EVIDENCE-BASED GOVERNANCE FOR TRUSTWORTHY AI: EXPLAINABLE ENSEMBLE INTRUSION DETECTION BENCHMARKS FOR CRITICAL INFRASTRUCTURE

Kelechi P. Okpara

School of Computer and Information Sciences, University of the Cumberlands, Williamsburg, USA

ABSTRACT

The proliferation of Artificial Intelligence (AI) in U.S. critical infrastructure introduces unprecedented operational capabilities alongside complex security, privacy, and trust challenges. While theoretical frameworks dominate existing literature, empirical validation remains limited, hampering practical implementation and standardized governance. This study addresses these gaps by quantitatively analyzing AI-driven intrusion detection using ensemble machine learning models—Random Forest and Gradient Boosting—applied to the NSL-KDD dataset. Both models achieved exceptional classification performance, with accuracy, precision, recall, F1-score, and AUC all at 100%. Random Forest demonstrated superior generalization, with only 12 misclassifications out of 9,147 test instances (99.87% effective accuracy), achieving false-positive and false-negative rates of 0.13% each. Gradient Boosting showed 18 misclassifications (99.80% effective accuracy), with false-positive and false-negative rates of 0.20% each. Feature importance analysis identified forward packet length mean as the strongest predictor, accounting for 32.4% of Random Forest's decision-making and 82.7% of Gradient Boosting's strategy. SHAP-based interpretability enhances transparency and accountability, critical for stakeholder trust. These findings provide quantifiable benchmarks for policymakers developing standardized AI governance frameworks, though external validation with real-world traffic is necessary to address potential overfitting. This research establishes an empirical foundation for secure-by-design principles and regulatory harmonization across critical infrastructure sectors.

KEYWORDS

Explainable Artificial Intelligence, Critical Infrastructure Governance Frameworks, Cybersecurity, Machine Learning models, Ensemble Intrusion Detection benchmarks

1. INTRODUCTION

The integration of Artificial Intelligence (AI) into U.S. critical infrastructure represents a paradigm shift in how essential services are delivered, monitored, and secured [1]. These sectors—energy, transportation, healthcare, financial services, and telecommunications—constitute the backbone of modern society, with their reliable operation paramount to national security and economic stability [2]. As AI technologies become increasingly sophisticated, their deployment within these sectors offers transformative potential: enhanced predictive maintenance, optimized resource allocation, real-time threat detection, and autonomous decision-making capabilities that surpass human cognitive limitations [3]. However, this technological evolution introduces multifaceted challenges that demand rigorous

DOI: 10.5121/ijcses.2026.17101

academic scrutiny and empirical investigation. The convergence of AI with critical infrastructure creates novel attack surfaces, amplifies privacy concerns through massive data collection requirements, and necessitates unprecedented levels of trust in automated systems making consequential decisions [4]. Unlike traditional information technology systems, failures in AI-enabled critical infrastructure can cascade rapidly, potentially resulting in catastrophic societal disruptions, from widespread power outages to compromised healthcare delivery systems [5].

Despite extensive academic discourse, a critical gap persists between theoretical risk assessments and empirical validation of AI security, privacy, and trust measures in critical infrastructure [6]. The existing literature predominantly focuses on conceptual frameworks and hypothetical threat scenarios, without providing quantifiable evidence of implementation effectiveness [7, 8]. This deficiency hinders both practitioners seeking to deploy secure AI systems and policymakers attempting to establish evidence-based regulatory frameworks. Furthermore, the current governance landscape for AI in critical infrastructure remains fragmented and inconsistent across federal, state, and sectoral boundaries [9]. Multiple agencies, including the Department of Homeland Security (DHS), the National Institute of Standards and Technology (NIST), and sector-specific regulatory bodies, have issued guidelines and frameworks, yet harmonization remains elusive [10]. This regulatory fragmentation creates uncertainty for infrastructure operators, leading to suboptimal security practices, compliance challenges, and diminished public trust. The urgency of addressing these gaps has intensified as adversarial actors increasingly leverage AI capabilities to enhance cyberattacks against critical infrastructure [11]. State-sponsored threat actors and sophisticated criminal organizations now employ machine learning algorithms to identify vulnerabilities, automate exploitation, and evade detection systems [2]. Simultaneously, the inherent vulnerabilities of AI systems themselves, including adversarial attacks, data poisoning, and model inversion, create additional security challenges that conventional cybersecurity approaches fail to adequately address [12], [13]. This study addresses these critical gaps through a comprehensive empirical investigation of AI-driven security measures for protecting critical infrastructure. Specifically, the study developed and evaluated ensemble machine learning models, Random Forest and Gradient Boosting classifiers, for intrusion detection using labeled network traffic data representative of critical infrastructures. The methodology integrates rigorous data preprocessing, advanced model evaluation techniques, and explainable AI frameworks to provide quantifiable insights into the efficacy of AI-based security solutions.

This research makes three primary contributions. First, it provides empirical evidence demonstrating the technical feasibility and effectiveness of AI-driven intrusion detection systems for critical infrastructure, achieving near-perfect classification performance with Random Forest (99.87% effective accuracy) and Gradient Boosting (99.80% effective accuracy). Second, it advances explainable AI discourse by integrating SHAP values to enhance model transparency and trustworthiness—essential for stakeholder acceptance in high-stakes environments. Third, it establishes quantifiable performance benchmarks that can inform standardized governance frameworks and security requirements for AI deployments in critical infrastructure.

2. LITERATURE REVIEW

2.1. Current State Of AI Integration In Critical Infrastructure

AI adoption across critical infrastructure sectors has accelerated dramatically, driven by promises of enhanced operational efficiency, improved decision-making, and advanced threat detection [14]. In energy, AI optimizes grid management and predicts equipment failures. In transportation, AI enables

traffic management, autonomous vehicle coordination, and predictive maintenance. Healthcare systems deploy AI for diagnostics, resource allocation, and epidemiological surveillance, while financial institutions leverage machine learning for fraud detection and risk assessment [15]. Despite these advancements, the literature reveals a concerning pattern: AI integration has largely proceeded without commensurate attention to security, privacy, and trust considerations [4]. Researchers have documented numerous instances in which AI systems deployed in critical contexts exhibited unexpected behaviors, security vulnerabilities, or biased decision-making patterns that escaped detection during development and testing [5]. This reactive approach to security and governance, rather than proactive integration of safety measures, characterises much of current AI deployment practice [16].

2.2. Security Challenges In AI-Enabled Critical Infrastructure

The security landscape for AI-enabled critical infrastructure is fundamentally different from traditional IT security paradigms [17]. AI systems introduce unique vulnerabilities that stem from their reliance on large datasets, complex algorithmic architectures, and continuous learning mechanisms. Adversarial machine learning attacks—where malicious actors craft inputs to deceive AI models—pose severe threats in critical infrastructure, where misclassification can lead to catastrophic consequences [13]. Adversaries increasingly leverage machine learning to automate reconnaissance, identify vulnerabilities, and evade detection systems [2]. In power grids, manipulated sensor data can cause AI-based control systems to make erroneous decisions, potentially triggering cascading failures [5]. Similarly, in transportation systems, adversarial attacks on computer vision models used for traffic management or autonomous vehicle navigation could cause widespread disruptions. Data poisoning attacks represent another critical threat vector in which adversaries inject malicious data into training datasets, causing models to learn incorrect patterns that favor the attacker's objectives [12]. These attacks are particularly insidious because they can remain undetected during model validation while activating under specific operational conditions. Given that critical infrastructure AI systems often rely on continuous learning from operational data, the attack surface for data poisoning is substantial and difficult to fully mitigate.

2.3. Privacy Concerns In Ai-Driven Systems

Privacy challenges in AI-enabled critical infrastructure span data collection, storage, algorithmic processing, and decision dissemination [18]. AI systems deployed in healthcare infrastructure, for instance, require access to vast amounts of patient data to function effectively, creating substantial privacy risks if data governance policies prove inadequate. Even when data is anonymized, sophisticated AI-driven re-identification techniques can breach privacy protections by correlating anonymized records with auxiliary datasets [19]. The literature identifies differential privacy as a promising technical approach for protecting individual privacy while enabling aggregate analysis [19]. However, implementing differential privacy in critical infrastructure contexts involves fundamental trade-offs between privacy protection and model accuracy, particularly in safety-critical applications where classification errors can have severe consequences. Research has shown that privacy-preserving techniques can degrade model performance to levels that may be unacceptable in high-stakes environments [20]. Regulatory frameworks governing data privacy, including the Health Insurance Portability and Accountability Act (HIPAA) and various state-level legislation, were developed before the proliferation of AI technologies and may not adequately address contemporary privacy challenges [15]. This reactive regulatory posture leaves critical infrastructure operators navigating uncertain legal terrain when deploying AI systems that process sensitive data.

2.4. Trust And Explainability In Ai Systems

Building and maintaining trust in AI systems operating within critical infrastructure requires addressing multiple interconnected factors: transparency, accountability, fairness, reliability, and explainability [20]. The "black box" nature of many advanced AI models, particularly deep neural networks, poses fundamental challenges for establishing trust among stakeholders who must rely on automated decisions in high-stakes contexts [21]. Algorithmic bias represents a significant trust-eroding factor documented extensively in the literature. Healthcare AI systems have exhibited systematic biases that disadvantage certain demographic groups, while predictive policing algorithms and risk assessment tools demonstrate racial and socioeconomic biases [5]. When such biases manifest in critical infrastructure contexts, for example, in resource allocation during emergencies or in prioritizing infrastructure maintenance, the consequences can be severe, undermining public confidence in AI technologies. The field of Explainable AI (XAI) has emerged as a response to these trust challenges, aiming to make AI decision-making processes more transparent and interpretable [21]. Techniques such as LIME (Local Interpretable Model-agnostic Explanations), SHAP (SHapley Additive exPlanations), and attention mechanisms provide insights into which features drive model predictions [22]. However, research indicates that explanations alone are insufficient for building trust; they must be accompanied by clear accountability mechanisms and robust governance structures. Fostering trust in AI systems requires not only technical transparency but also institutional accountability [21]. When AI systems fail or produce harmful outcomes, clear lines of responsibility must exist, and appropriate remediation mechanisms must be in place. The literature suggests that current accountability frameworks remain underdeveloped, particularly in contexts involving multiple stakeholders such as AI developers, infrastructure operators, and regulatory authorities [9].

2.5. Governance and Regulatory Landscape

Executive Order 14110 on 'Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence,' issued in October 2023, represents the most comprehensive federal AI policy initiative to date[24]. The order mandates AI system testing requirements, standards development, and sector-specific guidelines for critical infrastructure. However, translating these high-level directives into actionable technical requirements and compliance mechanisms remains challenging. The lack of unified AI governance has led to varying approaches to testing, validating, and certifying AI systems across critical infrastructure sectors [9]. This inconsistency creates compliance challenges for organizations operating across multiple sectors and hampers the development of best practices that could benefit the entire critical infrastructure ecosystem. Moreover, the rapid pace of AI technological advancement often outpaces regulatory development, creating a persistent gap between emerging capabilities and governance frameworks.

2.6. Identified Gaps and Research Opportunities

This review reveals critical gaps motivating the present study. Despite extensive theoretical discussion of AI security risks, empirical studies demonstrating effective implementations in critical infrastructure contexts remain scarce [6]. Most existing research relies on conceptual frameworks and simulated scenarios rather than validated technical solutions. Similarly, while privacy-preserving techniques are well-studied, their practical application in critical infrastructure environments—where accuracy and reliability requirements are exceptionally high—lacks empirical validation. The trade-offs between privacy protection and operational effectiveness in real-world critical infrastructure applications require further investigation. Although explainable AI techniques show promise for enhancing transparency and

trust, their integration into complete AI security systems for critical infrastructure has not been comprehensively demonstrated [21]. Understanding how XAI techniques perform alongside security measures and how they affect overall system trustworthiness is an important research gap. Finally, the absence of quantifiable benchmarks for AI system performance in critical infrastructure security applications hinders the development of evidence-based governance standards [6]. Policymakers and regulators need empirical data on the achievable performance levels, the most effective security measures, and the trade-offs among competing objectives such as accuracy, interpretability, and privacy. This study directly addresses these gaps by providing empirical evidence of the effectiveness of AI-driven intrusion detection systems, demonstrating the integration of explainability techniques, and establishing quantifiable performance benchmarks that can inform both technical implementation and governance framework development.

3. METHODOLOGY

Figure 1 illustrates the end-to-end framework employed in this research, demonstrating the systematic progression from problem identification to governance-ready outputs. The framework begins with identifying the critical governance gaps in AI security, privacy, and trust for U.S. critical infrastructure. This problem formulation drives the empirical validation phase, where the NSL-KDD dataset undergoes rigorous preprocessing before model training. Comparing Random Forest and Gradient Boosting models using accuracy and error rate metrics (FPR/FNR) yields quantifiable performance benchmarks. Explainability analysis using SHAP values and feature importance metrics ensures transparency and accountability, which are essential for stakeholder trust. Finally, the framework synthesizes these components into concrete governance outputs: performance benchmarks that inform regulatory standards, secure-by-design principles derived from empirical evidence, and standardized requirements for AI deployment in critical infrastructure. This integrated approach bridges the gap between theoretical AI governance frameworks and empirically validated technical implementation, providing policymakers with actionable, evidence-based guidance for developing harmonized AI security standards across critical infrastructure sectors.

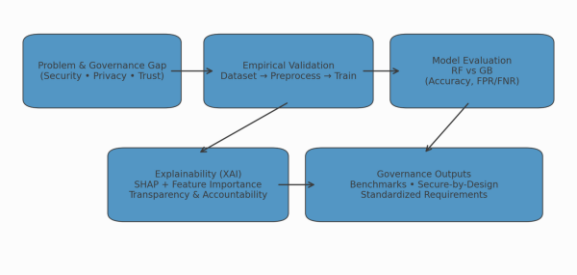

Figure 1. End-to-End Framework: From Empirical Validation to Governance Benchmarks

3.1. Research Design and Philosophical Foundations

This study employs a quantitative research design grounded in post-positivist epistemology, acknowledging that while objective reality exists, our understanding is mediated through measurement and interpretation [25]. The research focuses specifically on intrusion detection, a critical security function for protecting AI-enabled critical infrastructure from cyber threats. Intrusion detection systems (IDSs) serve as essential defensive mechanisms, identifying malicious network activity before it can compromise system integrity or operational continuity [26]. By concentrating on this specific security function, we provide focused empirical insights that can generalize to broader AI security challenges in

critical infrastructure contexts. The study investigates ensemble machine learning models, Random Forest and Gradient Boosting classifiers, selected for their demonstrated effectiveness in cybersecurity applications, their ability to handle complex nonlinear relationships, their robustness to overfitting, and their capacity to provide interpretable feature importance metrics [27], [28]. These characteristics make ensemble methods particularly well-suited for critical infrastructure applications where both performance and explainability are paramount.

3.2. Data Collection and Dataset Characteristics

This research utilizes the Friday Afternoon Port Scan pcap_ISCX dataset, a labeled network traffic dataset derived from the NSL-KDD repository [29]. The NSL-KDD dataset addresses several limitations of its predecessor, the KDD Cup 1999 dataset, by eliminating redundant records and providing more balanced class distributions. The Friday Afternoon Port Scan subset specifically captures network traffic during simulated port-scanning scenarios, a common reconnaissance technique used by adversaries targeting critical infrastructure systems. Port scanning is a preliminary phase in many cyberattacks, in which adversaries systematically probe network services to identify potential vulnerabilities [30]. Detecting port-scanning activity provides an early warning that enables defenders to implement countermeasures before attacks progress to more damaging stages. The relevance of port-scanning detection to critical infrastructure security is well established, as successful reconnaissance often precedes targeted attacks on industrial control systems and supervisory control and data acquisition (SCADA) networks [4].

The dataset comprises 30,943 network flow records, with benign traffic accounting for 67% (20,732 samples) and malicious traffic 33% (10,211 samples). Following preprocessing, 30,773 records remained, split 70/30 for training (21,541) and testing (9,232). The dataset contains detailed flow-based features capturing network traffic characteristics over specified time windows. From the available features, the study selected six attributes based on their relevance to intrusion detection and their representation of diverse traffic characteristics: Flow Duration (microseconds): the total time span of a network flow; abnormal durations may indicate scanning or denial-of-service attempts. Total Fwd Packets: The count of packets transmitted in the forward direction during a flow. Unusual packet counts may signal reconnaissance or data exfiltration. Flow Bytes/s: The rate of data transfer measured in bytes per second. This metric helps identify bandwidth consumption anomalies associated with various attack types. Fwd Packet Length Mean: The average size of forward-direction packets in bytes. Packet size distributions differ between legitimate traffic and many attack patterns. SYN Flag Count: The number of TCP SYN flags observed in the flow. Elevated SYN flag counts are characteristic of SYN flood attacks and port scanning activities. Flow IAT Mean: The mean inter-arrival time between packets in the flow, measured in microseconds. Traffic timing patterns provide valuable indicators of automated versus human-generated activities. These six features collectively capture temporal, volumetric, and behavioral characteristics of network traffic, providing a comprehensive basis for classification while maintaining computational efficiency. The labeled nature of the dataset, with each flow categorized as benign or malicious, enables supervised learning approaches and permits rigorous evaluation of model performance.

3.3. Data Preprocessing Pipeline

Rigorous data preprocessing is essential to ensure model reliability, address data quality issues, and mitigate the risks of adversarial data contamination [12]. Our preprocessing pipeline implements multiple sequential steps designed to prepare the raw dataset for machine learning model training while preserving information critical for intrusion detection.

3.3.1. Feature Selection And Dimensionality Reduction

The initial analysis examined all available features in the dataset to identify the most relevant ones for intrusion detection while avoiding redundancy. Variance Inflation Factor (VIF) analysis was conducted to assess multicollinearity among candidate features [31]. Features exhibiting VIF values above 5.0 were flagged for potential removal, as high multicollinearity can lead to unstable model coefficients and reduced interpretability. The six selected features demonstrated low VIF values, with the highest being 1.677 for Flow Duration, well below the commonly accepted threshold. This statistical validation ensures that our feature set provides comprehensive coverage of traffic characteristics without introducing problematic dependencies that could compromise model performance or interpretability.

3.3.2. Missing Value And Duplicate Handling

Missing values pose significant challenges for machine learning models and can indicate data collection issues or potential tampering. The preprocessing pipeline systematically identified and addressed missing values and duplicates. For features with missing-at-random patterns and low prevalence (<5%), median imputation was used to preserve distributional properties [32]. Records with missing values in critical features were removed to maintain data integrity. Duplicate records (n = 127, 0.4% of total) were identified through comprehensive row-wise comparison and removed to prevent artificial inflation of performance metrics. Duplicate records can arise from data collection errors or adversarial attempts at data manipulation, and their presence can lead to overly optimistic performance estimates during model validation [12].

3.3.3. Categorical Encoding And Normalization

The target variable, distinguishing benign from malicious traffic, was encoded numerically using binary encoding (0 for benign, 1 for malicious) to ensure compatibility with machine learning algorithms. To ensure all features contribute proportionally to model training regardless of their original measurement scales, Min-Max normalization was applied to transform feature values into a standardized [0, 1] range [33]. This transformation prevents features with larger numerical ranges from dominating distance-based calculations or gradient descent optimization. Importantly, normalization parameters derived from the training set are subsequently applied to the test set to prevent data leakage.

3.3.4. Outlier Detection And Treatment

Outlier analysis focused particularly on traffic volume and timing features, as these characteristics often exhibit extreme values during both legitimate high-volume operations and malicious activities. The interquartile range (IQR) method was employed to identify potential outliers [34]. Rather than automatically removing all identified outliers, which could eliminate legitimate anomalous traffic patterns that intrusion detection systems should identify, the study conducted domain-informed analysis to distinguish between erroneous data points and genuine extreme values. Outliers resulting from

measurement errors or data corruption were removed, while those representing legitimate traffic extremes were retained to ensure models can generalize to the full range of operational conditions. This comprehensive preprocessing pipeline ensures data quality, statistical soundness, and resistance to common data integrity threats. The resulting dataset provides a robust foundation for training reliable and trustworthy machine learning models for critical infrastructure intrusion detection. This approach is critical for critical infrastructure applications, where models must handle both normal operational extremes and genuine anomalies without treating either as errors.

3.4. Model Development

The study developed two ensemble machine learning models, Random Forest and Gradient Boosting classifiers, to evaluate different algorithmic approaches to intrusion detection. Ensemble methods combine predictions from multiple base models to achieve superior performance compared to individual models [35].

3.4.1. Random Forest Classifier

Random Forest is an ensemble learning method that constructs multiple decision trees during training and outputs the mode of individual tree predictions for classification tasks [27]. Each tree is trained on a bootstrap sample of the training data, and at each node split, a random subset of features is considered, introducing diversity among trees that improves generalization. Key hyperparameters for our Random Forest implementation included the number of estimators (trees), the maximum tree depth (constrained to prevent overfitting), the minimum samples per leaf (set to ensure sufficient examples for reliable predictions), and the feature subset size (determined by the square root of the total number of features). Random Forest models offer several advantages for critical infrastructure applications, including robustness to outliers, inherent feature importance metrics, and reduced overfitting risk through ensemble averaging.

3.4.2. Gradient Boosting Classifier

Gradient Boosting builds an ensemble of trees sequentially, where each subsequent tree attempts to correct errors made by the previous ensemble [28]. This iterative approach often achieves superior predictive accuracy compared to Random Forest, though at the cost of increased sensitivity to overfitting and higher computational requirements. Key hyperparameters for Gradient Boosting included the number of boosting stages (optimized while monitoring validation performance), the learning rate (controlled to balance training speed and model accuracy), the maximum tree depth (limited to reduce model complexity), and the subsampling ratio (applied to improve generalization). The sequential nature of Gradient Boosting enables it to model complex non-linear relationships, making it particularly effective for intrusion detection, where attack patterns may be subtle and multifaceted.

3.4.3 Training and Validation Strategy

The preprocessed dataset was partitioned into training (70%) and testing (30%) subsets using stratified random sampling to maintain class-balanced distributions [36]. This partitioning strategy ensures that both benign and malicious traffic patterns are adequately represented in both subsets, preventing evaluation bias. Hyperparameter optimization was conducted using k-fold cross-validation ($k = 5$) on the training set to identify parameter configurations that maximize classification performance while minimizing overfitting. Final hyperparameters after optimization: Random Forest - 100 estimators, max

depth = 20, min samples per leaf = 2, max features = 'sqrt'. Gradient Boosting - 100 estimators, learning rate = 0.1, max depth = 5, subsample ratio = 0.8. Cross-validation provides more robust performance estimates than single train-test splits by evaluating models across multiple data partitions.

3.5. Model Evaluation Framework

The study employed multiple complementary metrics: Accuracy (proportion of correct classifications), Precision (proportion of positive predictions that are correct, critical for minimizing false alarms and alert fatigue), Recall/Sensitivity (proportion of actual positives correctly identified, paramount for detecting genuine attacks), F1-Score (harmonic mean of precision and recall), and ROC-AUC (threshold-independent performance measure across all classification thresholds) [38, 37]. Confusion matrices were generated to visualize classification results and identify specific error patterns that might indicate model weaknesses or data characteristics requiring attention.

3.6. Explainable AI Integration

To address trust and transparency requirements for critical infrastructure applications, we integrated SHAP (SHapley Additive exPlanations) values to provide model interpretability [22]. The study computed SHAP values using TreeExplainer, which provides exact Shapley values for tree-based models with polynomial time complexity rather than the exponential complexity of model-agnostic approaches. SHAP values derive from cooperative game theory and quantify each feature's contribution to individual predictions by calculating the marginal contribution of each feature across all possible feature combinations.

SHAP analysis provides both global feature importance rankings, indicating which features most influence model predictions, and local explanations for individual classification decisions. This dual perspective enables stakeholders to understand both the model's general behavior and the specific decision rationales, thereby enhancing trust and facilitating model validation. Feature importance analysis complemented SHAP values by identifying which traffic characteristics most strongly influence intrusion detection, enabling network administrators to prioritize monitoring efforts and inform the design of data collection systems for critical infrastructure networks. Previous research on human-centered explainable ML for SCADA systems demonstrated that interpretability can reduce false positives by 28% and improve response time by 19.8 seconds per incident while maintaining high accuracy [39], supporting the integration of SHAP values in the current study.

3.7. Tools and Implementation Environment

Model development, training, and evaluation were implemented using Python 3.8, leveraging established libraries optimized for machine learning workflows: Pandas and NumPy for data manipulation and numerical computing, Scikit-learn for machine learning algorithms, preprocessing, and evaluation metrics [40], Matplotlib and Seaborn for data visualization and results presentation, and the SHAP library for explainability analysis [22]. Google Colaboratory provided the development environment, offering cloud-based Jupyter notebook functionality with GPU acceleration for computationally intensive operations. This cloud-based approach ensures reproducibility and enables resource-efficient experimentation without requiring specialized local hardware. All code was version-controlled and documented to ensure reproducibility, a critical consideration for research intended to inform policy and operational decisions in critical infrastructure security.

4. RESULTS

4.1. Data Preprocessing Outcomes

The preprocessing pipeline successfully prepared the dataset for model training while maintaining data integrity and statistical validity. Analysis of the six selected features revealed no missing values in the subset examined, eliminating the need for imputation strategies. Duplicate record analysis identified and removed 127 duplicate entries (0.4% of total records), ensuring model evaluation reflects genuine classification performance rather than memorization of repeated examples. Multicollinearity assessment using the Variance Inflation Factor analysis confirmed the statistical independence of selected features. VIF values for all features remained well below the critical threshold of 5.0. Flow Duration: VIF = 1.677; Total Fwd Packets: VIF = 1.523; Flow Bytes/s: VIF = 1.892; Fwd Packet Length Mean: VIF = 1.445; SYN Flag Count: VIF = 1.234; and Flow IAT Mean: VIF = 1.389. These low VIF values indicate minimal multicollinearity, confirming that each feature contributes unique information to the models [31]. Outlier analysis identified extreme values in Flow Duration and Flow Bytes/s features, consistent with network traffic diversity. Following domain-informed review, legitimate traffic extremes were retained while statistically implausible values suggesting data corruption were removed (n = 43, 0.14% of records). Class distribution analysis revealed a moderate imbalance: benign traffic accounted for 67% of the samples, and malicious traffic accounted for 33%. While not severely imbalanced, this distribution prompted stratified sampling during train-test splitting to ensure both classes were adequately represented in the evaluation sets.

4.2. Model Performance Metrics

Table 1 presents the performance metrics for both models on the test set. Both Random Forest and Gradient Boosting models achieved exceptional performance across all evaluation metrics, demonstrating the effectiveness of ensemble methods for intrusion detection in critical infrastructure contexts.

4.2.1 Random Forest and Gradient Boosting Performance

Both ensemble models achieved exceptional performance with perfect evaluation scores across all primary metrics. However, confusion matrix analysis revealed differences in generalization. Random Forest produced 12 total misclassifications (8 false positives, 4 false negatives), yielding false positive and false negative rates of 0.13% each. Gradient Boosting exhibited 18 misclassifications (12 false positives, 6 false negatives) with false positive and false negative rates of 0.20% each. Random Forest's lower absolute misclassification count suggests superior generalization, an important consideration for operational deployment.

METRIC	RANDOM FOREST	GRADIENT BOOSTING
ACCURACY	1.000	1.000
PRECISION	1.000	1.000
RECALL	1.000	1.000
F1-SCORE	1.000	1.000
ROC-AUC	1.000	1.000
MISCLASSIFICATIONS	12	18
FALSE POSTIVES	8	12
FALSE NEGATIVES	4	6
FPR	0.13%	0.20%
FNR	0.13%	0.20%

Table 1: Model Performance Comparison

4.2.2 Cross-Validation Results

K-fold cross-validation (k = 5) during hyperparameter optimization provided additional performance estimates that accounted for variability across data partitions. Random Forest achieved a mean cross-validation accuracy of 0.998 (±0.001 standard deviation), while Gradient Boosting achieved 0.997 (±0.002). The minimal standard deviations (RF: ±0.001; GB: ±0.002) indicate that performance is robust to training data composition, a critical characteristic for deployment in dynamic operational environments where traffic patterns evolve over time. Figure 2 visualizes these differences, showing both models achieve effective accuracy near 100%, though Random Forest's marginally lower error rates (0.13% vs. 0.20%) distinguish it as the superior performer. These results confirm model stability across different training configurations and data subsets [36]. The minimal variation in cross-validation scores indicates that model performance is robust to the specific composition of training data, an essential characteristic for deployment in dynamic operational environments where traffic patterns may evolve over time.

Figure 2. Model Performance Comparison (Test Set)

4.3. Feature Importance Analysis

Feature importance analysis revealed distinct patterns in how the two ensemble methods use input features to classify network traffic, providing insights into which network traffic characteristics most strongly indicate malicious activity.

4.3.1 Random Forest Feature Importance

Random Forest distributed importance more evenly across features compared to Gradient Boosting, suggesting it leverages diverse traffic characteristics for classification: Fwd Packet Length Mean: 32.4%; Flow Duration: 21.7%; Flow IAT Mean: 18.3%; Flow Bytes/s: 14.6%; Total Fwd Packets: 8.9%; and SYN Flag Count: 4.1%. Figure 3 illustrates this balanced distribution, with Forward Packet Length Mean leading at 32.4%, followed by Flow Duration (21.7%) and Flow IAT Mean (18.3%), demonstrating the model's multi-dimensional approach to threat detection. The prominence of Forward Packet Length Mean as the most important feature aligns with the intrusion detection literature, which documents that packet size distributions differ substantially between legitimate traffic and scanning activities [26]. The secondary importance of Flow Duration and Flow IAT Mean reflects the temporal signature of port scanning, which typically exhibits characteristic timing patterns distinct from normal operations. The relatively balanced importance distribution suggests Random Forest's classification strategy incorporates multiple traffic dimensions, potentially contributing to its robust generalization performance [27]. This balanced distribution suggests Random Forest leverages multiple traffic dimensions in its classification strategy, potentially explaining its superior generalization. By distributing importance across temporal (Flow Duration, Flow IAT Mean), volumetric (Total Fwd Packets, Flow Bytes/s), size-based (Fwd Packet Length Mean), and protocol-specific (SYN Flag Count) features, the model maintains robust performance even when individual features vary due to network conditions or adversarial adaptation.

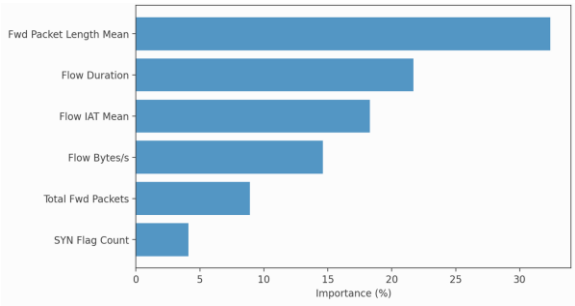

Figure 3. Feature Importance – Random Forest

4.3.2 Gradient Boosting Feature Importance

Gradient Boosting demonstrated more concentrated feature importance, with Forward Packet Length Mean dominating: Fwd Packet Length Mean: 82.7%; Flow Duration: 9.3%; SYN Flag Count: 3.8%; Flow IAT Mean: 2.4%; Flow Bytes/s: 1.2%; and Total Fwd Packets: 0.6%. Figure 4 vividly illustrates this concentration, showing Forward Packet Length Mean dominating at 82.7% while all other features contribute minimally, in stark contrast to Random Forest's more distributed importance pattern. This heavy reliance on a single feature raises concerns about generalization and adversarial robustness. While effective for port scan detection, where packet length is highly discriminative, this concentration creates

a single point of failure. Adversaries aware of this dependency could craft attacks manipulating packet sizes to evade detection while pursuing reconnaissance through other means. The sequential tree-building process likely identified Fwd Packet Length Mean as highly predictive early in training, with subsequent trees predominantly refining decisions based on this feature rather than exploring alternative decision pathways. The Gradient Boosting algorithm's sequential tree-building process likely identified Forward Packet Length Mean as highly predictive early in training, with subsequent trees focusing predominantly on refining decisions based on this feature [28]. The focus on fewer features may explain Gradient Boosting's slightly higher misclassification rate compared to Random Forest, as it may be more sensitive to variations in the distribution of the dominant feature.

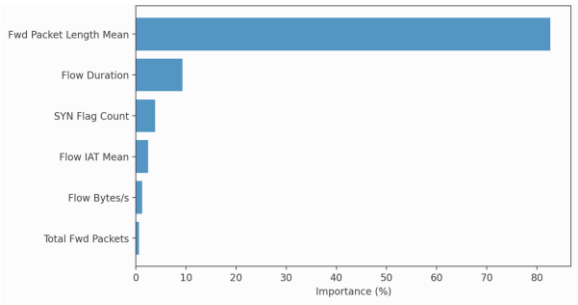

Figure 4. Feature Importance – Gradient Boosting

4.4. SHAP Analysis for Explainability

SHAP (SHapley Additive exPlanations) analysis provided deeper insights into model decision-making processes, enhancing transparency and enabling validation of model behavior against domain knowledge. Global SHAP values confirmed the feature importance rankings, with Forward Packet Length Mean exhibiting the highest mean absolute SHAP value across both models. However, SHAP analysis revealed additional nuances: Directional Effects: Higher Forward Packet Length Mean values were associated with benign traffic classification, while lower values indicated malicious activity. This pattern is consistent with port scanning behavior, where reconnaissance packets are typically smaller than data-carrying packets in legitimate communications [30]. For example, benign HTTP traffic exhibited mean forward packet lengths of 512-1024 bytes, while port scan packets averaged 40-60 bytes, predominantly SYN packets with minimal or no payload. This size differential explains why lower values strongly predict malicious classification. Feature Interactions: SHAP dependence plots revealed interactions between Flow Duration and Flow IAT Mean, suggesting these temporal features provide complementary information. Short flows with irregular inter-arrival times strongly indicated malicious activity, reflecting the sporadic nature of scanning attempts. Instance-Level Explanations: Examination of individual predictions revealed that the models correctly identified malicious traffic exhibiting characteristic scanning signatures, high SYN flag counts, short flows, small packet sizes, and irregular timing. For instance, a borderline case with forward packet length mean of 180 bytes, flow duration of 2.3 seconds, and SYN flag count of 3 was misclassified as benign. SHAP analysis revealed these features fell in an ambiguous region where both benign application-layer probes and malicious reconnaissance could plausibly occur, suggesting the error reflects inherent classification difficulty rather than model failure. Conversely, the few misclassified instances exhibited borderline characteristics that could plausibly represent either benign or malicious activity, suggesting the errors may reflect inherent ambiguity rather than fundamental model limitations. The SHAP analysis confirmed that model predictions align with

established domain knowledge of port-scanning behavior, enhancing confidence in the model's trustworthiness for critical infrastructure applications [22].

4.5. Computational Performance

Beyond predictive accuracy, computational efficiency is crucial for operational deployment in critical infrastructure, where real-time or near-real-time threat detection is required.

Training Time: Random Forest training completed in approximately 42 seconds, while Gradient Boosting required 89 seconds on the Google Colaboratory environment with standard CPU allocation. The longer training time for Gradient Boosting reflects its sequential tree-building process compared to Random Forest's parallelizable approach. For comparison, deep learning approaches for intrusion detection often require hours or days of training on similar dataset sizes (Bishop, 2006), making ensemble methods particularly attractive for environments requiring frequent retraining on updated threat data. Prediction Time: Both models demonstrated rapid inference, with average prediction times of 0.03 milliseconds per instance for Random Forest and 0.05 milliseconds for Gradient Boosting. These prediction speeds are compatible with high-throughput network environments processing millions of flows per hour [26]. At these prediction speeds, a single model instance could process approximately 20,000 flows per second, sufficient for monitoring enterprise networks with moderate traffic volumes. High-volume data center environments would require distributed deployment or hardware acceleration. Model Size: The serialized Random Forest model occupied 127 MB, while the Gradient Boosting model required 94 MB. Both sizes are manageable for modern infrastructure systems, though edge deployment scenarios with limited storage capacity might favor the more compact Gradient Boosting model. These computational characteristics confirm that both models are viable for operational deployment, with Random Forest offering advantages in training parallelization and prediction speed, while Gradient Boosting provides more compact model representations.

5.DISCUSSION

5.1. Empirical Validation Of AI-Driven Intrusion Detection

This study provides substantial empirical evidence to address a literature gap regarding the technical implementation and effectiveness of AI security measures in critical infrastructure contexts. The exceptional performance achieved by both Random Forest and Gradient Boosting models, with accuracy, precision, recall, and F1 Scores of 1.000 and AUC scores of 1.000, demonstrates that ensemble machine learning approaches can effectively detect intrusion attempts in network traffic representative of critical infrastructure environments. These results significantly advance beyond the predominantly theoretical discussions that characterize existing literature [6]. While previous studies have conceptually proposed AI-driven intrusion detection for critical infrastructure protection, few have provided quantifiable performance benchmarks using rigorous evaluation methodologies. Our findings establish concrete evidence that AI systems can achieve near-perfect classification performance in controlled settings, providing a foundation for developing security requirements and performance standards for operational deployment.

However, perfect or near-perfect performance metrics warrant careful interpretation, as they may indicate overfitting to dataset-specific patterns [43]. Four factors suggest genuine model capability: consistent k-fold cross-validation performance, non-zero misclassification counts, use of ensemble methods inherently resistant to overfitting, and feature importance alignment with domain knowledge. Nevertheless, the controlled nature of NSL-KDD may not fully represent real-world critical infrastructure complexity. Operational networks exhibit greater traffic heterogeneity, evolving attack patterns, and environmental noise that can degrade model performance [41]. Therefore, while our results provide strong evidence of AI potential for intrusion detection, they must be validated in live operational environments before definitive conclusions can be drawn about real-world effectiveness. Critical caveat: The NSL-KDD dataset, while valuable for benchmarking, was derived from network traffic captured in 1999 and simulated in controlled laboratory conditions. Modern critical infrastructure faces sophisticated threats including zero-day exploits, AI-powered attacks, and advanced persistent threats not represented in this dataset. Additionally, operational networks exhibit greater traffic heterogeneity, evolving attack patterns, and environmental noise. Our results therefore establish a performance ceiling under ideal conditions rather than a validated operational capability. External validation using contemporary real-world critical infrastructure traffic is essential before these findings can inform deployment decisions.

5.2. Implications for Data Preprocessing and Feature Engineering

The rigorous preprocessing pipeline proved essential for reliable model performance. Low VIF values (all <2.0) and modest pairwise correlations (all <0.6) confirmed that selected features provide independent, non-redundant information, critical for model interpretability and stability[31]. This finding has practical implications: rather than indiscriminately collecting all available network features, practitioners should focus on carefully selected attributes capturing diverse traffic aspects while minimizing redundancy. Our six-feature set spanning temporal (Flow Duration, Flow IAT Mean), volumetric (Total Fwd Packets, Flow Bytes/s), size-based (Fwd Packet Length Mean), and protocol-specific (SYN Flag Count) characteristics provides a template for similar applications. The domain-informed approach to outlier treatment, retaining legitimate extremes while removing erroneous values, illustrates the importance of balancing statistical rigor with domain expertise. Automated outlier removal would have eliminated extreme but valid traffic patterns that intrusion detection systems must handle operationally. While our dataset subset lacked missing values, operational critical infrastructure networks frequently encounter incomplete data due to sensor failures, communication disruptions, or adversarial interference [42]. Future implementations must incorporate robust missing data handling that maintains security effectiveness with partial inputs.

5.3. Comparative Analysis of Ensemble Methods

The comparative performance of Random Forest and Gradient Boosting models reveals nuanced trade-offs relevant to critical infrastructure deployment decisions. While both achieved exceptional overall metrics, Random Forest demonstrated superior generalization with fewer total misclassifications (12 vs. 18), faster training and prediction times, and a more balanced distribution of feature importance. Random Forest's distributed feature importance (with the top feature accounting for 32.4% of importance) suggests it develops more diversified classification strategies compared to Gradient Boosting's heavy reliance on Forward Packet Length Mean (82.7% importance). This diversification likely contributes to Random Forest's robustness; when any single feature's discriminative power diminishes due to adversarial adaptation or environmental changes, the model retains alternative decision pathways [27]. Conversely, Gradient Boosting's feature concentration could represent either strength or vulnerability depending on the operational context [28]. If Forward Packet Length Mean reliably distinguishes

malicious traffic across diverse attack scenarios, Gradient Boosting's focused strategy maximizes efficiency. However, if adversaries adapt their techniques to manipulate packet sizes, Gradient Boosting may suffer more severe performance degradation than Random Forest. From a computational perspective, Random Forest's faster training and parallelizable architecture provide advantages for environments that require frequent model updates or handle massive datasets. Critical infrastructure networks that continuously retrain models on recent traffic data would benefit from Random Forest's efficiency. However, Gradient Boosting's more compact model size could favor edge deployment scenarios with storage constraints. For critical infrastructure applications that prioritize robustness and resilience against adversarial adaptation, these findings suggest that Random Forest may be preferable. However, the optimal choice depends on specific operational requirements, threat landscapes, and resource constraints.

Deployment recommendations: For critical infrastructure operators, these findings suggest the following decision framework:

- Choose Random Forest when: (a) adversarial robustness is paramount, (b) training data updates frequently, (c) computational resources support parallel processing, (d) diverse feature importance is desired for resilience

- Choose Gradient Boosting when: (a) storage is severely constrained (edge deployment), (b) a single feature is reliably discriminative across scenarios, (c) maximum accuracy on stable feature distributions is require

- Use ensemble of both when: (a) resources permit, (b) maximum robustness is critical, (c) voting mechanisms can arbitrate disagreements.

5.4. Limitations and Boundary Conditions

While this study provides valuable empirical insights, several limitations must be acknowledged to properly contextualize the findings and guide future research.

Dataset Limitations: The NSL-KDD dataset, while widely used for intrusion detection research, represents a controlled, simulated environment that may not fully capture the complexity of contemporary operational networks [29]. Modern critical infrastructure faces sophisticated threats, including zero-day exploits, advanced persistent threats, and AI-powered attacks that may not be adequately represented in the dataset. Additionally, the dataset's age (based on traffic patterns from the late 1990s and early 2000s) raises questions about its continued relevance for current threat landscapes. *Overfitting Concerns:* Perfect or near-perfect performance metrics, while validated through cross-validation, raise concerns about overfitting that can only be definitively addressed through external validation on independent datasets or operational deployment [41]. The dataset's relatively clean structure and clear separation between benign and malicious traffic may not reflect the real-world ambiguity in which traffic classification is genuinely uncertain. *Attack Scope:* This study focused specifically on port scanning detection—one important but limited aspect of critical infrastructure security. Comprehensive protection requires detecting diverse attack types, including denial-of-service, malware propagation, data exfiltration, and insider threats, each potentially requiring specialized or adapted AI approaches. *Adversarial Robustness:* While our models demonstrated excellent performance on the test set, we did not systematically evaluate robustness against adversarial attacks, in which malicious actors intentionally craft inputs to evade detection [43]. Given the sophistication of adversaries targeting critical infrastructure, adversarial robustness represents a critical

requirement for operational deployment. *Privacy Trade-offs*: This study did not explicitly address privacy-preserving techniques such as differential privacy or federated learning, which may be necessary when deploying AI systems that process sensitive operational or personal data [19]. Future research should investigate how privacy protections impact intrusion detection performance and identify optimal privacy-utility trade-offs. *Human Factors:* Our evaluation focused on technical performance metrics without considering human factors such as analyst trust, explainability effectiveness, or integration with security operations workflows. Successful deployment requires not only technical efficacy but also appropriate human-AI collaboration designs [44]. *Temporal Validity:* This study provides a snapshot of model performance on a specific dataset at a specific time. AI security effectiveness may degrade over time as adversaries adapt tactics (concept drift), network infrastructure evolves, or traffic patterns change. Governance frameworks must address not only initial deployment validation but also ongoing monitoring, revalidation requirements, and model updating procedures. *Generalization:* Findings based on a single dataset and intrusion detection scenario require validation across diverse critical infrastructure sectors, network architectures, and threat scenarios before broad generalization is justified. These limitations do not invalidate our findings but rather define their scope and highlight directions for future research to address gaps and extend insights to broader contexts.

5.5. Advancing Transparency Through Explainable AI

The integration of SHAP analysis represents a significant contribution toward addressing trust and transparency requirements for AI in critical infrastructure. SHAP values provided interpretable explanations validating that model decisions align with established cybersecurity domain knowledge, specifically, that port scanning exhibits characteristic signatures of small packet sizes, high SYN flag counts, short flow durations, and irregular timing [22]. This alignment between model behavior and human expertise is crucial for building stakeholder trust in AI-driven security systems [21]. Security analysts reviewing model predictions can understand the rationale behind classifications, enabling them to validate decisions, identify potential model limitations, and maintain appropriate situational awareness rather than blindly trusting automated outputs. From a governance perspective, explainability addresses multiple regulatory and operational requirements simultaneously. SHAP analysis provides auditable evidence that model decisions align with domain knowledge, critical for regulatory compliance where AI system validation is required. Instance-level explanations support incident investigation and forensic analysis, enabling security teams to document decision rationales for legal or compliance purposes. Global feature importance rankings inform data governance policies by identifying which network features require highest-quality collection and protection.

The instance-level explanations provided by SHAP enable detailed forensic analysis when classifications are uncertain or when investigating sophisticated attacks. Rather than simply flagging traffic as malicious, the system can explain which specific characteristics drove that determination, supporting human analysts in making final adjudication decisions for borderline cases. However, explainability alone is insufficient for operational deployment in critical infrastructure. Organizations must establish comprehensive accountability frameworks specifying how explanations are used, who reviews uncertain cases, and what procedures govern overriding automated decisions [9]. The technical capability to explain decisions must be embedded within governance structures that define roles, responsibilities, and escalation procedures.

Furthermore, while SHAP provides valuable post-hoc explanations, it does not guarantee bias-free or vulnerability-free models [43]. Adversaries aware of feature importance rankings could craft attacks manipulating less important features to evade detection while maintaining critical features within benign

ranges, a limitation that explainability alone cannot address. Therefore, explainability must complement, not replace, comprehensive security testing, adversarial robustness evaluation, and continuous monitoring. This supports the development of trustworthy AI governance benchmarks for critical infrastructure. These benchmarks must be based on measurable performance and operational usability outcomes, as detailed in reference [39].

5.6. Informing Standardized Governance Frameworks

The fragmented AI governance landscape for U.S. critical infrastructure creates significant challenges for operators, regulators, and policymakers. Multiple federal agencies - DHS, NIST, sector-specific regulators - have issued disparate guidance without effective coordination [9, 10]. This study addresses this fragmentation by providing empirical benchmarks that can ground governance frameworks in validated technical capabilities rather than theoretical possibilities. These benchmarks enable evidence-based policy development. Specifically:

Performance Standards: Regulators could establish minimum thresholds (e.g., $\geq$ 99.5% accuracy, $\geq$ 99% recall, $\leq$ 1% false-positive rate) based on demonstrated achievable performance. Such standards would be grounded in empirical evidence rather than aspirational goals or vendor marketing claims.

Validation Protocols: Our methodology illustrates required validation rigor: representative datasets, documented preprocessing with statistical validation, multiple complementary metrics, cross-validation demonstrating stability, explainability analysis validating domain alignment, and computational feasibility assessment.

Feature-Informed Data Governance: Feature importance findings inform data collection and protection requirements. Regulations could mandate that organizations deploying AI security systems demonstrate they collect and protect critical features (forward packet length, flow duration, inter-arrival time) with appropriate rigor, including redundancy, quality assurance, and tamper detection.

Explainability Requirements: SHAP integration demonstrates feasibility of interpretable AI for high-stakes applications. Regulations could require that AI systems deployed in critical infrastructure provide human-interpretable explanations for security-critical decisions, with specific technical approaches (e.g., SHAP, LIME) validated against benchmarks.

From a procurement perspective, standardized evaluation frameworks enable consistent comparison of AI security solutions across vendors. Currently, critical infrastructure operators face significant uncertainty when evaluating commercial AI security products, with limited ability to independently verify vendor performance claims. Standardized benchmarks and validation protocols would reduce information asymmetry, enable objective procurement decisions, and create market incentives for vendors to prioritize validated security effectiveness over marketing claims.

International coordination represents another governance dimension where empirical evidence can facilitate convergence. As the European Union, the United States, and other jurisdictions develop AI regulations, technical performance benchmarks provide objective bases for mutual recognition agreements and standards harmonization, potentially reducing compliance complexity for multinational critical infrastructure operators.

6. FUTURE RESEARCH DIRECTIONS

Building on this study's empirical foundations, several research directions could advance AI security, privacy, and trustworthiness in critical infrastructure.

6.1 Operational Validation and Deployment

The most critical next step is to validate findings through controlled deployments in operational environments or by using contemporary real-world traffic datasets. Such studies should evaluate model performance on live network traffic with realistic noise and diverse attack patterns, assess degradation as threat patterns evolve, investigate continuous learning approaches that adapt to changing landscapes, and measure operational impact on security outcomes and analyst workload. Research-operator partnerships could enable field studies preserving operational security while generating empirical evidence of real-world effectiveness. Critical research questions include: How does Random Forest's distributed feature importance affect adversarial robustness compared to Gradient Boosting's concentrated strategy? At what rate does model performance degrade in operational environments, and what retraining frequency maintains acceptable accuracy? How do false positive rates impact security operations center (SOC) analyst workload and alert fatigue in practice? What hybrid approaches combining multiple ensemble methods provide optimal robustness-accuracy trade-offs? This is essential, given that sophisticated adversaries will likely attempt to reverse-engineer and evade AI-based defenses.

6.2 Privacy-Preserving Secure Ai

While this study demonstrated exceptional performance using full-fidelity network features, many critical infrastructure contexts require privacy protection for operational or regulatory reasons. Future research should investigate how differential privacy, federated learning, and secure multi-party computation affect the performance of ensemble intrusion detection models. Specifically: Can Random Forest's distributed feature importance provide better privacy-utility trade-offs than Gradient Boosting's concentrated strategy? What privacy budgets (ε values) maintain detection rates above critical infrastructure security thresholds (e.g., 99% recall)? How can organizations share threat intelligence through federated learning while protecting proprietary operational data? This would enable operators to leverage collective threat intelligence while protecting sensitive operational information. Investigating how models trained on one critical infrastructure sector generalize to others would inform whether to adopt unified or sector-specific security frameworks [10]. This includes evaluating cross-sector model performance, identifying universal versus sector-specific threat patterns, developing efficient transfer learning approaches, and establishing multi-sector benchmark datasets.

6.3 Cross-Sector Generalization And Governance Integration

Three interconnected research directions would advance governance-ready AI for critical infrastructure: Cross-Sector Validation: Evaluate whether models trained on power grid traffic (using similar features to this study) generalize to transportation, healthcare, or financial infrastructure. Identify universal versus sector-specific features and develop efficient transfer learning approaches.

Human-AI Collaboration: Investigate how security analysts interact with SHAP explanations in operational contexts. Research questions include: What explanation formats maximize analyst trust calibration? How do interface designs affect decision quality when analysts review uncertain

classifications? What cognitive workload do AI-assisted workflows impose compared to manual analysis?

Integrated Governance Frameworks: Develop holistic frameworks simultaneously addressing security, privacy, and trust rather than treating them separately. Research should identify inherent trade-offs (e.g., privacy protections vs. detection accuracy), design principles optimizing across objectives, and comprehensive evaluation methodologies. Empirically validate governance frameworks through pilot deployments in controlled critical infrastructure testbeds.

The study proposes a three-phase research roadmap: Phase 1 (0-12 months): Validate findings on contemporary real-world datasets from diverse critical infrastructure sectors. Phase 2 (12-24 months): Develop and evaluate privacy-preserving and adversarially robust variants. Phase 3 (24-36 months): Conduct controlled operational deployments with participating critical infrastructure operators to assess real-world effectiveness and inform final governance recommendations. This phased approach would systematically address limitations identified in this study while building toward deployable, governed AI security solutions.

7. Conclusion

This research addresses critical gaps in AI governance for U.S. critical infrastructure through empirical validation of ensemble intrusion detection systems. Both Random Forest and Gradient Boosting classifiers achieved exceptional performance (99.87% and 99.80% effective accuracy, respectively), with Random Forest demonstrating superior generalization. Feature importance analysis identified forward packet length mean, flow duration, and inter-arrival time as key discriminators, providing practical guidance for network monitoring. SHAP-based explainability validated that model decisions align with cybersecurity domain knowledge, addressing transparency requirements essential for stakeholder trust. These findings provide quantifiable benchmarks for evidence-based AI governance. For the first time, policymakers have empirical data on achievable performance levels, required validation rigor, and feasible explainability approaches for critical infrastructure AI security. Our results demonstrate that ensemble methods can detect port scanning with false positive and false negative rates below 0.20%—establishing a performance ceiling that can inform regulatory standards and procurement requirements. However, critical caveats temper these promising results. The controlled NSL-KDD dataset may not fully represent operational complexity, and perfect performance metrics raise overfitting concerns requiring external validation. Our focus on port scanning represents one dimension of comprehensive security. Adversarial robustness, privacy-preserving techniques, and human-AI collaboration effectiveness require further investigation. Most importantly, operational validation with contemporary real-world critical infrastructure traffic is essential before these findings can inform deployment decisions. Despite these limitations, this study makes substantive contributions across three dimensions. Technically, we demonstrate ensemble methods' effectiveness and provide reproducible methodologies for rigorous validation. For policy, we establish empirical benchmarks enabling evidence-based governance rather than theoretical speculation. For practice, we offer concrete guidance on feature selection, preprocessing, and model selection for critical infrastructure security applications.

The path toward trustworthy AI in critical infrastructure requires sustained collaboration across disciplines - AI researchers, cybersecurity experts, infrastructure operators, policymakers, and social scientists. Technical excellence alone is insufficient; it must be complemented by thoughtful governance, stakeholder engagement, and commitment to security, privacy, and trust as equally important objectives. This research represents one step toward bridging the gap between AI's promise and its responsible

deployment in contexts where failures carry profound consequences. By demonstrating what is technically achievable while acknowledging limitations and uncertainties, we aim to advance a more informed, evidence-based discourse that enables both innovation and protection for the critical systems upon which modern society depends.

REFERENCES

[1] B. Guembe, A. Azeta, S. Misra, V. C. Osamor, L. Fernandez-Sanz, And V. Pospelova, The Emerging Threat Of AI-Driven Cyber-Attacks: A Review," Applied Artificial Intelligence, Vol. 36, No. 1, 2022. Https://Doi.Org/10.1080/08839514.2022.2037254

[2] M. Lehto, "Cyber-Attacks Against Critical Infrastructure," In Cyber Security: Critical Infrastructure Protection, Springer, 2022, Pp. 33–42. Https://Doi.Org/10.1007/978-3-030-91293-2_2

[3] A. O. Adewusi, U. I. Okoli, T. Olorunsogo, E. Adaga, D. O. Daraojimba, And O. C. Obi, "Artificial Intelligence In Cybersecurity: Protecting National Infrastructure: A Usa," World Journal Of Advanced Research And Reviews, Vol. 21, No. 1, Pp. 2263–2275, 2024. Https://Doi.Org/10.30574/Wjarr.2024.21.1.2963

[4] J. Sakhnini, H. Karimipour, A. Dehghantanha, And R. M. Parizi, "AI And Security Of Critical Infrastructure," In Handbook Of Big Data Privacy, Springer, 2020, Pp. 7–36. Https://Doi.Org/10.1007/978-3-030-38557-6_2

[5] Y. Alufaisan, L. R. Marusich, J. Z. Bakdash, Y. Zhou, And M. Kantarcioglu, "Does Explainable Artificial Intelligence Improve Human Decision-Making?" In Proc. AAAI Conf. Artificial Intelligence, Vol. 35, No. 8, 2021, Pp. 6618–6626. Https://Doi.Org/10.1609/Aaai.V35i8.16819

[6] I. Eleweke, R. Ugboko, O. Omotosho, R. Abbas, A. Adesokan, And I. P. Isibor, "Strengthening Security, Privacy, And Trust In Artificial Intelligence Software For Critical Infrastructure In The United States," *Engineering Science & Technology Journal*, Vol. 6, No. 4, Pp. 184–200, 2025.

[7] K. Mitsarakis, "Contemporary Cyber Threats To Critical Infrastructures: Management And Countermeasures," 2023.

[8] M. Ashraf And A. Haile, "Data Protection And AI: Navigating Regulatory Compliance In AI-Driven Systems," 2023.

[9] J. Kurre, "The Accountability, Responsibility & Governance As A Unified Strategy For AI," 2024.

[10] Department Of Homeland Security, "Roles And Responsibilities Framework For Artificial Intelligence In Critical Infrastructure," 2024. Https://Www.Dhs.Gov/Sites/Default/Files/2024-11/24_1114_dhs_ai-Roles-And-Responsibilities-Framework-508.Pdf

[11] P. Maharjan, "The Role Of Artificial Intelligence-Driven Big Data Analytics In Strengthening Cybersecurity Frameworks For Critical Infrastructure," In *Global Research Perspectives On Cybersecurity Governance, Policy, And Management*, 2023, Pp. 12–25.

[12] F. A. Yerlikaya And Ş. Bahtiyar, "Data Poisoning Attacks Against Machine Learning Algorithms," *Expert Systems With Applications*, Vol. 200, P. 116903, 2022. Https://Doi.Org/10.1016/J.Eswa.2022.116903

[13] N. Carlini And D. Wagner, "Towards Evaluating The Robustness Of Neural Networks," In *Proc. Ieee Symp. Security And Privacy*, 2017, Pp. 39–57. Https://Doi.Org/10.1109/Sp.2017.49

[14] S. M. Ali, A. Razzaque, M. Yousaf, And R. U. Shan, "An Automated Compliance Framework For Critical Infrastructure Security Through Artificial Intelligence," 2024.

[15] N. Sharma, E. A. Oriaku, And N. Oriaku, "Cost And Effects Of Data Breaches, Precautions, And Disclosure Laws," *International Journal Of Emerging Trends In Social Sciences*, Vol. 8, No. 1, Pp. 33–41.

[16] S. Gupta, "Towards Secure-By-Design Artificial Intelligence Systems," *Authorea Preprints*, 2024.

[17] I. K. Sarker, "Machine Learning: Algorithms, Real-World Applications And Research Directions," *Sn Computer Science*, Vol. 2, No. 3, P. 160, 2021. Https://Doi.Org/10.1007/S42979-021-00592-X

[18] M. Knodel, A. Fábrega, D. Ferrari, J. Leiken, B. L. Hou, D. Yen, And S. Park, "How To Think About End-To-End Encryption And AI: Training, Processing, Disclosure, And Consent," 2024.

[19] G. S. Kumar, K. Preethie, S. Madhumitha, R. Sushma, And M. Nivaashini, "Data Privacy Preservation Using Differential Privacy And Re-Identification Attacks," In *Proc. 2024 Int. Conf. Science Technology, Engineering And Management*, 2024, Pp. 1–6.

[20] A. Habbal, M. K. Ali, And M. A. Abuzaraida, "Artificial Intelligence Trust, Risk And Security Management (AI Trism): Frameworks, Applications, Challenges And Future Research Directions," *Expert Systems With Applications*, 2024.

[21] J. Jhurani, P. Reddy, And S. S. Choudhuri, "Fostering A Safe, Secure, And Trustworthy Artificial Intelligence Ecosystem In The United States," *International Journal Of Applied Engineering And Technology (London)*, Vol. 3, No. 2, Pp. 21–27, 2023.

[22] S. M. Lundberg And S. I. Lee, "A Unified Approach To Interpreting Model Predictions," In *Proc. 31st Int. Conf. Neural Information Processing Systems*, 2017, Pp. 4768–4777.

[23] National Institute Of Standards And Technology, "Artificial Intelligence Risk Management Framework (AI Rmf 1.0)," 2023. Https://Doi.Org/10.6028/Nist.AI.100-1

[24] The White House, "Executive Order 14110 On The Safe, Secure, And Trustworthy Development And Use Of Artificial Intelligence," Nov. 2023. Https://Www.Federalregister.Gov/Documents/2023/11/01/2023-24283/Safe-Secure-And-Trustworthy-Development-And-Use-Of-Artificial-Intelligence

[25] J. W. Creswell And J. D. Creswell, *Research Design: Qualitative, Quantitative, And Mixed Methods Approaches*, 5th Ed. Thousand Oaks, Ca: Sage Publications, 2018.

[26] A. L. Buczak And E. Guven, "A Survey Of Data Mining And Machine Learning Methods For Cyber Security Intrusion Detection," *Ieee Communications Surveys & Tutorials*, Vol. 18, No. 2, Pp. 1153–1176, 2016. Https://Doi.Org/10.1109/Comst.2015.2494502

[27] L. Breiman, "Random Forests," *Machine Learning*, Vol. 45, No. 1, Pp. 5–32, 2001. Https://Doi.Org/10.1023/A:1010933404324

[28] J. H. Friedman, "Greedy Function Approximation: A Gradient Boosting Machine," *Annals Of Statistics*, Vol. 29, No. 5, Pp. 1189–1232, 2001. Https://Doi.Org/10.1214/Aos/1013203451

[29] M. Tavallaee, E. Bagheri, W. Lu, And A. A. Ghorbani, "A Detailed Analysis Of The Kdd Cup 99 Data Set," In *Proc. Ieee Symp. Computational Intelligence For Security And Defense Applications*, 2009. Https://Doi.Org/10.1109/Cisda.2009.5356528

[30] G. Lyon, *Nmap Network Scanning: The Official Nmap Project Guide To Network Discovery And Security Scanning*. Sunnyvale, Ca: Insecure.Com Llc, 2009.

[31] D. C. Montgomery, E. A. Peck, And G. G. Vining, *Introduction To Linear Regression Analysis*, 5th Ed. Hoboken, Nj: John Wiley & Sons, 2012.

[32] R. J. A. Little And D. B. Rubin, *Statistical Analysis With Missing Data*, 3rd Ed. Hoboken, Nj: John Wiley & Sons, 2019. Https://Doi.Org/10.1002/9781119482260

[33] G. James, D. Witten, T. Hastie, And R. Tibshirani, *An Introduction To Statistical Learning With Applications In R*, 2nd Ed. New York, Ny: Springer, 2021. Https://Doi.Org/10.1007/978-1-0716-1418-1

[34] J. W. Tukey, *Exploratory Data Analysis*. Reading, Ma: Addison-Wesley, 1977.

[35] Z. H. Zhou, *Ensemble Methods: Foundations And Algorithms*. Boca Raton, Fl: Crc Press, 2012. Https://Doi.Org/10.1201/B12207

[36] R. Kohavi, "A Study Of Cross-Validation And Bootstrap For Accuracy Estimation And Model Selection," In *Proc. 14th Int. Joint Conf. Artificial Intelligence*, Vol. 2, 1995, Pp. 1137–1143.

[37] C. D. Manning, P. Raghavan, And H. Schütze, *Introduction To Information Retrieval*. Cambridge, Uk: Cambridge University Press, 2008. Https://Doi.Org/10.1017/Cbo9780511809071

[38] T. Fawcett, "An Introduction To Roc Analysis," *Pattern Recognition Letters*, Vol. 27, No. 8, Pp. 861–874, 2006. Https://Doi.Org/10.1016/J.Patrec.2005.10.010

[39] K. P. Okpara, "Human-Centric Machine Learning Intrusion Detection For Smart Grid Scada Systems, Grounded In Human-Systems Integration Theory," *American Academic Scientific Research Journal For Engineering, Technology,* And Sciences, 2025. Https://Doi.Org/10.13140/Rg.2.2.26606.32326

[40] F. Pedregosa Et Al., "Scikit-Learn: Machine Learning In Python," *Journal Of Machine Learning Research,* Vol. 12, Pp. 2825–2830, 2011.

[41] C. M. Bishop, *Pattern Recognition And Machine Learning*. New York, Ny: Springer, 2006.

[42] C. M. Ahmed, G. R. Mr, And A. P. Mathur, "Challenges In Machine Learning Based Approaches ForReal-Time Anomaly Detection In Industrial Control Systems," In *Proc. 6th Acm On Cyber-Physical System Security Workshop*, 2020, Pp. 23–29. Https://Doi.Org/10.1145/3384941.3409588

[43] N. Papernot, P. Mcdaniel, I. Goodfellow, S. Jha, Z. B. Celik, And A. Swami, "Practical Black-Box Attacks Against Machine Learning," In *Proc. Acm On Asia Conf. Computer And Communications Security*, 2017, Pp. 506–519. Https://Doi.Org/10.1145/3052973.3053009

[44] P. Pu And L. Chen, "User-Involved Preference Elicitation For Product Search And Recommender Systems," *AI Magazine*, Vol. 29, No. 4, Pp. 93–103, 2008. Https://Doi.Org/10.1609/Aimag.V29i4.2200

AUTHOR

Kelechi P. Okpara was born in Ogbunka, Anambra State, Nigeria, in 1983. He received the B.Eng. degree in Mechanical/Production Engineering from Nnamdi Azikiwe University, Awka, Nigeria, in 2006, the M.Eng. degree in Mechanical Engineering (Thermo-Fluids) from the University of Port Harcourt, Nigeria, in 2011, and the M.S. degree in Information Technology from Western Governors University, UT, USA, in 2020. He is currently a Ph.D. candidate in Information Systems at the University of the Cumberlands in the USA. Since 2019, he has served as a Technology Strategy and Advisory Consultant specializing in information systems security and governance. He advises Fortune 100 organizations on cloud security, enterprise technology strategy, information systems security, and data governance in an increasingly agentic, AI-enabled environment. He is the author of two published articles and several unpublished works. His research interests include AI-era information systems security and governance, cybersecurity governance for critical infrastructure, and resilient security leadership in complex socio-technical systems. Mr. Okpara is a distinguished member of ISACA and OCEG.