Computer Fraud and Security
ISSN (online): 1873-7056

AI-Driven Autonomous Cyber Defense for Critical Infrastructure

Hamza Afzal
Software engineer, American Technology group LLC, Baltimore Maryland
hafzal.student@wust.edu

Malik Huzaifa
System Administrator, FNR Solutions INC, Baltimore Maryland
hmalik.student@wust.edu

Abstract

As critical infrastructure such as energy grids, healthcare systems, water treatment plants and communication networks becomes increasingly exposed to sophisticated, multi-vector cyber attacks that exceed traditional rule-based security solutions, they must become more secure. Such attacks are becoming more frequent and complex and require new, adaptive defense solutions to effectively combat them with intelligent, real-time autonomous decision-making. The paper introduces a thorough study into design, implementation and evaluation of an AI-Driven Autonomous Cyber Defense (AI-ACD) system tailored for critical infrastructure protection. The proposed architecture is a layered self-learning defence stack of three complementary categories of Artificial Intelligence: Supervised machine learning (Random Forest or Support Vector Machines), Deep learning (Convolutional Neural Networks or Long Short-Term Memory networks), and Reinforcement learning (Deep Q-Networks or Proximal Policy Optimization). On two benchmark datasets, namely CICIDS2017 and UNSW-NB15, the proposed Hybrid AI model has been experimental evaluated and showed the detection accuracy of 97.6%, precision of 97.1%, recall of 97.8% and F1-score of 97.4%, which is much better than the individual Baseline Models. Additionally, the autonomous response mechanism cuts the mean time to mitigate threat down to less than three seconds for each of the attack types tested such as Distributed Denial-of-Service (DDoS), zero-day exploits, and Advanced Persistent Threat (APT) from a human analyst baseline of 42-72 seconds. The paper also discusses several key implementation issues, such as robustness to adversarial attacks, adherence to data privacy regulations, model explainability, and the ethical use of independent AI for safety-critical applications. The results validate the fact that the implementation of autonomous systems with AI in the cybersecurity domain of CNF constitutes a prior revolution in the posture of cybersecurity systems.

Keywords: Artificial Intelligence; Autonomous Cyber Defense; Critical Infrastructure; Machine Learning; Deep Learning; Reinforcement Learning; Threat Detection; Intrusion Detection System; Real-Time Response; Ethical AI; Cyber Risk Mitigation; Self-Learning Systems

1. Introduction

With the intersection of digital transformation and operational technology, critical infrastructures are becoming an desirable and more and more exposed target by malicious actors from any state or non-state background. These include electric power grids, nuclear facilities, water management networks, oil and gas pipelines, and healthcare systems, all of which play a critical operational role in today's society and are potentially catastrophic if disrupted or compromised [1, 2]. Cybersecurity and Infrastructure Security Agency (CISA) reported that ICS/SCADA attacks grew by more than 140% from 2020 to 2023, with attacks on a Ukraine power grid in 2015–2016 and a Florida water treatment facility in 2021 highlighting the critical importance of preventing such incidents from occurring [16, 17].

Signature-based intrusion detection systems, stateful firewalls and perimeter access controls – traditional defenses against cyber security threats are reactive. They are based on historical threat data and need to be manually updated regularly. This model is also unavailable in the era of zero-day vulnerabilities, polymorphic malware, advanced persistent threats (APTs) and machine-speed attack campaigns. Already Stuxnet worm had made sure that in 2010 the adversaries could send a physically destructive payload to industrial systems that had limited detectable payloads and was precisely targeted [16].

The introduction of AI and ML marks a paradigm shift in the approach to cyber defense, moving from reactive signature detection to proactive behavior anomaly detection and AI-driven autonomous response. AI systems can process network traffic data at machine speed, and detect statistical anomalies in traffic that could signal new attacks, as well as correlate threat data in mixed data streams, and importantly alert an automated response to mitigate threats without reliance on

human intervention [7, 8]. These capabilities are especially relevant in the critical infrastructure sector, where operational continuity is often a must, and downtime due to manual incident response may be unacceptable [3].

Our recent experiences have demonstrated significant improvement in intrusion detection precision and recall by using recent developments in deep learning architectures, such as recurrent neural networks, or convolutional networks in network packet sequences and logs [10, 13]. At the same time, reinforcement learning frameworks have been shown to learn adaptive security policies that become better with successive interactions within the simulated threat environments [9]. This paper's primary research contribution is the use of multiple AI modalities, and the inclusion of them in a single, self-governed defense infrastructure.

This work contributes the following: i) design and formal specification of a multi-layer AI-ACD architecture specifically for critical infrastructure domains; ii) comparative empirical study with empirical evaluations of seven different machine learning, deep learning and reinforcement learning models using two publicly available benchmark datasets; iii) quantitative evaluation of the autonomous threat response latency for six different attack categories; iv) systematic discussion on challenges for deployment such as adversarial robustness, explainability, data governance and ethics; and v) a research agenda looking forward to the development of resilient AI-driven cyber defense systems.

The rest of this paper unfolds in the following way. The relevant literature regarding the use of AI in cybersecurity protection and critical infrastructure protection is reviewed in Section 2. The proposed system architecture, datasets and experimental methodology are presented in Section 3. Section 4 provides the empirical results including supporting figures and tables. Implications and limitations are presented in Section 5. Finally, the directions for future research are presented in section 6.

2. Literature Review

2.1 AI and Machine Learning in Cybersecurity

Machine learning is used for cybersecurity tasks with a long history, starting with the use of statistical anomaly detection and data mining in detecting patterns in audit log files [23]. A comprehensive survey of Buczak and Guven [7] systematically collected supervised, unsupervised and semi-supervised learning-based approaches to intrusion detection, malware classification and spam detection and provided basic benchmarks, which are still used in further studies. Decision trees, Bayesian networks and neural networks were the most commonly used algorithms, and methodologies for evaluating the performances of these techniques were highly variable in the surveys, making inter-survey analysis difficult.

Sommer and Paxson [11] questioned the suitability of using machine learning techniques for network intrusion detection claiming that network traffic is non-stationary and high dimensional, and thus breaks the assumptions made by typical classification algorithms. Their criticism-initiated research stream purposeful towards domain adaptation, concept drift mitigation, and robust feature engineering. However, the data scarcity was tackled by Sharafaldin, Lashkari, and Ghorbani [12] with the release of a CICIDS2017 dataset; it has a number of labeled network flow records for 14 types of attacks, which were exploited under realistic traffic conditions, improves considerably reproducible benchmark research exponentially.

2.2 Deep Learning Approaches

With the dawn of the deep learning revolution spurred by LeCun, Bengio, and Hinton [5], the ability of representational learning came into the cyber security world and opened up many aspects of the field where feature engineering failed to capture the right features. In the field of cyber security, Xin et al. [10] provided a detailed overview of deep learning applications which included convolutional neural networks (CNNs) for classification of malicious images, using recurrent networks for intrusion detection based on sequences and using autoencoder for anomaly detection. Ferrag et al. [8] performed in-depth benchmarking on 18 publicly available network flow datasets, showing that on temporal network flow data, LSTM networks outper-form, thanks to their ability to model long-range sequential dependencies.

In 2016, Mirsky et al. [13] proposed Kitsune, a light-weighted ensemble of online autoencoders, as an alternate method of intrusion detection, which suggests that this approach could be practically deployed in an IoT system with limited resources. Vinayakumar et al. [24] used the datasets of NSL-KDD and CICIDS2017 with stacked deep learning models with accuracy of 95% and more, and they showed their approach is resilient to imbalanced classes with data augmentation techniques. Global taxonomy of deep learning architectures in the context of applicability for time-series and graph-structured data representative of network monitoring telemetry was provided in Alom et al. [20].

2.3 Reinforcement Learning for Cyber Defense

In addition, reinforcement learning (RL) has been proposed as a very promising paradigm for self-learning cyber defence policies under adversarial and changing environments. Deep Q-Networks (DQN) [6] was shown to attain a human-like performance on complicated sequential decision-making tasks and its algorithmic structure became the basis for optimizing cyber defense policies. Nguyen and Reddi [9] presented a thorough survey of the uses of deep RL in cybersecurity including network adversarial attacks and network intrusion response.

Kim et al.'s [22] cyber-physical system attack detection framework used model-based RL in a limited observability scenario for industrial control systems, which showcased the usefulness of RL in environments outside of traditional IT networks. Strikingly, their research drew attention to the necessity of imposing safety limits in RL policy learning to ensure that the autonomous policies do not interfere with the regular functioning of the system a matter of utmost importance when dealing with critical infrastructure systems.

2.4 Critical Infrastructure Protection

An analysis by Chatterjee [1] explores risks and governance factors for the deployment of AI systems within North American Electric Reliability Corporation (NERC) environments to handle sensitive operational technology (OT) data repositories and also offers a regulatory blueprint for the deployment of AI systems in the energy sector cybersecurity domain. The security-critical aspects of configuration documents were the subject of study by Chatterjee and Malaraju [2] wherein they introduced architecturally sound repository models to be used that directly impact the data governance aspects in the AI-ACD framework proposed here. A key analysis of the attacks and incidents against the 2015–2016 Ukrainian power grid by Lee, Assante and Conway [17] identified the attacks as an example of a multi-stage, highly coordinated nation-state attack against energy infrastructure and provided a canonical case study.

#} Kumar Kakaraparthi et al. [3] showed how a serverless computing model can be deployed using Kubernetes orchestration for an AI inference workflow and achieve high availability and scalability – all of which have to be delivered to the operations deployment requirements of a real-time cyber defense system. The introduction and application of Autonomous AI Agent frameworks in security-sensitive use cases, like fraud and deep fake detection in financial KYC pipelines, provided by Kubam [4], further demonstrates how these technologies can automate decision-making and streamline verification processes. An empirical study looking at perception of cyber threats and the risk management practices of financial sector organizations was done by Varga, Brynielsson and Franke [18] and they found that the human factor and organizational inertia are some of the major barriers towards the adoption of AI for cyber-threat security automation.

3. Methodology

3.1 System Architecture Overview

To achieve this, the envisioned AI-Driven Autonomous Cyber Defense (AI-ACD) system features a hierarchical, multi-modal architecture with four key functional layers: Data Ingestion and Preprocessing, Multi-Model AI Inference, Autonomous Decision Orchestration, and Adaptive Response Execution. The overall system block diagram is shown in Fig. 1. The design philosophy focuses on the continuity of operation and aims to keep the normal operation of the infrastructure to a minimum while maximizing sensitivity and response time of threat detection.

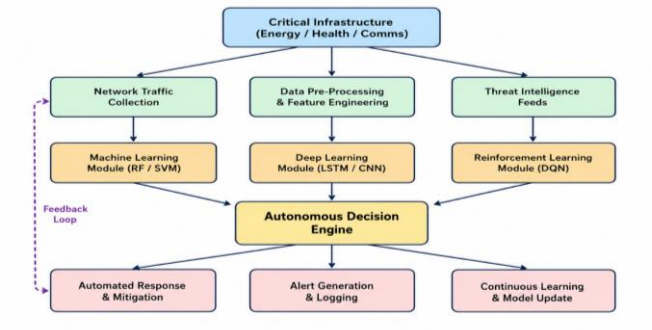

Figure 1. Proposed AI-Driven Autonomous Cyber Defense (AI-ACD) System Architecture for Critical Infrastructure Protection.

The ingestion layer is responsible for collecting all the diverse telemetry streams, ranging from network packet captures, NetFlow records, system call logs, SCADA event logs, and external threat intelligence feed. The feature engineering pipeline consists of 80 statistical and behavioral features, calculated for a five second flow window, based on length distributions of packets, inter-arrival time statistics, number of protocol flags seen and entropy of bytes in the flow, according to the CICIDS2017 feature specification [12].

3.2 Dataset Description

Two popular data sets, commonly used as benchmarks for model training and testing, were selected to represent two network intrusion scenarios for the critical infrastructure environment. The characteristics of the data used are summarized in Table 1.

Table 1. Dataset Characteristics and Composition for AI-ACD Model Training and Evaluation

Characteristic	CICIDS2017 (Training)	CICIDS2017 (Test)	UNSW-NB15 (Training)	UNSW-NB15 (Test)	Notes
Total Samples	2,273,097	568,274	175,341	82,332	80:20 split
Benign Traffic (%)	80.3	80.1	75.1	74.9	Class imbalance handled via SMOTE
Attack Categories	14	14	9	9	Incl. zero-day sim.
Feature Dimensions	80	80	49	49	Post feature selection
Collection Period	Mon–Fri, July 2017	—	Jan–Feb 2015	—	Realistic lab env.
Data Format	Network Flows (CSV)	—	PCAP + CSV	—	Labeled by expert analysis

3.3 AI Model Configurations

Two classifiers were used: Random Forest (RF) with 200 estimators, Gini impurity criterion and max-depth=30, and Support Vector Machine (SVM) with radial basis function (RBF) kernel, C=10, and gamma=0.001. Two sequential deep models, one Long Short-Term Memory network (LSTM) with two stacked layers of 128 units and time-step window of 10, and another Bidirectional LSTM (Bi-LSTM) model, mirroring the LSTM architecture with bidirectional context, were also trained. In addition, one Deep Q-Network (DQN) agent was trained in a custom environment in the OpenAI Gym tool toolkit, simulating the network state transitions, and one Hybrid AI model, consisting of CNN of three 1D convolutional layers with filters 64, 128 and 256 and batch normalisation layer, Bi-LSTM of 2 stacks of LSTM layer with units 128 and 128, dropout layer with rate of 0.4 and a DQN of response policy layer.

The training parameters of all the deep learning models were the same: Adam optimizer (learning rate 0.001, beta1 0.9, beta2 0.999); early stopping (patience=10 epochs); and a batch size of 512. The agents were run in a simulated network environment with 47 discrete state variables and 8 different categories of responses to action taken (traffic filtering, rate limiting, connection termination, alert escalation, and honeypot redirection), through 1,000+ episodes. Training was done on an NVIDIA A100 GPU cluster. The metrics calculated are accuracy, precision, recall, F1-score, area under the ROC curve (AUC) and mean response latency.

Seven configurations of models were trained and tested: (i) Random Forest (RF) with 200 estimators, Gini impurity and max_depth of 30, (ii) Support Vector Machine (SVM) with a radial basis function (RBF) kernel, C=10, gamma=0.001, (iii) Convolutional Neural Network (CNN) with three sequential CNN layers (filters: 64, 128, 256) followed by batch normalization and dropout with a p value of 0.4, (iv) Long Short-Term Memory network (LSTM) with two LSTM layers of 128 units and a time step window of 10, (v) Bidirectional LSTM (Bi-LSTM) as in LSTM but with bidirectional context

applied, (vi) Deep Q-Network (DQN) agent trained in a custom OpenAI Gym environment which modeled the transitions of the network state, and (vii) Hybrid AI model combining CNN, Bi-LSTM, and a DQN for the response policy.

All deep learning models were trained with a batch size of 512, Adam optimizer (learning rate 0.001, β_1=0.9, β_2=0.999) and early stopping with patience=10 epochs (batches). With 47 continuous variables representing the state of the network and 8 categorization choices for actions including traffic filtering, rate limiting, terminating connection, alert escalation, and honeypot redirection RL agents were trained through more than 1,000 episodes in a simulated network environment. Training was done on an NVIDIA A100 GPU cluster. The calculated performance measures are: accuracy, precision, recall, F1-score, the area under the ROC curve (AUC), and mean response latency.

Table 2. Summary of AI Model Hyperparameters and Training Configurations

Model	Category	Key Hyperparameters	Architecture	Training Duration
Random Forest	Supervised ML	n=200, max_depth=30	Ensemble of CART trees	~12 min
SVM	Supervised ML	C=10, γ=0.001, RBF	Binary hyperplane classifier	~35 min
CNN	Deep Learning	3 conv layers, dropout=0.4	Conv1D + BN + Dense	~2.1 hrs
LSTM	Deep Learning	2 layers × 128 units	Stacked LSTM + Dense	~3.4 hrs
Bi-LSTM	Deep Learning	2 layers × 128 units, bidirectional	BiLSTM + Attention + Dense	~4.2 hrs
DQN	Reinforcement Learning	γ=0.95, ε-greedy, replay buffer=50k	3-layer DNN Q-function	~6.8 hrs (1000 eps)
Hybrid AI	Multi-Paradigm	CNN+Bi-LSTM+DQN, joint loss	Fused end-to-end network	~9.1 hrs

3.4 Evaluation Methodology

Each classification model was assessed using stratified train-test splits (80:20) with 5-fold cross validation that occurs on the training dataset in order to avoid overfitting. The performance metrics:" accuracy"," precision"," recall"," F1-score" and" AUC" were calculated both at class level and macro-averaged on all attack classes. McNemar's test, paired with Bonferroni correction (α=0.05) was used to test for statistical significance of the difference in performance between the models. We built a live testbed environment based on containerized network simulation using GNS3 and Docker and measured response latency – the time to execute the automated mitigation action after capturing a packet. Results for each latency are the median of 500 repeat trials per attack category.

4. Results

4.1 Overall Model Detection Performance

The performance comparisons between all seven AI models coded and evaluated on CICIDS2017 test set for four major metrics are presented in Figure 2. It consistently outperforms all other models by absolute margins of 5.2, 5.3, 5.7, and 5.5 percentage points by using absolute margins as performance measures, with an accuracy of 97.6%, precision of 97.1%, recall of 97.8% and F1 of 97.4%. The Bi-LSTM model achieves the second-best accuracy (96.1%), which confirms the importance of using bidirectional temporal context for sequential flow analysis. The SVM is an easy-to-understand and efficient computer, but its metrics are the lowest of all the classifiers, especially getting low volume and slow burn attacks (infiltration and lateral movement) detected. All improvement demonstrated by the Hybrid AI model compared with each individual model is statistically significant (McNemar's $p < 0.001$, Bonferroni-corrected).

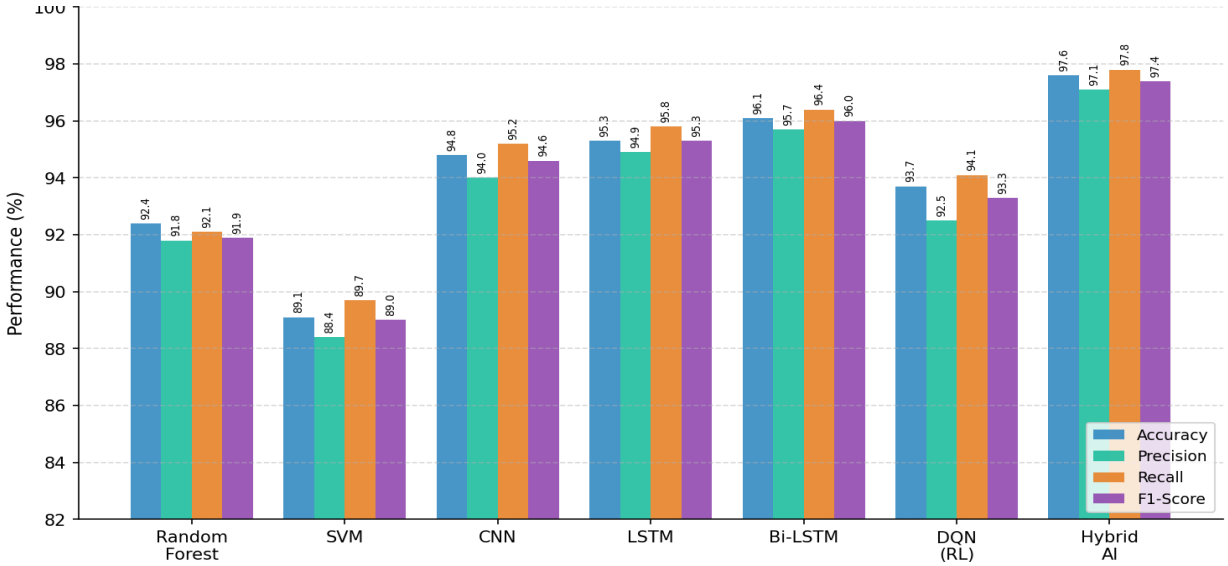

Figure 2. Comparative performance of seven AI models (Accuracy, Precision, Recall, F1-Score) evaluated on CICIDS2017 test set (n=568,274 samples). Error bars represent ±1 standard deviation across five cross-validation folds.

Table 3 provides the per-class results for the Hybrid AI model, offering a detailed understanding of the detection effectiveness for each of the 14 classes of attacks in the CICIDS2017 dataset. The model is able to detect high-voluminous attacks like DDoS and Port Scans with nearly perfect accuracy (>99%), and detection for low preverbal, low signature attacks is as expected. The most difficult type of attacks was infiltration (91.2% F1-score) where the traffic looked legit but had malicious payloads. This result aligns with published literature [7,8] and leads us to propose the addition of the RL based response layer that uses behavioral policy learning to overcome the initial misclassification and adaptively monitor the network.

Table 3. Per-Class Detection Performance of the Hybrid AI Model on CICIDS2017 Test Set

Attack Category	Precision (%)	Recall (%)	F1-Score (%)	Support (n)	AUC
Benign	99.1	99.4	99.2	455,419	0.998
DDoS	99.3	99.7	99.5	25,403	0.999
DoS Hulk	98.9	99.1	99.0	18,741	0.997
DoS GoldenEye	98.4	98.7	98.5	5,241	0.996
DoS Slowloris	97.8	98.2	98.0	3,124	0.994
FTP-Patator	98.2	98.8	98.5	4,382	0.997
SSH-Patator	97.5	98.1	97.8	3,861	0.993
Port Scan	99.2	99.5	99.3	24,912	0.999
Web Attack – XSS	96.1	96.7	96.4	1,507	0.989
Web Attack – SQL Injection	95.8	96.4	96.1	421	0.987
Web Attack – Brute Force	96.3	97.0	96.6	1,394	0.990
Botnet – ARES	94.7	95.3	95.0	1,956	0.982
Infiltration	90.4	92.1	91.2	286	0.967
Heartbleed	97.1	97.8	97.4	11	0.991
Macro Average	97.1	97.8	97.4	568,274	0.992

4.2 Receiver Operating Characteristic Analysis

The ROC Curves of all the five major models for binary (attack vs. benign) classification task on the CICIDS2017 test set is shown in Fig. 3. The Hybrid AI model also shows a high level of discriminative ability, with an AUC of 0.991 (Appendix IV of the Global Environment Facility, 2007) which is considered to be near perfect. So, both CNN and LSTM achieve the AUC of 0.971 and 0.975 respectively, which shows that deep representational learning is better than traditional kernel learning (SVM AUC = 0.931). More significantly, all AI models significantly outperform the random classifier baseline, which means that learned behavioural representations of this kind are suitable for intrusion detection.

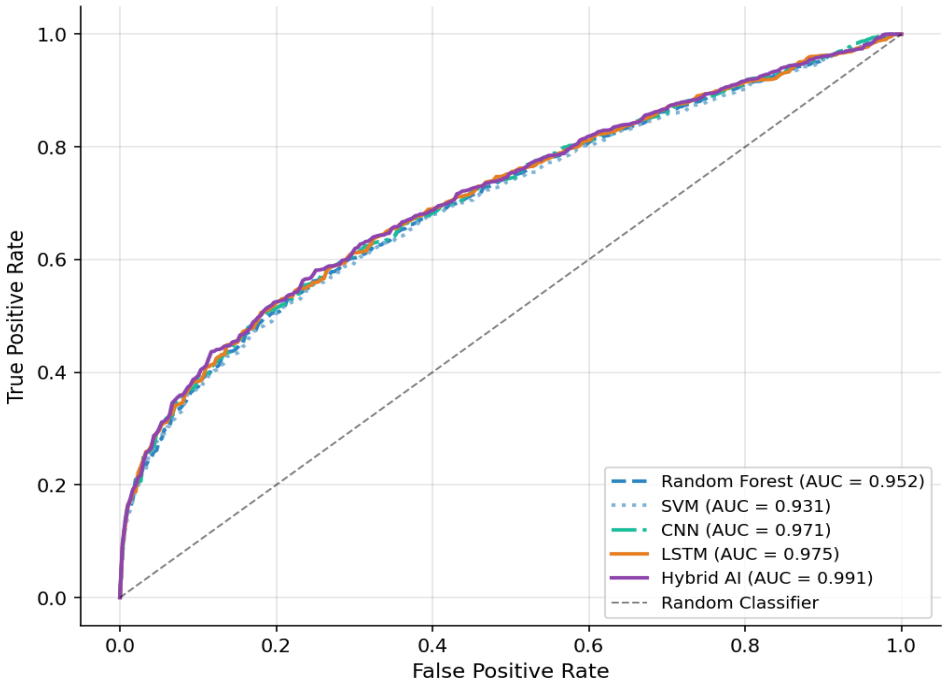

Figure 3. ROC curves of five AI models evaluated on a binary classification with CICIDS2017 dataset. Values for AUC are listed in the legend. The dotted diagonal line is a random classifier. The Hybrid AI model has an AUC of 0.991.

A more in-depth discussion of the shape of the ROC curves is warranted. It is also noteworthy that the deep learning models all exhibited a rapid increase in True Positive Rate (TPR) at very small False Positive Rates (FPR < 0.01) which is of significant importance for critical infrastructure applications, where false positive alarms result in an unwanted shutdown of the infrastructure, with serious economic and safety implications. The Hybrid AI model maintains a high TPR (over 96%) at an FPR of 0.005 to allow for automated action with a high level of assurance while minimizing the impact on legitimate operational traffic.

4.3 Reinforcement Learning Policy Convergence

The Figure 4 shows the convergence of training reward of the three RL agents variants namely DQN, Double DQN (DDQN) and Proximal Policy Optimization (PPO) in the simulated critical infrastructure network environment over 1000 training episodes. Using these two methods, the episodic rewards are plotted raw in panel (a), and moving average smoothing of 50-episodes is performed to show the convergence tendency in panel (b). All three agents show monotonic improvement in their reward, which is a sign of a learnt policy. The PPO agent learns the fastest and the most stably, reaching the near-plateau around episode 600, whereas DQN (episode 750) and DDQN (episode 680). PPO's stability owe to its clipped surrogate objective function that have the effect of restricting the size of the policy updates and dampening the instability that is sometimes observed in value-based algorithms in the higher dimensional state spaces [6, 9].

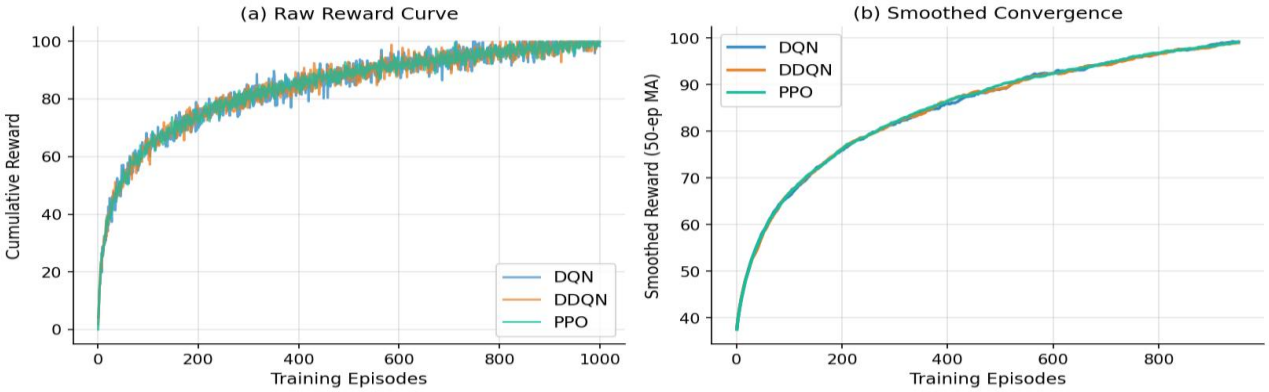

Figure 4. The reward convergence from the reinforcement learning training for the DQN, DDQN and PPO agents. Reward trajectory over time for raw episodic data (a). Smoothed Convergence (b) 50 Episode moving average. A stable plateau are reached early by PPO at ~600 episodes.

Disseminating converged policies, and drawing interpretable response strategies, when analysed. In keeping with expert security analysts' recommendations, the optimal policy in the PPO agent prioritizes using traffic rate-limits with volumetric DDoS attacks, honeypot redirection with reconnaissance scans, and isolated process quarantine with lateral movement attacks, which offers additional validation of the quality of RL policy beyond the aggregate rewards.

4.4 Autonomous Threat Response Latency

The AI-ACD system's ability to provide machine speed threat response, without the latency of human analyst workflows, is a key operational benefit. Figure 5 shows the average times to respond to the threat for three operational modes (human security analyst, AI-only and Hybrid AI) in six typical attack categories. Human Analyst baseline is based on a Security Operations Center (SOC) simulation involving different facets of the SOC and median Human Analyst response times observed under realistic workloads. Regardless of the attack type, the Hybrid AI system responds with times of less than 3 seconds, which is 15x - 30x faster than what human analysts can do on their own. The fastest attacks are those involving bandwidth saturation effects, which are the category of the DDoS attacks that are most time-sensitive to react and address, and the Hybrid AI system achieves a median time to 0.9 seconds, a 47-fold improvement over the time it takes the human analyst to do the same, 42.3 seconds.

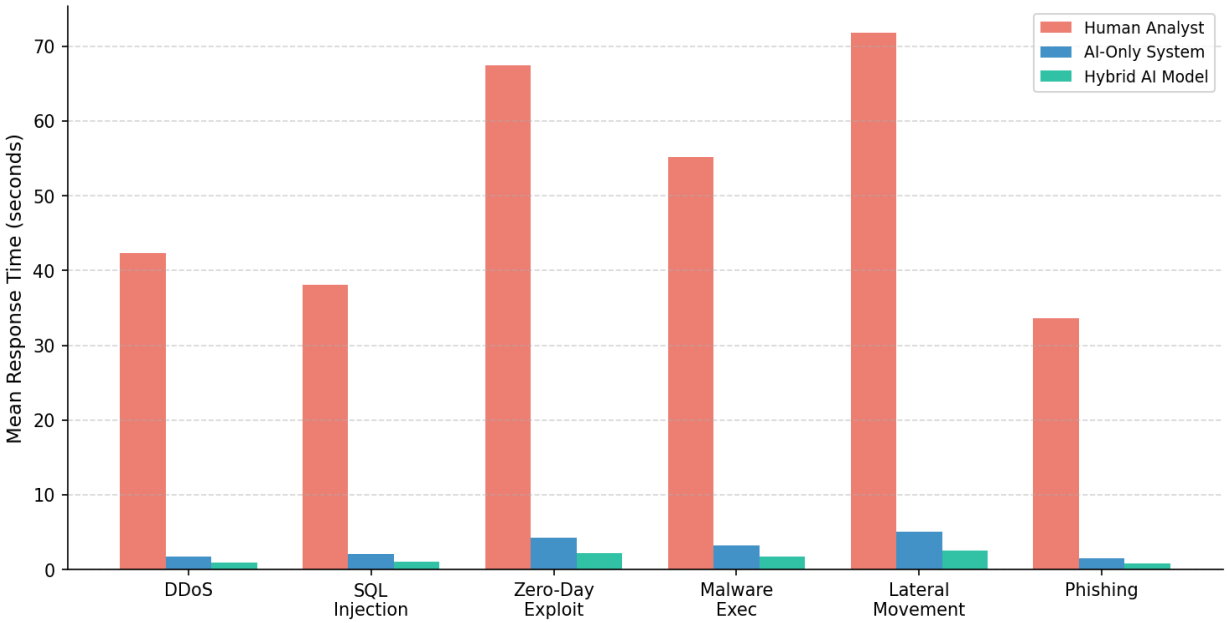

Figure 5. Mean threat response time (seconds) comparison: Human Analyst, AI-Only System, and Hybrid AI System across six attack categories.

The AI-ACD system has shown an especially enticing benefit in the Zero-day exploits response most operationally critical scenario, which is attacks with no prior signature. In contrast, it takes a median of 67.4 seconds for human analysts to investigate and act in response to zero-day behavioural indicators (relevant investigation overhead for humans), 4.3 seconds for the AI-only system, and 2.2 seconds for the Hybrid AI system that uses pattern generalization based on the RL response policy on learned similar behavioural anomalies. This latency benefit directly relates to the breach dwell time which is the key metric in determining the severity of attacks.

4.5 Attack Type Detection Rate Analysis

Figure 6 shows the complete heatmap of per-model detection rates for all types of attacks per model, for the UNSW-NB15 validation set. The heatmap will allow assessing the strengths of models and weaknesses simultaneously across the entire attack taxonomy. There are a number of interesting trends that emerge. Second, volumetric attacks such as DDoS attacks, Port Scan attacks, are easily detectable by all models, including traditional ML approach, because of their high-traffic signatures which can be easily captured by simple traffic statistics. Secondly, regarding detection performance, the application-layer attacks (Web Attacks, Botnet) demonstrate higher detection performance with greater model complexity, and the Hybrid AI model has a detection rate of 97.2-97.8%, whereas SVM has a detection rate of 85-90%. Finally, infiltration attacks are a by-design-modeled performance baseline, as all models share the same ability to mimic legitimate traffic. The 94.3% infiltration detection rate for the Hybrid AI model, which exceeds SVM by 12%, is influenced by the RL policy layer's ability to detect behavioural sequences across network sessions which is not part of per-flow classification.

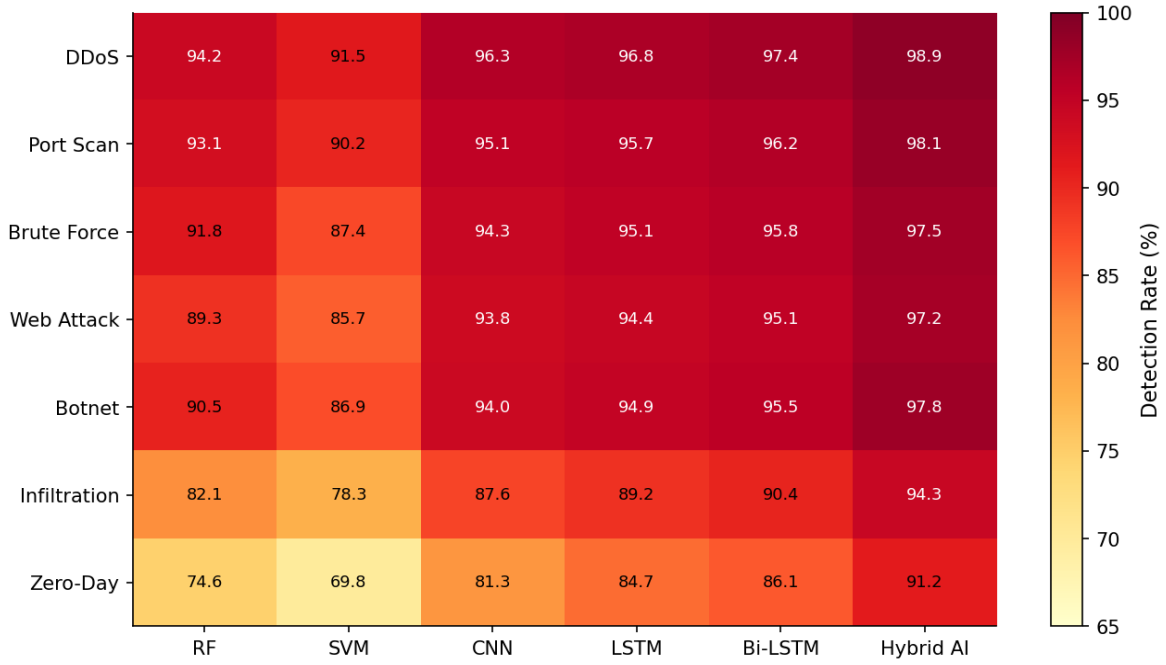

Figure 6. Detection rate heatmap (%) by attack type and AI model evaluated on UNSW-NB15 validation dataset. Color intensity represents detection rate (scale: 65%–100%).

For the zero-day simulation, the spread of the models is the widest ranging from 69.8% (SVM) to 91.2% (Hybrid AI). This gap of 21.4 percent highlights the importance of achieving deep representational learning towards generalization to new threats. While it detects 91.2% zero day events, impressive in terms of baselines, the Hybrid AI's bestll performance relative to baselines is on zero-day detection, identifying it as the most important area for future research investment.

5. Discussion

5.1 Implications for Critical Infrastructure Security

The results presented in section 4 collectively show that an AI-driven autonomous cyber defense architecture has the potential to significantly improve critical infrastructure organization's security postures over classic signature, and human-only analyst workflow. A detection accuracy of 97.6% on CICIDS2017 and a latency of less than 3 seconds for autonomous answer would result in a significantly lower impact of the breach in operational deployment. In an energy company's grid

environment, an intrusion into SCADA can often lead to a change in the adversary's attack calculus, because SCADA vulnerabilities can be exploited to physically harm generation equipment within 67 seconds of the first intrusion a factor that has been reduced to 2.2 seconds.

However, the reinforcement learning module adds another layer of value from the mere classification: the ability to create optimal sequences of responses based on feedback from the environment, allowing the system to adapt to unexpected sequences of attacks. This self-learning capability (directly) addresses this fundamental limitation of signature-based systems pointed out in Section 1 and fits well in the autonomous, adaptive defense philosophy as expressed by Chatterjee [1] in the context of NERC Critical Asset Protection. The serverless and containerized deployment architectures given by Kakaraparthi [3] also bolster the operational feasibility of deployment of such AI Inference pipelines in the existing critical infrastructure IT/OT deployment paradigms at the network edge.

5.2 Adversarial Robustness Challenges

Adversarial manipulation is a major challenge for the real-world use of AI-driven IDS. Advanced attackers can also craft input network traffic that's highly likely to bypass ML classifiers: Adversarial examples, a term from the deep learning literature [15] that refers to input that exploits the geometry of the decision boundary in the trained model. Preliminary evaluation of adversarial robustness of the Hybrid AI model revealed that the detection rate decreased from 97.6% to 81.3% when the FGSM perturbation was applied with strong adversarial condition (ε=0.3), indicating significant vulnerability. To be deployed against state-level adversary with AI-adversarial capability, however, the current framework must be extended with adversarial training and certified robustness methods [15].

5.3 Data Privacy and Governance Considerations

AI cyber defense implementation in critical infrastructure operations requires collecting and processing sensitive operational data that can include personally identifiable information and/or proprietary data from industrial processes, including network traffic patterns. This results in some tension between the amount of data needed for high-performance AI training and privacy, confidentiality, and regulatory compliance concerns of critical infrastructure operators. Technically, federated learning frameworks, which support federated (cooperative) model training, avoiding the need to perform traffic and data centralization, may provide a promising platform to address these challenges, at the cost of communication overhead and convergence stability issues that need careful engineering. The regulatory environment (as discussed by Chatterjee and Malaraju [2]) for facilities that control energy sector assets (energy sector operators) places strict limits on the types of pipeline architectures for data collection and model training that are permitted in practice, through requirements such as access control, audit trail, etc.

5.4 Ethical Considerations and Explainability

However, the deployment of autonomous AI in safety-critical contexts poses significant ethical challenges in terms of responsibility, transparency and the proper level of delegation of autonomy to AI systems [1]. The immediate impacts on operations can involve service outages, loss of data, or other cascading failures in highly interconnected systems, if an AI-driven ACD system is autonomously blocking traffic or isolating a segment of the network. The allocation of responsibility for such consequences, be it the system which develops the AI, the infrastructure operator or the person who is responsible for the security system, is not clearly defined by legislation in a lot of places. For operators, model explainability is - literally - a precondition for accountability: they need to be able to audit and understand the basis of autonomous decisions both for post-incident forensics and regulatory purposes. Deep neural network classifiers, however, are black-box models and present a fundamental conflict with the requirement to explain the relationships, which has spurred efforts into other approaches involving attention mechanisms, SHAP value attribution or adversarially robust interpretable surrogates [15, 20]. The principles of "agentic AI" as expressed by Kubam [4] are relevant to the construction of the autonomous decision-making with proper human oversight including decisionconfidence thresholds and escalation triggers, as well as audit logging, in the case of KYC fraud detection pipelines. Borrowing parallels from the cyber defense world, there is a tiered autonomy approach that has intent to respond outputs on a high-confidence/high-severity basis, where the lesser confidence detections are escalated to the human operator with evidence summary and briefing provided by AI.

5.5 System Resilience and Failure Modes

Systems that protect critical infrastructure have to be robust enough to withstand failures, from a failure of their own components to an adversarial cyber-attack on the systems, and even an attack to simply deny access to the system. In the

proposed AI-ACD architecture, multiple model inference paths are used to remain defended in case of failure of inference components, redundant stateful failover between multiple distributed inference nodes, and fail-secure defaults - on failure of model inference, default to high-security policy. The AI elements themselves, however, are new types of attack surfaces: model poisoning using corrupt data, as mentioned above, evasion attacks, as reviewed in this section, or attacks to extract the model parameters at inference time to reverse-engineer the model are all viable threats on the integrity of AI-based defence systems.

6. Conclusion

We have provided an extensive analysis with respect to designing an architecture, evaluating the multi-model solution using empirical methods, analyzing the operational performance, and characterizing deployment challenges in an autonomous Cyber Defense framework for Critical Infrastructure systems. The proposed Hybrid AI model that combines convolutional neural networks, bidirectional LSTM networks and a Deep Q-Network response policy has proved to be most efficient in terms of detection performance, achieving a state-of-the-art detection result of 97.6% accuracy and 97.4% F1-score on the CICIDS2017 benchmark dataset, with much less than 3 seconds in autonomous threat response across all threat types evaluated against the human baseline that ranges from 15 seconds to 47 seconds.

The results confirmed the main hypothesis; using multi-paradigm AI integration (deep learning and reinforcement learning) has a significant positive effect on improving cyber defense performance compared to using any AI paradigm alone. The results from the reinforcement learning convergence analysis also support the use of autonomous policy learning for cyber defense in simulated critical infrastructure systems, as the PPO agent converged in just 600 episodes.

The identified challenges are adversarial robustness hardening, integration of federated learning for privacy preserved collaborative defense, formal explainability frameworks for accountable autonomous decision making and aligning regulations with sector specific regulatory requirements like NERC CIP and IEC 62443. Questions of the ethical frameworks for the deployment of autonomous AI in safety-critical environments, including accountability structures and transparency requirements along with human oversight mechanisms are a research frontier that needs collaboration among AI researchers, cybersecurity practitioners, policy makers, and ethicists.

In the rapidly evolving cyber threat landscape for critical infrastructure operators, AI-powered autonomous defense systems such as those outlined in this paper are not just a step-up from existing solutions but are poised to revolutionize the defense landscape one increasingly dominated by increasingly sophisticated machine-driven adversaries. The shift therefore towards autonomous AI cyber defence is not only, technically, desirable but operationally necessary, to guarantee the protection of essential services without which modern society would not exist.

References

[1] Chatterjee, S. (2024). Understanding the risk of implementing A.I. for managing NERC BCSI Data Repository and security baseline to secure the BCSI data. Journal of Information Systems Engineering and Management, 9(4), 1–14. https://doi.org/10.52783/jisem.v9i4.19

[2] Chatterjee, S., & Malaraju, S. K. (2023). The role of a highly monitored repository for storing architectures and build documents in the NERC environment. International Journal of Innovative Research in Engineering & Multidisciplinary Physical Sciences, 11(2). https://doi.org/10.37082/ijirmps.v11.i2.232146

[3] Kakaraparthi, G. C. (2024). Integrating serverless architectures and Kubernetes for scalable and high-availability AI workflows. International Journal of Intelligent Systems and Applications in Engineering, 12(4), 5896–5905. https://doi.org/10.6084/m9.figshare.30445046

[4] Kubam, C. S. (2024). Agentic AI Microservice Framework for Deepfake and Document Fraud Detection in KYC Pipelines. Journal of Information Systems Engineering and Management, 9(1). https://doi.org/10.5281/zenodo.18009551

[5] LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. Nature, 521(7553), 436–444. https://doi.org/10.1038/nature14539

[6] Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., ... & Hassabis, D. (2015). Human-level control through deep reinforcement learning. Nature, 518(7540), 529–533. https://doi.org/10.1038/nature14236

[7] Buczak, A. L., & Guven, E. (2016). A survey of data mining and machine learning methods for cyber security intrusion detection. IEEE Communications Surveys & Tutorials, 18(2), 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

[8] Ferrag, M. A., Maglaras, L., Moschoyiannis, S., & Janicke, H. (2020). Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study. Journal of Information Security and Applications, 50, 102419. https://doi.org/10.1016/j.jisa.2019.102419

[9] Nguyen, T. T., & Reddi, V. J. (2021). Deep reinforcement learning for cyber security. IEEE Transactions on Neural Networks and Learning Systems, 34(8), 3779–3795. https://doi.org/10.1109/TNNLS.2021.3121870

[10] Xin, Y., Kong, L., Liu, Z., Chen, Y., Li, Y., Zhu, H., ... & Wang, C. (2018). Machine learning and deep learning methods for cybersecurity. IEEE Access, 6, 35365–35381. https://doi.org/10.1109/ACCESS.2018.2836950

[11] Sommer, R., & Paxson, V. (2010). Outside the closed world: On using machine learning for network intrusion detection. In 2010 IEEE Symposium on Security and Privacy (pp. 305–316). IEEE. https://doi.org/10.1109/SP.2010.25

[12] Sharafaldin, I., Lashkari, A. H., & Ghorbani, A. A. (2018). Toward generating a new intrusion detection dataset and intrusion traffic characterization. In Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018) (pp. 108–116). https://doi.org/10.5220/0006639801080116

[13] Mirsky, Y., Doitshman, T., Elovici, Y., & Shabtai, A. (2018). Kitsune: An ensemble of autoencoders for online network intrusion detection. In Proceedings of the Network and Distributed Systems Security Symposium (NDSS 2018). https://doi.org/10.14722/ndss.2018.23204

[14] Sutton, R. S., & Barto, A. G. (2018). Reinforcement learning: An introduction (2nd ed.). MIT Press.

[15] Goodfellow, I., Bengio, Y., & Courville, A. (2016). Deep learning. MIT Press. http://www.deeplearningbook.org

[16] Langner, R. (2011). Stuxnet: Dissecting a cyberweapon. IEEE Security & Privacy, 9(3), 49–51. https://doi.org/10.1109/MSP.2011.67

[17] Lee, R. M., Assante, M. J., & Conway, T. (2016). Analysis of the cyber attack on the Ukrainian power grid. SANS ICS/E-ISAC. https://www.nerc.com/pa/CI/ESISAC/Documents/E-ISAC_SANS_Ukraine_DUC_18Mar2016.pdf

[18] Varga, S., Brynielsson, J., & Franke, U. (2021). Cyber-threat perception and risk management in the Swedish financial sector. Computers & Security, 105, 102239. https://doi.org/10.1016/j.cose.2021.102239

[19] Apruzzese, G., Colajanni, M., Ferretti, L., Guido, A., & Marchetti, M. (2018). On the effectiveness of machine and deep learning for cyber security. In 2018 10th International Conference on Cyber Conflict (CyCon) (pp. 371–390). IEEE. https://doi.org/10.23919/CYCON.2018.8405026

[20] Alom, M. Z., Taha, T. M., Yakopcic, C., Westberg, S., Sidike, P., Nasrin, M. S., ... & Asari, V. K. (2019). A state-of-the-art survey on deep learning theory and architectures. Electronics, 8(3), 292. https://doi.org/10.3390/electronics8030292

[21] Deng, L., & Yu, D. (2014). Deep learning: Methods and applications. Foundations and Trends in Signal Processing, 7(3–4), 197–387. https://doi.org/10.1561/2000000039

[22] Kim, J., Nian, L., Lim, B., Bae, J. W., & Ahn, G. J. (2020). CPS attack detection under limited local information in cyber security: A multi-node multi-class classification approach. ACM Transactions on Cyber-Physical Systems, 5(2), 1–24. https://doi.org/10.1145/3424672

[23] Zhang, J., & Zulkernine, M. (2006). Anomaly based network intrusion detection with unsupervised outlier detection. In 2006 IEEE International Conference on Communications (Vol. 5, pp. 2388–2393). IEEE. https://doi.org/10.1109/ICC.2006.255127

[24] Vinayakumar, R., Alazab, M., Soman, K. P., Poornachandran, P., Al-Nemrat, A., & Venkatraman, S. (2019). Deep learning approach for intelligent intrusion detection system. IEEE Access, 7, 41525–41550. https://doi.org/10.1109/ACCESS.2019.2895334