearer assignment of ownership, implementation scope, and assessment expectations [2]. Even so, literature and practice converge on a recurring failure mode: control catalogs become documentation artifacts if programs do not measure control effectiveness and operational performance over time.

Risk management guidance provides a foundation for why measurable execution matters. NIST SP 800-30 Rev. 1 emphasizes risk assessment as decision support, not a compliance deliverable, motivating prioritization based on what reduces the most material risk [5]. NIST SP 800-37 Rev. 2 (RMF) emphasizes that control selection and assessment must be followed by continuous monitoring as an ongoing management practice tied to mission and business objectives [6]. At the enterprise governance layer, NIST IR 8286 emphasizes integrating cybersecurity risk into ERM to support consistent leadership decision-making and accountability [12]. Collectively, these sources support an applied perspective: resilience is defensible when alignment is translated into measurable capability—coverage, timeliness, and validated recovery—reviewed through governance routines, not when alignment is asserted through mappings alone [1], [2], [6], [12].

For critical infrastructure contexts, baseline practice guidance further reinforces this focus on implementable, verifiable actions. CISA's Cross-Sector Cybersecurity Performance Goals provide a voluntary set of prioritized practices intended to reduce significant risks across sectors [16]. While not a replacement for CSF or 800-53, the CPGs reflect a literature-consistent direction: emphasize concrete practices that can be implemented and assessed, and avoid treating frameworks as endpoints rather than as structures for continuous improvement [1], [16].

2.2 Ransomware readiness, incident response, and continuous monitoring

Ransomware preparedness literature emphasizes that resilience depends on sustained execution across Protect, Detect, Respond, and Recover rather than on prevention alone. CISA's StopRansomware guidance foregrounds implementable actions such as hardening remote access, strengthening identity controls, improving monitoring and readiness, and validating backups and restores [3]. NIST's incident response guidance frames response as a lifecycle integrated into cybersecurity risk management—preparation; detection and analysis; containment, eradication, and recovery; and post-incident activity—reinforcing that response capability must be practiced, measured, and improved [4]. Together, these sources support an evidence-driven interpretation of resilience: the ability to detect and contain quickly and to restore reliably is demonstrated through time-based measures and validated recovery outcomes, not through narrative assertions [3], [4].

Continuous monitoring literature provides the measurement bridge between framework outcomes and operational performance. NIST SP 800-137 describes information security continuous monitoring as a program for maintaining visibility into assets, vulnerabilities, threats, and control effectiveness [7]. NIST SP 800-137A provides guidance for assessing ISCM programs, reinforcing that organizations should be able to show that monitoring is effective rather than merely present [8]. This monitoring-centric framing is particularly relevant to ransomware, where delays in detection and containment increase the probability of escalation toward disruptive impact and complicate recovery. In critical infrastructure and ICS-adjacent environments, constraints on availability and safety also shape what monitoring and response actions are feasible, reinforcing the need for tailored planning and validated recovery readiness [11]. The combined implication across these sources is that resilience programs require measurable signals of execution (e.g., identity control coverage, patch/configuration discipline, detection timeliness, containment timeliness, and restore success) that can be reviewed and improved over time [3], [4], [7], [8], [11].

Threat-informed defense literature provides additional structure for linking observed adversary behavior to defensive priorities. MITRE ATT&CK is widely used to describe adversary tactics and techniques based on real-world observations, supporting more coherent prioritization of telemetry sources, detections, and mitigations [17]. ATT&CK does not itself provide resilience metrics, but it complements stage-based planning by helping organizations justify why specific detection and response capabilities matter for commonly observed adversary behaviors [17].

2.3 AI/ML intrusion detection and UNSW-NB15 benchmarking

A substantial body of work evaluates AI/ML for intrusion detection using public datasets, demonstrating discriminative capability under controlled settings. Survey literature consistently notes that reported performance is sensitive to preprocessing choices, imbalance, feature representation, split policies, and metric selection, and that operational translation is often weak when decision policies and error tradeoffs are not reported clearly [19], [20], [21]. From an evaluation standpoint, this motivates reporting both threshold-independent metrics (e.g., ROC-AUC) and metrics that are informative under class imbalance and operational alerting constraints (e.g., precision-recall behavior and average precision) [18]. The relationship between ROC curves and precision-recall curves has been analyzed formally, supporting the practice of reporting both views to avoid misleading interpretations when the positive class is relatively rare or when false positives have high operational cost [18].

Benchmark datasets remain central because they enable comparative evaluation and reproducibility, but they also introduce limitations regarding generalization. UNSW-NB15 was introduced to address limitations of older datasets by incorporating contemporary benign traffic and synthesized attack behaviors, with dataset construction and evaluation documented in peer-reviewed work [14], [15]. Subsequent studies on UNSW-NB15 demonstrate that representation and feature selection can materially affect results, reinforcing the need for explicit, conservative reporting of experimental conditions and outcomes [22]. In applied contexts, this supports a disciplined stance: benchmark results can provide bounded baseline evidence, but they do not justify deployment claims without external validation across environments and collection pipelines [14], [15], [22].

Finally, the IDS literature increasingly emphasizes the “model-to-operations” integration problem. Even strong benchmark classifiers do not automatically improve resilience unless integrated into monitoring, triage, and incident handling workflows with explicit decision rules and measurable operational outcomes (e.g., time-to-detect and time-to-contain) [4], [7], [8]. This motivates studies that report conservative baseline performance under fixed conditions and explicitly connect analytic outputs to operational governance metrics aligned to recognized frameworks [1]–[4], [7], [8], [18], [14], [15]. Several applied ML studies in other high-stakes domains (e.g., clinical risk stratification and medical image analysis) emphasize conservative reporting practices such as explicit metric definitions, dataset-bounded conclusions, and transparent evaluation protocols, which are directly relevant to avoiding over-claiming in security analytics [23],[24]. While these works are not intrusion detection studies, they reinforce the broader methodological point that performance claims should be anchored to clearly stated data sources, decision policies, and verifiable evaluation outcomes [23], [24].

3. Materials and Experimental Setup

3.1 Dataset and study scope

This study uses the UNSW-NB15 intrusion detection benchmark dataset and is restricted to a single confirmed source file: UNSW_NB15_training-set.csv. UNSW-NB15 was developed to provide a contemporary benchmark for network intrusion detection research, and its dataset design and evaluation have been documented in prior work [14], [15]. In this paper, UNSW-NB15 is used strictly as a controlled benchmark for baseline evaluation under fixed conditions. Accordingly, results are interpreted as dataset-bounded and are not presented as evidence of performance in specific U.S. critical infrastructure networks without external validation [14], [15].

The confirmed dataset properties are: 175,341 records and 45 columns. The binary target is the column label with class counts 1 = 119,341 and 0 = 56,000. No additional files, labels, or multi-class attack categorization are introduced beyond these confirmed elements.

3.2 Experimental design and reproducibility constraints

The evaluation uses a single hold-out test design to provide a transparent baseline comparison across model families. The dataset is split into training and test partitions using an 80/20 stratified split by label with random seed 42. Stratification is applied due to class imbalance to maintain similar class proportions across partitions and reduce avoidable variance in threshold-dependent metrics.

A fixed classification decision threshold of 0.5 is applied for all threshold-based reporting for all models. The study does not perform threshold tuning, cost-sensitive threshold selection, or probability calibration. This constraint is intentional: it ensures that reported confusion counts and precision/recall tradeoffs are attributable to the baseline models under a fixed decision policy rather than to post-hoc optimization [18]. As a result, the reported performance should be read as baseline behavior under the stated split and threshold, not as an optimized operating point.

3.3 Preprocessing and feature handling

UNSW-NB15 includes mixed numeric and categorical fields. The preprocessing is implemented as a sparse-friendly pipeline to support one-hot encoding without forcing dense matrix conversion. A ColumnTransformer applies separate transformations by feature type and produces a sparse output suitable for both linear models and tree-based models.

Numeric features are processed using median imputation followed by StandardScaler(with_mean=False). Categorical features are processed using most_frequent imputation followed by OneHotEncoder(handle_unknown="ignore") with sparse output. The handle_unknown="ignore" setting prevents failures when categories appear in the test partition that were not observed during training.

No manual feature selection, handcrafted rules, or rebalancing procedures are introduced beyond the stated preprocessing. This restriction is maintained to preserve interpretability of the baseline comparison and to avoid introducing unverified assumptions about which variables should be retained or discarded. Prior UNSW-NB15 benchmarking work indicates that preprocessing and feature choices can materially affect outcomes; therefore, the study keeps these steps explicit and fixed [14], [15], [22].

3.4 Baseline models

Two baseline classifiers are evaluated under identical split, preprocessing, and threshold conditions.

Logistic Regression (logreg_saga). This model serves as a linear baseline and is appropriate for high-dimensional sparse feature spaces arising from one-hot encoding. The saga solver is used as specified.

XGBoost Classifier (xgboost). This model serves as a non-linear baseline for structured/tabular classification and is widely used to capture interaction effects without manual feature engineering. The XGBoost method is referenced as the underlying algorithmic system [8]. In this paper it is treated as a baseline model family, not as a tuned state-of-the-art configuration.

The purpose of including both models is to provide a controlled comparison between a sparse linear approach and a boosted-tree approach under the same preprocessing and fixed decision policy, consistent with conservative benchmarking practice on public IDS datasets [14], [15], [22].

3.5 Evaluation metrics and reporting rules

Performance is reported using both threshold-independent and threshold-dependent metrics.

Threshold-independent metrics include ROC-AUC and average precision, which summarize ranking/separability across thresholds. Average precision is included because it is informative under imbalance and complements ROC-AUC; the relationship between ROC and precision-recall perspectives is well established in the evaluation literature [18].

Threshold-dependent metrics are computed at the fixed threshold of 0.5 and include F1-score, precision, recall, and confusion matrix counts (TN, FP, FN, TP). Confusion counts are reported explicitly because they represent operationally meaningful tradeoffs: false positives influence alert volume and analyst workload, while false negatives represent missed detections under the fixed decision rule. In the manuscript, scalar metrics are rounded to three decimals in narrative text, while TN/FP/FN/TP are reported as integers.

3.6 Section linkage

Section 4 presents the NIST-aligned resilience mapping and defines the operational metrics used to measure program execution. Section 5 reports baseline model performance under the fixed split and fixed threshold conditions and interprets results conservatively within the bounds of the dataset and experimental design [14], [15].

4. Methodology

This section presents an applied method that connects NIST-aligned resilience planning to measurable execution and a conservative baseline AI/ML threat detection evaluation. The method is designed for environments where governance must be defensible and operational outcomes must be measurable, which is consistent with the outcome orientation of NIST CSF 2.0 and the implementation structure of NIST SP 800-53 Rev. 5 [1], [2]. The approach also reflects practical ransomware readiness expectations and incident-handling lifecycle integration emphasized in U.S. guidance [3], [4]. The contribution is not a new framework; it is a traceable mechanism that an organization can audit: threat stage → NIST function(s) → 800-53 family domain → resilience action → measurable metric, with an analytic baseline reported under fixed conditions to avoid overstated claims.

4.1 Design principles and boundary conditions

Two principles guide the method. First, resilience claims must be evidenced through measurable indicators that can be sourced from systems of record and reviewed over time. This principle is aligned with continuous monitoring guidance, which treats cybersecurity effectiveness as something that must be observed and managed continuously, not inferred from one-time mappings or tool inventories [7], [8]. Second, analytic results must be reported conservatively with explicit decision policy. Many intrusion detection studies report strong separability metrics while leaving threshold policy ambiguous; this paper fixes the threshold at 0.5 and reports confusion counts to expose the operational tradeoff between alert volume and missed detections, consistent with evaluation best practices under imbalance [18].

The method is explicitly bounded. The analytic baseline uses a single confirmed file from UNSW-NB15 (UNSW_NB15_training-set.csv) and a single 80/20 stratified split (seed = 42) with fixed preprocessing steps and two baseline models. These constraints are intentionally imposed to keep the study reproducible and to avoid introducing unverified assumptions. The resilience mapping and metric definitions are presented as implementable artifacts; they do not claim that the listed targets are universally optimal for all sectors or organizations. Instead, they provide a compact measurement set that is commonly feasible and aligns naturally with CSF functions and widely recognized control-family implementation domains [1]–[4], [7], [12], [16].

4.2 Threat-stage model for applied resilience planning

Ransomware and comparable intrusion campaigns commonly progress through recognizable objectives: establishing access, elevating privileges, expanding reach, staging or moving data, and triggering disruptive impact. Applied planning benefits from a stage model because it reduces ambiguity in accountability. If a disruption occurs, leaders need to ask targeted questions: which stage was enabled by missing controls or weak execution, which function outcome failed to materialize, and which operational indicator should have signaled risk earlier.

This study uses five stages: Initial Access, Privilege Escalation, Lateral Movement, Exfiltration, and Impact (Encrypt/Wipe). These stages are used as planning anchors rather than as a complete taxonomy. They reflect common operational inflection points emphasized in ransomware readiness guidance, where the difference between manageable disruption and severe impact is often determined by identity posture, configuration and patch discipline, monitoring quality, containment readiness, and validated recovery capability [3], [4]. The stage framing is compatible with threat-informed defense practice, where adversary behaviors are described in structured terms to support prioritization of telemetry and mitigations [17].

4.3 Mapping to NIST CSF 2.0 and NIST SP 800-53 Rev. 5 families

NIST CSF 2.0 provides the outcome structure that a critical infrastructure program can communicate and govern: it separates what should be achieved (outcomes) from how it is achieved (implementation) [1]. NIST SP 800-53 Rev. 5 provides a control-family structure that can be used to assign implementation ownership and assessment responsibility across domains such as access control, configuration management, logging, incident response, and contingency planning [2]. In this method, the mapping is not presented as an exhaustive crosswalk. It is designed to be minimal and operational: each stage is associated with the function(s) most directly responsible for reducing risk at that stage and with 800-53 families that reflect practical implementation areas.

Early-stage risk reduction is anchored in Protect. Initial Access is mapped to Protect and anchored to AC/IA because strong authentication and privileged access governance directly reduce the most common credential and remote access pathways [2], [3]. Privilege Escalation remains in Protect and is anchored to AC/CM because least privilege and controlled baselines reduce the likelihood that a foothold becomes an administrative takeover and reduce exploitability through unmanaged drift [2]. Mid-stage intrusion activity emphasizes Detect and Respond. Lateral Movement is mapped to Detect and anchored to AU/SI because detection depends on telemetry completeness, log integrity, and monitoring depth [2], [7]. Exfiltration spans Detect/Respond and is anchored to SC/IR because monitoring for data movement and enforcing communications protections must be paired with triage and containment execution to prevent or limit loss [2], [3], [4]. Finally, Impact emphasizes Recover and is anchored to CP because the organization's ability to restore and meet recovery objectives is the defining resilience criterion once disruption occurs [2], [3], [4].

4.4 End-to-end workflow overview

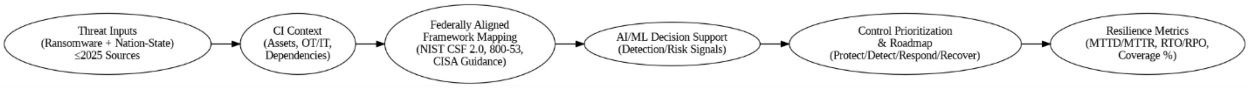

Fig. 1. End-to-end applied resilience-to-ML pipeline for NIST-aligned control mapping, metric definition, and baseline threat detection evaluation (Fig1_Resilience_Pipeline).

Figure 1 (inserted by the author) is intended to show the method as a closed loop. The resilience program begins by identifying the threat stages that must be managed, then maps those stages to CSF functions and 800-53 families to clarify accountability. Next, operational metrics are defined so that execution can be measured using systems of record and reviewed in governance cycles. The analytic baseline is then evaluated under fixed conditions to provide transparent performance and error tradeoffs. Finally, results are interpreted through the resilience measurement lens: model metrics are treated as supportive evidence that must be governed through operational outcomes such as detection and containment timeliness and validated recovery, consistent with CSF and incident lifecycle expectations [1], [4], [7], [12].

4.5 Framework alignment artifact

Threat Stage	NIST CSF Function	800-53 Families	Resilience Action	Metric
Initial Access	Protect	AC/IA	MFA. PAM, hardened remote access	MFA coverage (%)
Privilege Escalation	Protect	AC/CM	Least privilege, patching, config baselines	Petch compliance (%)
Lateral Movement	Detect	AU/SI	Central logging, EDR, segmentation monitor	MTTD (hours)
Exfiltration	Detect/Respond	SC/IR	DLP signals, alert triage, containment playback	Containment time (hours)
Impact (Encrypt/Wipe)	Recover	CP	Immutable backups, restore testing	RTP/RPO achieved

Table 1. Threat-stage alignment to NIST CSF functions, NIST SP 800-53 families, and measurable resilience actions (Table1_Framework_Alignment).

Table 1 is the primary translation layer from threat progression to operational accountability. It is designed to be used in program governance: each row points to a measurable activity and a metric that can be reviewed on a defined cadence. The table is intentionally compact so that it remains actionable; comprehensive crosswalks often fail in practice because they are too large to govern and too diffuse to assign ownership. The focus is on the parts of resilience that determine whether an organization can prevent common entry paths, detect intrusions early, constrain damage quickly, and restore services reliably under disruption [1]–[4], [7].

4.6 Operational metric definitions and evidence sources

Metric	Definition	Data Source	Target	NIST CSF Function
MFA Coverage	Percent of privileged + remote access protected	IdP/IAM logs	≥ 95%	Protect
Patch	Compliance Percent of assets patched within policy window	Vuln scanner/CMDB	≥ 90%	Protect
MTTD	Mean time to detect suspicious activity	SIEM alerts	↓	Detect
MTTR	Mean time to contain/eradicate	IR tickets/SOAR	↓	Respond
Backup	Restore Success Percent of restore tests passing	Backup platform reports	≥ 95%	Recover
RTO/RPO	Achieved Recovery objectives achieved during exercises/incidents	DR tests/postmortems	Yes	Recover

Table 2. Operational definitions and data sources for resilience metrics aligned to NIST CSF functions (Table2_Resilience_Metrics).

Table 2 defines what “measurable” means in the method. Metrics were selected to satisfy three conditions. They must be retrievable from systems of record, interpretable by both technical and governance stakeholders, and aligned to CSF function outcomes so that trends can be used to justify prioritization and investment decisions. Coverage metrics such as MFA coverage and patch compliance function as leading indicators of exposure. Timeliness metrics such as MTTD and MTTR function as operational performance indicators that reflect whether monitoring and response execution are effective. Recovery validation metrics such as restore success and RTO/RPO achievement function as proof that Recover outcomes are achievable in practice, consistent with ransomware readiness guidance emphasizing validated backups and recovery planning [3], [4]. This measurement design aligns with continuous monitoring principles, where the goal is not to assert capability but to demonstrate it repeatedly through evidence and trends [7], [8].

4.7 Baseline analytic pipeline integration

The analytic component is included to provide a conservative baseline for threat detection that can be governed through the same resilience measurement model. The baseline is evaluated on UNSW-NB15 under the confirmed constraints described in Section 3 and grounded in UNSW-NB15 documentation literature [14], [15]. Two model families are used to represent common baseline choices for tabular intrusion detection: a sparse linear classifier (logreg_saga) and a boosted tree classifier (xgboost) [8]. The fixed threshold of 0.5 is a deliberate policy constraint that forces reporting of operationally meaningful error tradeoffs, which is important because model outputs influence alert volume and missed detection risk. The interpretation framework in this method is strict: model metrics do not substitute for resilience outcomes. Instead, they are treated as analytic evidence that must ultimately be validated by operational performance measures such as MTTD and MTTR and by recovery validation measures such as restore success and RTO/RPO achievement, consistent with the program logic of CSF and incident response lifecycle guidance [1], [4], [7], [12].

5. Results and Analysis

This section reports (i) the NIST-aligned framework artifacts (Tables 1–2), (ii) the provided threat-stage versus CSF function heatmap values (Figure 2), and (iii) baseline model performance on UNSW-NB15 under the fixed split, fixed preprocessing, and fixed threshold policy (Figures 3–4). All interpretations are conservative and bounded to the single confirmed dataset source file (UNSW_NB15_training-set.csv), the 80/20 stratified split (seed = 42), and the fixed decision threshold of 0.5 [14], [15]. The analytic results are presented as baseline evidence supporting a measurement-governed resilience discussion, not as deployment-grade claims for U.S. critical infrastructure networks.

The framework alignment and metric definitions are summarized in Tables 1–2.

5.1 Framework artifacts and operational interpretation

Tables 1 and 2 operationalize the paper’s central applied claim: resilience must be measurable to be governable. NIST CSF 2.0 provides outcome structure across Protect, Detect, Respond, and Recover, while NIST SP 800-53 Rev. 5 provides an implementation vocabulary that supports assignment of ownership and assessment scope [1], [2]. However, framework mapping alone is insufficient unless it is coupled to metrics that can be validated from systems of record and reviewed over time. Table 1 provides a compact alignment from threat stages to CSF functions and 800-53 family domains. Table 2 defines a minimal metric set that can be retrieved from enterprise systems such as identity logs, vulnerability management tools, SIEM/EDR alert records, incident ticketing/SOAR, and backup/DR testing reports.

Two operational patterns emerge from the mapping that are consistent with ransomware readiness guidance and incident lifecycle expectations. First, early-stage resilience emphasizes exposure reduction through Protect outcomes: strengthening identity and privileged access controls, reducing exploitable conditions through patching and baseline enforcement, and hardening common entry paths [3], [4]. This is why Table 1 associates Initial Access and Privilege Escalation with Protect and anchors them to AC/IA/CM family domains [2]. Second, once an adversary is active internally, timeliness becomes the dominant operational determinant: Detect and Respond outcomes are reflected in how quickly suspicious activity is recognized (MTTD) and how quickly containment/eradication actions are executed (MTTR or containment time) [4], [7], [8]. Finally, the mapping treats Recover as a validated capability rather than a policy statement, emphasizing restore testing and achievement of recovery objectives (RTO/RPO) in exercises or real incidents, consistent with ransomware recovery guidance [3], [4]. In critical infrastructure contexts, this "validated recovery" posture is essential because disruption tolerance is low and recovery failures can produce extended service impact [12], [16].

5.2 Threat-stage versus CSF function emphasis (heatmap)

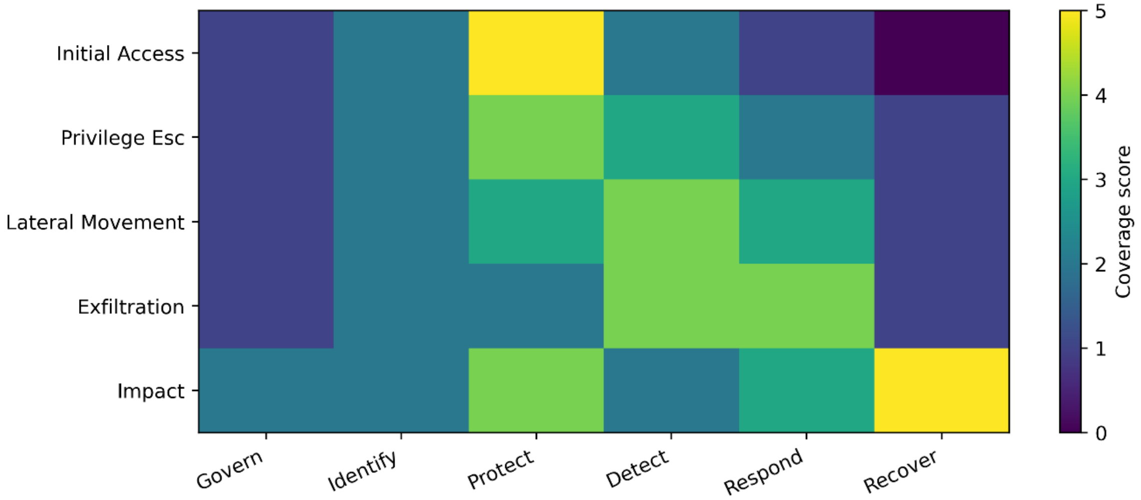

Fig. 2. Stage-to-function emphasis across Govern, Identify, Protect, Detect, Respond, and Recover for the five threat stages using the provided ordinal values (Fig2_ThreatxCSF_Heatmap).

Figure 2 reports the provided ordinal values as a planning visualization. The most prominent pattern is the shift from Protect-dominant emphasis in early stages to Detect/Respond dominance in mid-stages and Recover dominance at Impact. For Initial Access, Protect is highest (5), consistent with the centrality of identity hardening and remote access protection in reducing the probability of entry [2], [3]. For Lateral Movement and Exfiltration, Detect and Respond increase (Detect up to 4; Respond up to 4), reflecting that internal propagation and data staging require monitoring depth, signal quality, triage discipline, and rapid containment to prevent escalation [4], [7]. For Impact, Recover is highest (5), reinforcing that resilience is ultimately expressed as a validated ability to restore and meet recovery objectives, not solely as the absence of incidents [1], [3], [4]. The heatmap is not presented as a measured effectiveness result; it is used to communicate a structured emphasis model for capability planning aligned to CSF functions.

5.3 Baseline model performance on UNSW-NB15

UNSW-NB15 is used here as a controlled benchmark dataset for baseline threat detection evaluation. Dataset design and evaluation analyses have been documented in prior work, supporting its use when experimental conditions are stated explicitly and claims are bounded to the benchmark setting [14], [15]. Consistent with conservative benchmarking practice, this study reports both threshold-independent and threshold-dependent metrics and includes confusion matrix counts at a fixed decision threshold to make error tradeoffs explicit [18]. Two baseline model families are evaluated under identical preprocessing and split conditions: logistic regression (saga) and XGBoost [8]. No threshold tuning or calibration is performed.

At the fixed threshold of 0.5, xgboost achieves ROC-AUC = 0.993 and average precision = 0.997. Threshold-based performance is F1 = 0.969, precision = 0.962, and recall = 0.976. The corresponding confusion matrix counts are TN = 10,279; FP = 921; FN = 575; TP = 23,294.

Under the same fixed conditions, logreg_saga achieves ROC-AUC = 0.984 and average precision = 0.992. Threshold-based performance is F1 = 0.954, precision = 0.923, and recall = 0.988. The confusion matrix counts are TN = 9,230; FP = 1,970; FN = 281; TP = 23,588.

These outcomes show a clear baseline tradeoff under a fixed decision policy. Logistic regression produces higher recall (0.988 vs. 0.976) and fewer false negatives (281 vs. 575), which is favorable if the dominant objective is minimizing missed detections at the fixed threshold. However, the false positive burden is materially higher (1,970 vs. 921), which implies increased alert volume and potentially increased triage workload. XGBoost produces higher precision (0.962 vs. 0.923) and fewer false positives, which is favorable for reducing alert volume and improving triage feasibility, but it produces more false negatives under the same fixed threshold. Because the threshold is fixed and not tuned, these results are reported as baseline behavior rather than as optimized operational operating points.

5.4 Score distribution and ROC-AUC visualizations

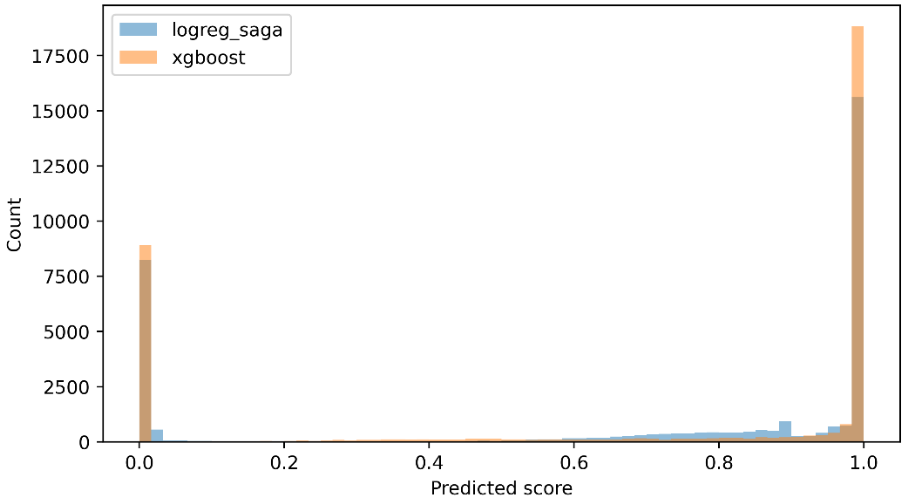

Fig. 3. Distribution of predicted scores by class for the two baseline models under the fixed split and preprocessing pipeline (Fig3_Score_Distributions).

Figure 3 provides a qualitative view of score separation and overlap, which helps interpret why two models with high ROC-AUC and average precision can produce different confusion outcomes at a fixed threshold. In imbalanced and high-stakes detection settings, small differences in score overlap around the decision threshold can materially affect alert volume and miss risk. Reporting score distributions alongside ROC-AUC and average precision is consistent with the general recommendation to interpret performance through both ranking metrics and decision-policy outcomes rather than relying on a single headline metric [18].

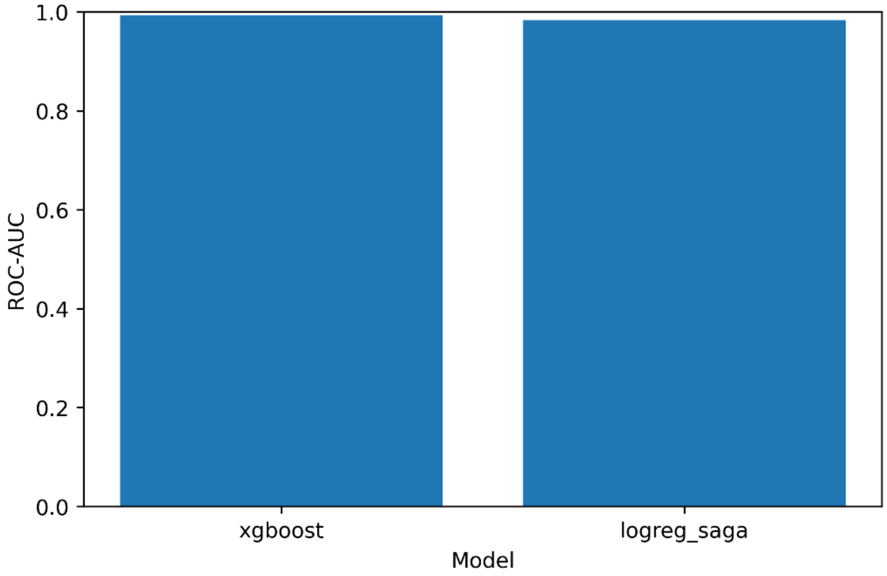

Fig. 4. ROC-AUC curves for xgboost and logreg_saga on the test partition under the fixed split and preprocessing pipeline (Fig4_Model_ROCAUC).

Figure 4 reports ROC curves and summarizes separability via ROC-AUC. Both baselines show strong separability on the benchmark partition (0.993 and 0.984). This does not imply direct transfer to operational critical infrastructure telemetry. Benchmark datasets differ from operational environments in collection pipelines, feature distributions, and adversary behaviors. Therefore, the results are interpreted as bounded evidence under fixed benchmark conditions and are not used to claim deployment readiness without external validation [14], [15].

5.5 Program relevance under conservative assumptions

The applied relevance of these results is in how they can be governed through resilience metrics rather than in the absolute magnitude of ROC-AUC. CSF 2.0 and RMF-aligned governance expect measurable program outcomes, and continuous monitoring guidance emphasizes sustained visibility into control effectiveness [1], [6], [7]. If an organization integrates an ML-based detector into security operations, its value should ultimately be reflected in operational metrics such as MTTD and MTTR and in validated recovery outcomes, not in model metrics alone [4], [7], [8]. Under the fixed 0.5 threshold, the two baselines imply different operational risks: the higher false positive rate of logreg_saga would likely increase analyst workload and could degrade containment timeliness if triage becomes saturated; the higher false negative count of xgboost relative to logreg_saga at the fixed threshold implies higher miss risk that could delay detection and allow further progression before response activation. This paper does not claim these effects occurred; it provides a measurement-governed interpretation framework so that such effects can be evaluated in practice through the metrics defined in Table 2 and the incident lifecycle evidence emphasized in U.S. guidance [3], [4], [7].

6. Discussion

This paper is intentionally positioned as an applied baseline study: it does not propose a new cybersecurity framework, and it does not claim deployment-ready ML performance for U.S. critical infrastructure. Its contribution is the linkage between two layers that are frequently treated separately in practice: (i) NIST-aligned resilience planning that is governable only when tied to measurable execution and evidence, and (ii) an AI/ML threat-detection baseline that is interpretable only when the decision policy and error tradeoffs are reported explicitly. In critical infrastructure environments, “alignment” that cannot be measured and defended during an incident is operationally weak. NIST CSF 2.0 provides outcome structure, and NIST SP 800-53 provides implementation domains, but neither guarantees that a program has measurable signals proving that Protect, Detect, Respond, and Recover outcomes are being achieved [1], [2]. The mapping and metrics in this paper are therefore designed to be used as governance artifacts: they allow a program owner to define what evidence is expected and to monitor that evidence over time, consistent with risk management and continuous monitoring expectations [6]–[8], [12].

6.1 Interpreting the framework-to-metrics layer

The stage-based mapping (Initial Access through Impact) is used because it aligns to how ransomware and comparable intrusions manifest operationally. It improves practical accountability by making it harder to hide behind generic statements such as "we are CSF-aligned." Under this method, each stage has an associated function emphasis, implementation domain, and metric. That structure matters for governance: it turns strategic intent into reviewable indicators. For example, if Initial Access is repeatedly enabled through credential abuse or weak remote access controls, a program should not merely report "Protect controls exist"; it should show MFA coverage trends from IAM logs and demonstrate that privileged access controls are enforced consistently [2], [3]. If incidents show delayed discovery of lateral movement, the program must explain telemetry coverage, logging centralization quality, and MTTD evidence derived from alert timelines rather than asserting "we have a SIEM" [4], [7]. If impact events cause extended disruption, the program should be able to show restore testing success rates and whether recovery objectives are actually met in exercises and incidents, not simply that backups exist [3], [4]. This aligns with the core premise of continuous monitoring guidance: effectiveness is demonstrated through measurable outcomes and sustained visibility, not assumed from tool presence or policy documentation [7], [8].

6.2 Interpreting the ML baseline under a fixed decision policy

The ML baseline is deliberately conservative. The study uses a single benchmark dataset file (UNSW-NB15 training-set file), a single stratified hold-out split (seed = 42), and a fixed threshold of 0.5. These constraints are not limitations by accident; they are constraints by design to avoid overfitting the narrative to tuned operating points or to unreported decision-policy changes. Under the fixed threshold, xgboost produces fewer false positives (921 vs. 1,970) and higher precision (0.962 vs. 0.923), while logreg_saga produces fewer false negatives (281 vs. 575) and higher recall (0.988 vs. 0.976). These are not "better vs. worse" outcomes; they are different risk postures at a fixed policy. Higher false positives imply higher alert volume and higher triage workload; higher false negatives imply higher miss risk at the decision threshold. Because critical infrastructure security operations are constrained by staffing, response playbooks, and time, the tradeoff is program-relevant even in a benchmark study: the decision to prioritize fewer false positives or fewer false negatives affects triage load, containment speed, and potentially the likelihood that events progress into more damaging stages [4], [7], [18].

The method in this paper explicitly rejects a common misstep: treating high ROC-AUC as sufficient proof of operational value. Ranking metrics are informative, but operational impact is determined by thresholded decisions and workflow integration. That is why confusion counts at the fixed threshold are treated as first-class evidence. This is consistent with evaluation literature emphasizing that ROC and precision-recall views can tell different stories under imbalance and that decision policy must be explicit when results are intended to inform real alerting systems [18]. In the applied resilience framing, the correct question is not "Which model has a higher ROC-AUC?" but "Under our fixed decision policy and constraints, what error tradeoff do we accept, and how will we measure whether it improves or degrades operational outcomes such as MTTD and MTTR?" [4], [7], [8].

6.3 Benchmark validity and why conservative claims matter

UNSW-NB15 is widely used and has documented dataset design and evaluation analyses, which justifies its use for controlled baseline comparisons [14], [15]. However, benchmark validity does not equal operational validity. Real critical infrastructure networks differ in asset composition, telemetry availability, traffic patterns, change management discipline, and adversary behaviors. These factors can shift feature distributions and degrade model performance relative to benchmark conditions. The literature on UNSW-NB15 and broader IDS evaluation repeatedly reinforces that results are sensitive to preprocessing, representation, and split policy, which motivates explicit reporting and conservative claims [14], [15], [22]. Accordingly, this paper treats the ML results as bounded evidence and does not claim direct transferability to operational environments. In a submission context, this restraint improves credibility: it shows reviewers that the paper understands the boundary between benchmark experimentation and operational assurance. This conservative stance is consistent with applied ML reporting patterns in other high-stakes domains, where clarity on evaluation conditions and bounded claims is treated as a credibility requirement [23],[24].

6.4 Practical implications for resilience programs

The most practical implication of this paper is a template for governance. A critical infrastructure operator can adopt the stage-based mapping and metrics as a minimum viable measurement layer and then govern both control execution and analytics integration through those metrics. Protect can be governed through measurable coverage and compliance indicators (e.g., MFA coverage and patch compliance). Detect and Respond can be governed through timeliness indicators (e.g., MTTD and MTTR) derived from SIEM/EDR and IR records. Recover can be governed through restore success rates and whether RTO/RPO objectives are consistently achieved in tests and incidents. This provides an auditable mechanism for continuous improvement that is consistent with CSF's outcome orientation, RMF's continuous monitoring expectations, and incident response lifecycle management [1], [4], [6]–[8].

From an AI/ML perspective, the implication is also practical: any model that is introduced into security operations should have a documented decision policy and should be evaluated not only by model metrics but by program metrics that reflect operational outcomes. A model that reduces false positives may improve triage feasibility and containment speed; a model that reduces false negatives may reduce missed detections. Which outcome is preferable depends on mission risk tolerance and operational capacity. This paper provides a structured way to connect that choice to measurable program performance rather than subjective preference.

6.5 Limitations and future work

This study has clear limitations that bound its conclusions. The analytic evaluation uses a single benchmark dataset file and a single hold-out split; therefore, results reflect baseline performance under those fixed conditions and should not be generalized as operational assurance for critical infrastructure networks [14], [15]. The threshold is fixed at 0.5 and is not tuned; therefore, reported precision/recall tradeoffs reflect the fixed policy rather than an optimized operating point. Additionally, the paper does not include external validation across different environments, data sources, or time periods, which would be required for stronger claims about robustness and deployment suitability. Future work should therefore focus on extending the evaluation to additional datasets or real-world telemetry under privacy-preserving governance, conducting robustness checks under distribution shift, and evaluating how model integration affects operational metrics such as MTTD and MTTR in practice, consistent with continuous monitoring and incident response lifecycle expectations [4], [7], [8].

7. Conclusion

This paper presented an applied approach for strengthening U.S. critical infrastructure resilience by linking NIST-aligned frameworks to measurable execution and by reporting a conservative AI/ML threat-detection baseline under fixed, transparent conditions. The method translates five intrusion stages into CSF 2.0 function emphasis and SP 800-53 family implementation domains, then defines auditable metrics that can be sourced from enterprise systems of record to govern Protect, Detect, Respond, and Recover outcomes [1]–[4]. Using UNSW-NB15 as a benchmark dataset, two baseline classifiers were evaluated under a fixed 80/20 stratified split (seed = 42) and a fixed threshold of 0.5. XGBoost achieved ROC-AUC 0.993 and average precision 0.997 with F1 0.969 (TN=10,279; FP=921; FN=575; TP=23,294), while logistic regression achieved ROC-AUC 0.984 and average precision 0.992 with F1 0.954 (TN=9,230; FP=1,970; FN=281; TP=23,588). These results illustrate baseline error tradeoffs under a fixed decision policy and reinforce that model metrics must be interpreted through operational risk and program constraints rather than treated as deployment-ready guarantees.

The central takeaway is practical: resilience should be governed through measurable evidence, and analytics should be treated as a support component whose value must be validated through operational outcomes such as detection timeliness, containment timeliness, and validated recovery capability. By providing compact mapping from threat stages to NIST-aligned actions and metrics and by reporting a conservative baseline detector with explicit decision policy, this paper offers a submission-ready, audit-oriented foundation for critical infrastructure organizations seeking defensible and measurable resilience improvement.

References

[1] National Institute of Standards and Technology, "The NIST Cybersecurity Framework (CSF) 2.0," NIST CSWP 29, Feb. 2024. https://doi.org/10.6028/NIST.CSWP.29 (NIST Publications)

[2] National Institute of Standards and Technology, "Security and Privacy Controls for Information Systems and Organizations," NIST SP 800-53 Rev. 5, Sep. 2020. https://doi.org/10.6028/NIST.SP.800-53r5 (NIST Computer Security Resource Center)

[3] Cybersecurity and Infrastructure Security Agency (CISA), "StopRansomware Guide," Oct. 2023. https://www.cisa.gov/stopransomware/ransomware-guide (CISA)

[4] National Institute of Standards and Technology, "Incident Response Recommendations and Considerations for Cyber Risk Management," NIST SP 800-61 Rev. 3, Apr. 2025. https://doi.org/10.6028/NIST.SP.800-61r3 (NIST Publications)

[5] National Institute of Standards and Technology, "Guide for Conducting Risk Assessments," NIST SP 800-30 Rev. 1, Sep. 2012. https://doi.org/10.6028/NIST.SP.800-30r1 (NIST Computer Security Resource Center)

[6] National Institute of Standards and Technology, "Risk Management Framework for Information Systems and Organizations," NIST SP 800-37 Rev. 2, Dec. 2018. https://doi.org/10.6028/NIST.SP.800-37r2 (NIST Publications)

[7] National Institute of Standards and Technology, "Information Security Continuous Monitoring (ISCM) for Federal Information Systems and Organizations," NIST SP 800-137, Sep. 2011. https://doi.org/10.6028/NIST.SP.800-137 (NIST Computer Security Resource Center)

[8] National Institute of Standards and Technology, "Assessing Information Security Continuous Monitoring (ISCM) Programs: Developing an ISCM Program Assessment," NIST SP 800-137A, May 2020. https://doi.org/10.6028/NIST.SP.800-137A (NIST Computer Security Resource Center)

[9] National Institute of Standards and Technology, "Guide to Enterprise Patch Management Planning: Preventive Maintenance for Technology," NIST SP 800-40 Rev. 4, Apr. 2022. https://doi.org/10.6028/NIST.SP.800-40r4 (NIST Computer Security Resource Center)

[10] National Institute of Standards and Technology, "Guide for Security-Focused Configuration Management of Information Systems," NIST SP 800-128, Aug. 2011. https://doi.org/10.6028/NIST.SP.800-128 (NIST Computer Security Resource Center)

[11] National Institute of Standards and Technology, "Guide to Industrial Control Systems (ICS) Security," NIST SP 800-82 Rev. 2, May 2015. https://doi.org/10.6028/NIST.SP.800-82r2 (NIST Computer Security Resource Center)

[12] K. Stine, G. Witte, S. Quinn, R. K. Gardner, and K. M. Stine, "Integrating Cybersecurity and Enterprise Risk Management (ERM)," NIST IR 8286, Oct. 2020. https://doi.org/10.6028/NIST.IR.8286 (NIST Computer Security Resource Center)

[13] T. Chen and C. Guestrin, "XGBoost: A Scalable Tree Boosting System," in Proc. 22nd ACM SIGKDD Int. Conf. Knowledge Discovery and Data Mining (KDD), 2016. https://doi.org/10.1145/2939672.2939785 (ACM Digital Library)

[14] N. Moustafa and J. Slay, "UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems (UNSW-NB15 Network Data Set)," in Proc. MilCIS, 2015. https://doi.org/10.1109/MilCIS.2015.7348942 (NIST Publications)

[15] N. Moustafa and J. Slay, "The evaluation of network anomaly detection systems: Statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set," Information Security Journal: A Global Perspective, 2016. https://doi.org/10.1080/19393555.2015.1125974 (CISA)

[16] Cybersecurity and Infrastructure Security Agency (CISA), "Cross-Sector Cybersecurity Performance Goals (CPGs)," official website. https://www.cisa.gov/cross-sector-cybersecurity-performance-goals (CISA)

[17] MITRE, "MITRE ATT&CK®," official website. https://attack.mitre.org/ (MITRE ATT&CK)

[18] J. Davis and M. Goadrich, "The Relationship Between Precision-Recall and ROC Curves," in Proc. 23rd Int. Conf. Machine Learning (ICML), 2006. https://doi.org/10.1145/1143844.1143874 (ACM Digital Library)

[19] A. H. Ali, M. Charfeddine, B. Ammar, B. B. Hamed, F. Albalwy, A. Alqarafi, and A. Hussain, "Unveiling machine learning strategies and considerations in intrusion detection systems: a comprehensive survey," *Frontiers in Computer Science*, 2024. https://doi.org/10.3389/fcomp.2024.1387354

[20] L. Ashiku and C. Dagli, "Network Intrusion Detection System using Deep Learning," Procedia Computer Science, 2021. https://doi.org/10.1016/j.procs.2021.05.025 (ScienceDirect)

[21] "Analysis of deep learning-based intrusion detection systems for improving IoT security," *Cybernetics and Systems*, May 2025. https://doi.org/10.1080/07366981.2025.2498222

[22] S. M. Kasongo and Y. Sun, "Performance Analysis of Intrusion Detection Systems Using a Feature Selection Method on the UNSW-NB15 Dataset," Journal of Big Data, 2020. https://doi.org/10.1186/s40537-020-00379-6 (Springer)

[23] N. B. Asha et al., "A Transfer Learning–Based Deep Convolutional Neural Network Framework for Automated Multi-Class Eye Disease Classification in the USA Using Retinal Fundus Image," Journal of Medical and Health Studies, 2023. https://doi.org/10.32996/jmhs.2023.4.4.24

[24] S. K. R. U. I. Rahat et al., "Deep Learning–Based Skin Cancer Diagnosis in the United States: Advances, Challenges, and Clinical Translation," Journal of Medical and Health Studies, 2023. https://doi.org/10.32996/jmhs.2023.4.6.18