Grenze International Journal of Engineering and Technology, June Issue

Sustainable AI for Cyber Defense: A Multi-Objective Benchmark

Hrishabh Bhandari[1], Jayant Ranjan Jha[2], Amit Kumar[3], Rohit Kumar[4], Devendra Gautam[5] and Neda Ahmad[6]
[1-6]Dept. of Computer Science & Engineering, JIMS Engineering Management Technical Campus, Greater Noida, Uttar Pradesh, India
Email: hrishabh07042005@gmail.com, jayantranjanjha986@gmail.com, amitkumar312765@gmail.com, kumarrkrohit.1324@gmail.com, devendragautam.gn@jagannath.org, neda.gn@jagannath.org

***Abstract*— Intrusion Detection Systems (IDS) are extensively used in modern networking, to ensure safety over the internet. These systems are aggressively switching towards machine learning approaches to improve their performance, which in turn increases computational demands for deploying such systems. Machine learning approaches also require robustness evaluation so that minor changes in data do not threaten the IDS. Most evaluations today focus only on improving performance, neglecting the robustness and efficiency prospects of these systems. This study aims to benchmark supervised learning models for these multi-objective tradeoffs. It evaluates models based on three axes: classification performance using standard performance metrics, robustness under Gaussian noise injection, and efficiency using temporal and memory metrics. The results highlight tree-based models as strong performers with reasonable efficiency, while neural nets exhibit more stable performance under variance at higher computational costs. These findings put into light the importance of multi-criteria evaluation for practical and sustainable AI-driven cyber defense.**

***Index Terms*— Intrusion Detection System, Cyber Defense, Machine Learning, Robustness, Efficiency, Sustainability.**

I. INTRODUCTION

Cyber attacks are increasingly becoming a major issue in modern networking. Their rapid growth in volume, diversi-ty, and strength puts networks in a highly vulnerable zone [1], [2]. Modern day networks create huge amounts of traffic every single day. Monitoring these using rule-based systems is becoming more impractical and unfeasible [3], [4]. As a result, machine learning-based cyber security systems show rapid growth due to their ability to identify and classify unseen threat patterns with high accuracy. Recent research has been focusing on improving their detection performance even further [5], [6], [7].

Most Intrusion Detection System (IDS) studies show enhanced model performance in terms of accuracy, F1-score, and detection rate, on various datasets [8], [9], [10]. However, such high performance models are generally highly computationally expensive as well. They require huge amounts of compute which is impractical for many network-ing scenarios and unfeasible at large scale deployment [11], [12]. Also, models are only evaluated under ideal test conditions and not under noisy inputs which resemble real world data more closely [13]. Models trained on ideal conditions may lack and falter on noisy inputs making them unfavorable for deployment [14].

Grenze ID: 01.GIJET.12.2.513_20

Further, computation-al efficiency of such models are often completely neglected giving an unfair overall comparison [15].

In real world, cyber security systems operate continuously which require enormous amounts of computation and energy [11], [16]. If models are expensive, they would engulf majority of the budget just to operate. If they are not given adequate computation, their non-ideal performance will in turn threaten networks. Also, robust behavior of models under uncertain and noisy data is critical for reliable cyber defense systems. Thus, there is an uprising need to balance cyber security models on three dimensions: Performance, Robustness, and Efficiency.

Existing work generally focus on optimizing single objective, typically model performance. Other dimensions involv-ing robustness and efficiency are often ignored, and if evaluated, they are recorded in isolation to each other [9], [13], [15]. Such studies often fail to capture the trade-offs between these properties. Thus, there is limited guidance on selecting model under deployment constraints. This raises a need for a unified benchmark on all three perspec-tives.

This work aims to produce a framework that assess cyber defense models jointly on performance, robustness, and efficiency, which gives a fair comparison of overall model sustainability under identical conditions. It analyzes trade-offs rather than optimizing any isolated feature and gives objective-based guidance on model selection under specific deployment constraints.

The study evaluates classical machine learning models which are generally used in cyber defense systems. All mod-els are trained and tested using the same dataset, with identical preprocessing and handling to ensure fairness. Model performance is evaluated using standard classification metrics. Robustness is evaluated by injecting controlled noise perturbations that imitate real world condition. Model efficiency is measured using temporal and memory metrics. These evaluations allow a fair comparison of models parallel.

The remaining paper is organized in the following manner. Section II reviews previous related works on machine learning-based cyber security systems with sustainability considerations. Section III describes the proposed bench-marking architecture. Section IV presents the experiment results and analysis. Section V discusses observations and trade-offs. Section VI concludes the paper and discusses future directions.

II. Literature Review

A. Performance Evaluation in Intrusion Detection Systems

Performance is undoubtedly the most important aspect of an IDS model. Good performance ensures reliable data security over networks. This aspect is accomplished perfectly by machine learning models that display strong per-formance on various benchmark datasets. Studies report that classical and deep learning models like Support Vector Machines, Logistic Regression, Random Forests, Convolutional Neural Networks, Recurrent Neural Networks, and Long Short Term Memory Networks display high accuracy and F1-scores on datasets such as CIC-IDS2017, UNSW-NB15, and Edge-IIoT [5], [6], [7], [8], [9], [10].

However, most of these studies evaluate models only under idealized conditions which give an incomplete compari-son between models. Real world data includes characteristics such as class imbalances and data variability which are often neglected. Thus, these performance metrics do not translate to real world deployability completely [8], [9], [13].

B. Robustness Evaluation in Intrusion Detection Systems

Robustness is a critical aspect in development of an IDS. Several studies evaluated robustness against adversarial attacks like FGSM, PGD, and DeepFool [14]. Many others used noise-based perturbations and zero-day attacks [11], [13]. These studies report that many machine learning models that perform well under ideal conditions falter when data becomes noisy or distribution shifts. This is due to some machine learning models being highly susceptible to variance in data [6], [14].

Researchers treat robustness as a secondary trait rather than a core property. Thus, robustness studies have no stand-ard practices for evaluation, which makes it difficult to compare robustness across models [5], [9].

C. Efficiency Evaluation in Intrusion Detection Systems

Computational efficiency is an equally crucial requirement for the deployment of IDS. Resource constraints directly impact a system's feasibility and performance. Heavy and expensive machine learning models are impractical to be deployed at large scale [11], [15]. Multiple studies claim sustainable models, but rarely explicitly mention efficiency metrics such as training time, inference latency, memory, energy consumption, or model complexity [12], [16].

Recent studies have focused on improving sustainability metrics using techniques like feature reduction and adaptive optimization [11], [15]. These led to reduced GPU utilization and energy consumption without significant reduction in detection accuracy. But efficiency studies are still evaluated in isolated environments and require performance and robustness integration [7], [11].

III. PROPOSED SYSTEM

This study proposes a benchmarking framework for evaluating and comparing different machine learning-based IDS models under identical conditions and network traffic dataset. The objective is to analyze performance, robustness, and efficiency trade-offs to guide objective-based IDS development and deployment.
Fig. 1 describes the framework step by step, in order

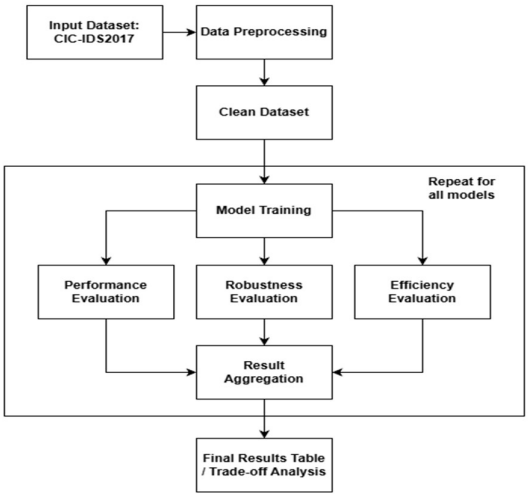

Figure 1. Proposed model benchmarking architecture

A. Dataset Preparation

The widely used public benchmark dataset CIC-IDS2017, collected over multiple days, is used. First, the data is merged into a single data frame for consistent analysis of all scenarios. All column names are standardized and cleaned using standard practices and redundant attributes were removed.
Then the label column is cleaned and multiple fine-grained attack categories were mapped to fewer, broader attack classes to reduce fragmentation. The "infiltration" class, containing only 36 samples, is removed due to extreme class imbalance. Infinite values are replaced with NaN, which are subsequently replaced using medians of respective fea-tures. Features with zero variance are removed as they contribute no additional information for evaluation and only increase computational costs.
This clean dataset is then split into stratified training and testing sets using an 80:20 ratio to preserve the class propor-tions. After splitting, winsorization is applied using the 1st and 99th percentiles only on the training dataset, to prevent information leakage. An unscaled version of the dataset is kept for the ensemble models, while a standardized ver-sion is maintained for scale-sensitive models.

B. Model Training and Evaluation Framework

After data preprocessing, multiple supervised machine learning models were trained and evaluated to compare effec-tiveness for IDS. A set of models including Logistic Regression, Random Forest, XGBoost, and a two-layer Multi-Layer Perceptron is selected. These models represent linear, ensemble, gradient boosted, neural network approaches, thus allowing a diversified comparison under identical settings. All models were trained on the same, respectively scaled versions of the dataset to ensure an unbiased comparison.
Model performance was evaluated using multiple metrics to correctly assess performance under class imbalance, typical to network intrusion datasets. The metrics include Macro-F1 and balanced accuracy for equal importance among classes, and Macro-AUC and Micro-AUC to capture per-class and overall detection capability. This provides an all-inclusive assessment of classification performance, and not just inflated accuracy.

To test the models' robustness, controlled noise perturbations were introduced to replicate real-world variability in network traffic. Gaussian noise was injected into feature values at three noise levels, with standard deviations of 0.05, 0.10, and 0.20. Each model's performance was captured at each noise level. Robustness is then calculated by measuring relative drop in Macro-F1 and AUC under increasing noise. Finally, an aggregate robustness score is cal-culated based on percentage change in Macro-F1 and AUC, giving an overall stability comparison under noise.

Efficiency is evaluated by measuring training time in seconds, inference latency in milliseconds per 1,000 samples, and peak memory during training and trained model size in megabytes. These metrics capture real resource con-straints often considered for deploying IDS.

IV. RESULTS

This section reports the experimental results obtained from the evaluated models.

A. Performance Evaluation Results

Table I shows the recorded performance results of the evaluated models.

TABLE I. PERFORMANCE RESULTS

Model	Macro-F1	Balanced Accuracy	Macro-AUC	Micro-AUC
Logistic Regression	0.665	0.973	0.997	0.997
Random Forest	0.965	0.950	0.999	1.000
XGBoost	0.971	0.996	1.000	1.000
MLP (2-layer)	0.791	0.763	0.999	1.000

XGBoost achieved the highest Macro-F1 and balanced accuracy, while Logistic Regression exhibited lower Macro-F1.

B. Robustness Evaluation Results

Table II shows the recorded robustness results of the evaluated models.

TABLE II. ROBUSTNESS RESULTS

Model	σ = 0	σ = 0.05	σ = 0.10	σ = 0.20	Aggregate Ro-bustness Score
Logistic Regression	0.665	0.640	0.585	0.485	0.8504
Random Forest	0.965	0.140	0.129	0.128	0.0757
XGBoost	0.971	0.185	0.155	0.140	0.0638
MLP (2-layer)	0.791	0.791	0.766	0.679	0.9344

Fig. 2 shows the combined stability curves of all models under increasing noise levels

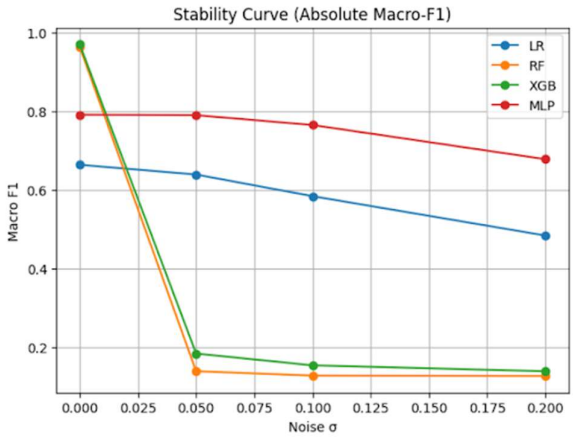

Figure 2. Combined stability curve

Logistic Regression and MLP exhibit smooth and minimal performance degradation, while Random Forest and XGBoost drop performance sharply when introduced to noise.

C. Efficiency Evaluation Results

Table III shows the recorded efficiency metrics of the evaluated models.

TABLE III. EFFICIENCY METRICS

Model	Train Time (sec)	Inference Time (ms/1k)	Peak RAM (MB)	Size (MB)
Logistic Regression	375	0.79	1862	0.005
Random Forest	3064	43.47	4038	270.207
XGBoost	1710	93.50	3532	6.761
MLP (2-layer)	173	1.74	3399	0.074

Logistic Regression shows highest efficiency across all metrics, while Random forest takes more time and memory. XGBoost takes reasonable amount of time and exhibits highest inference time.

V. DISCUSSION

The experiment results clearly display trade-offs among performance, robustness, and efficiency of evaluated mod-els.

XGBoost exhibits the highest Macro-F1 and balanced accuracy, with Random Forest closely catching up. This dis-plays their ability to handle class imbalances and complex decision boundaries well. Logistic Regression exhibits comparatively much lower Macro-F1 indicating poor performance in capturing network threats. The MLP model achieves moderate Macro-F1 and balanced accuracy indicating average threat detection performance.

Robustness evaluations on the other hand display contrasting results. MLP and Logistics Regression exhibit relatively stable performance across increasing noise, with smaller drops in Macro-F1. Both the ensemble methods show sharp performance drop under noise, indicating high sensitivity to input perturbations. This highlights that simpler decision boundaries may offer better resistance to noise-based perturbations. This creates a significant accuracy-robustness trade-off.

In terms of efficiency, Logistic Regression and MLP display significantly lower training time, inference latency, and memory consumption. This makes them suitable for deployment under resource-constrained environments. Random Forest displays the highest time, memory and disk size consumption across all models, which highlights their limiting nature. XGBoost exhibits reasonably average efficiency metrics.

Considering all three dimensions simultaneously, no single model dominates across all three dimensions. While, XGBoost ensures high predictive performance with reasonable computational costs, it is fragile to robustness issues. Logistic Regression provides a cheap and efficient model with strong robustness at the cost of reduced performance. Random Forest shows strong performance but poor robustness and expensive computation. Finally, MLP gives a balance of all three dimensions with moderate performance and strong robustness with comparatively cheaper re-quirements.

VI. CONCLUSION

This study presented a multi-objective benchmark of machine learning models for IDS. It evaluates performance, robustness, and efficiency within unified experimental conditions. The results of the study explicitly show that a model cannot be evaluated solely on the basis of its performance. Robustness to input noise and computational costs are critical to a complete evaluation.

This emphasizes that model development must be suitable to deployment and sustainability constraints and not just performance. Simpler models may offer better robustness and efficiency, but complex models offer significantly better performance at increased costs.

Overall, the study showed the importance of multi-objective evaluation in design and deployment of sustainable models for cyber defense systems. Future work can extend this study to explore hybrid or adaptive approaches to balance these dimensions, or integrate other robustness analysis techniques such as adversarial perturbations.

REFERENCES

[1] M. H. Hamzah, “An Advanced Security System for Detecting Cyber-Attacks in Computer Networks Based on Machine Learning,” Al-Rafidain J. Eng. Sci., vol. 4, no.1, pp. 74-89, 2026.

[2] L. Kendrick, “Research on Machine Learning-Based Threat Detection and Security Defense Systems for Telecom Operator Networks,” Huaxia Xinzhi, vol. 2, no.1, pp. 9-19, 2026.

[3] S. Panta and S. Pandey, “Artificial Intelligence for Cyber Threat Detection in Government and Enterprise Sys-tems,” in Proc. Conf. Cyber Threat Detection, 2024.
[4] R. Ram and S. Pal, “Review Paper on Autonomous Detection of IoT Botnet Attacks Using Deep Learning Tech-nique,” Int. J. Adv. Res. Multidiscip. Trends, vol. 3, no.1, pp. 44-52, 2026.
[5] N. Purwaha, “Comparative Study of Cybersecurity AI Models: Performance, Safety, Trust, and Explainability,” Int. J. Eng. Technol. Res. Manag., vol. 10, no.1, pp. 199-205, 2026.
[6] J. Wilkie, H. Hindy, C. Michie, C. Tachtatzis, J. Irvine, and R. Atkinson, “A Novel Contrastive Loss for Zero-Day Network Intrusion Detection,” IEEE Trans. Netw. Serv. Manag., 2026.
[7] Y. Li, H. Wang, and G. Xu, “Federated Reinforcement Learning-Driven Multi-Task Optimization for Robust and Ethical Edge Internet of Things Security,” Sci. Rep., in press, 2026.
[8] C. B. Sabdana, N. D. Salyasari, I. N. Z. Aliya, and A. M. Shiddiqi, “Multi-Task Temporal Deep Learning Model for Real-Time Intrusion Detection System,” JUTI J. Ilm. Teknol. Inf., vol. 24, no. 1, pp. 149-164, 2026.
[9] O. Bhulakshmi, M. Anusha, R. Somesh, S. Badrinath, M. M. Gopal, and P. T. Khan, “Adaptive Machine Learning for Real-Time Intrusion Detection in IoT,” in Proc. ICRDICCT’25, 2025, pp. 654-662.
[10] D. C. Nijhum, F. B. F. Ananna, M. H. Khan, and M. N. Uddin, “Enhancing Intrusion Detection in Cyber-Physical Systems: A Hybrid Graph Neural Network and Explainable AI Approach for Classification,” in Proc. IEEE COMPAS, 2025.
[11] T. D. Nguyen, A. Alazab, A. Khraisat, and T. Jan, “Feature Reduction in Federated Learning for Intrusion De-tection in IoT Networks,” Cybersecurity, vol. 9, no. 102, 2026.
[12] A. Bashir, J. Iqbal, A. Elahi, and A. Akram, “ADAPT-IDS: Self Learning Anomaly Detection with Dynamic Thresholding and SHAP Interpretability,” Spectrum Eng. Sci., vol. 4, no. 1, pp. 159-172, 2026.
[13] S. Sajid, J. Iqbal, and A. Akram, “An Adaptive Zero-Day Intrusion Detection and Traffic Classification Frame-work for IoT Networks Using Autoencoder-Based Anomaly Isolation,” Spectrum Eng. Sci., vol. 4, no. 1, pp. 201-209, 2026.
[14] J. Zhu, Z. Chen, R. Cong, H. Sun, and Y. Dong, “STS-AT: A Structured Tensor Flow Adversarial Training Framework for Robust Intrusion Detection,” Sensors, vol. 26, no. 536, pp. 1-21, 2026.
[15] A. V. Aanandaram, P. Deepalakshmi, and V. P. Rajendaran, “Quantum-Assisted Hybrid Security Framework for Secure Authentication with Intrusion Detection for 5G Networks,” J. Theor. Appl. Inf. Technol., vol. 104, no. 1, pp. 455-460, 2026.
[16] E. Obunezi and N. Nwiabu, “A Model for Emerging Threats and User Behaviour Variability Detection in Smart Homes Using Hybrid Technique,” Int. J. Comput. Sci. Math. Theory, vol. 11, no. 9, pp. 1-11, 2025.