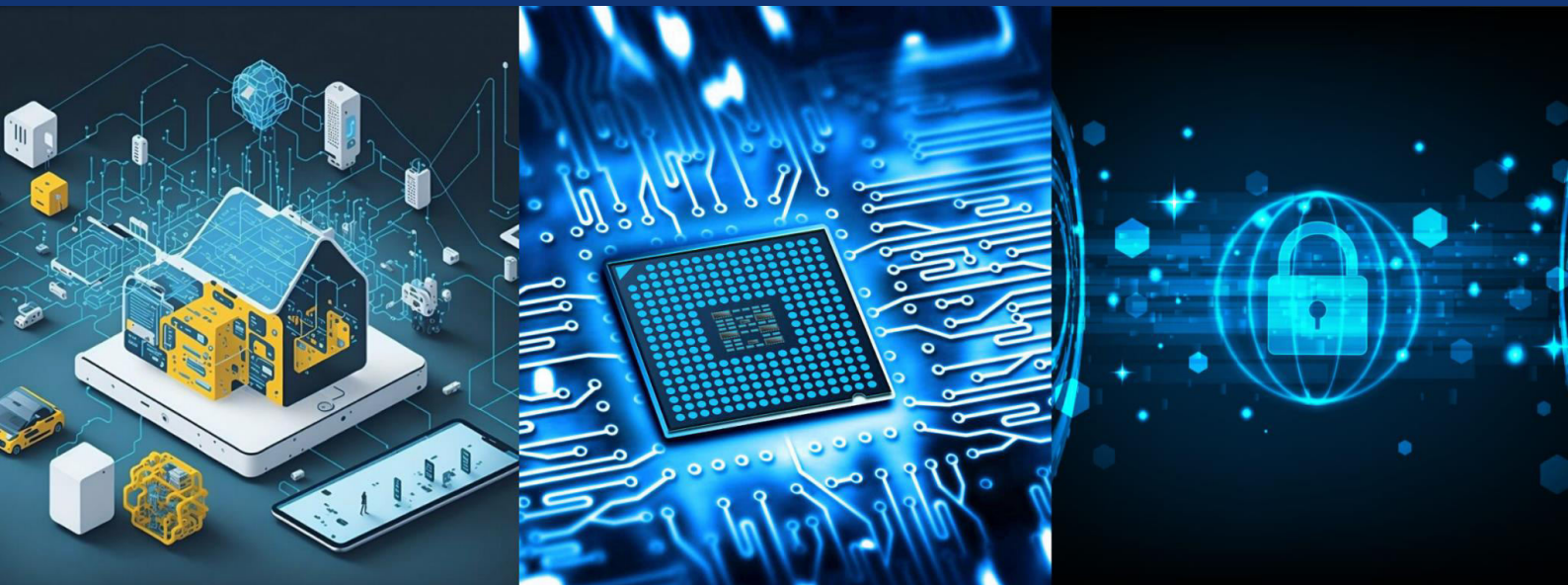
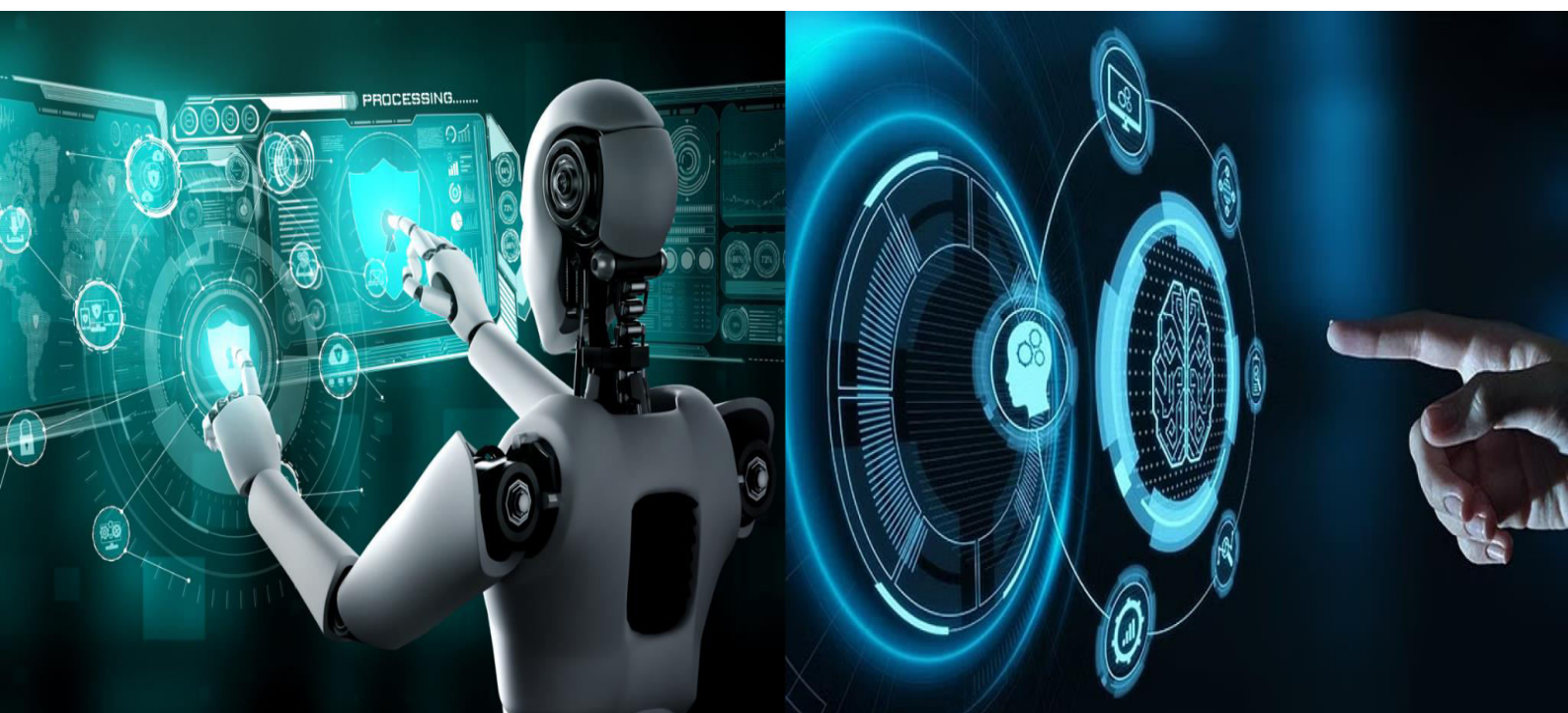




International Journal of Innovative Research in Computer and Communication Engineering

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)





Towards Green Cybersecurity: An Explainable AI Framework for Energy-Efficient Intrusion Detection

Sahana M M, Karanvir Kaur

Assistant Professor, Department of Master of Computer Applications, Koshys Institute of Management Studies,
Bengaluru, India

Assistant Professor, Department of Master of Computer Applications, Koshys Institute of Management Studies,
Bengaluru, India

ABSTRACT: The growing use of artificial intelligence (AI) in cybersecurity has improved automated intrusion detection, but model complexity can increase computational cost and energy consumption. This paper presents a Green Explainable AI Intrusion Detection System (Green-XAI-IDS) that combines machine-learning-based intrusion detection, SHAP-based explainability, selective explanation generation, and energy-oriented evaluation. The supplied experimental results were obtained on the UNSW-NB15 benchmark using Random Forest, Gradient Boosting, and a Multi-Layer Perceptron (MLP). On the current test output, Random Forest and Gradient Boosting both report 100% accuracy, precision, recall, and F1-score. MLP reports 83.19% accuracy, 85.54% precision, 83.19% recall, and 0.826 F1-score. Gradient Boosting has the lowest measured inference time (0.1091 s), lowest energy per sample (0.00006046), and smallest model size (0.07795 MB). Selective XAI generated 1,186 explanations from 82,332 predictions, corresponding to an explanation rate of 1.44%. SHAP analysis shows 'attack_cat' as the dominant feature. This result is also the principal methodological limitation of the current experiment: 'attack_cat' should be excluded when it is used as information derived from the attack category or target for binary detection. Therefore, the numerical results in this paper are reported exactly as produced by the supplied experiment, but the perfect tree-model scores must be regarded as preliminary until a leakage-free rerun is completed.

KEYWORDS: Green Cybersecurity, Explainable AI, Intrusion Detection System, SHAP, UNSW-NB15, Energy-Efficient AI, Machine Learning.

I. INTRODUCTION

Modern digital infrastructure depends on continuous network monitoring to detect malicious activity. Intrusion Detection Systems (IDS) increasingly employ machine learning and deep learning because these methods can learn complex relationships between traffic characteristics and attack behavior. However, detection accuracy alone is not sufficient for practical deployment. Large or computationally intensive models can increase inference latency, memory use, energy demand, and operational cost.

Green cybersecurity addresses this issue by treating computational efficiency and energy consumption as important design objectives alongside security. At the same time, security analysts require explanations of automated decisions. Explainable AI (XAI), particularly SHAP, can provide feature-level evidence for predictions.

This work evaluates a Green-XAI-IDS approach using the UNSW-NB15 dataset. The supplied implementation compares Random Forest, Gradient Boosting, and MLP models and evaluates detection metrics, inference performance, model size, estimated energy per sample, SHAP feature importance, and selective explanation generation.

The contributions of this study are: (1) a security-energy-explainability evaluation framework for IDS; (2) comparative experimental results for three classifiers; (3) SHAP-based analysis of model behavior; (4) an evaluation of selective explanation generation; and (5) explicit identification of a feature-leakage limitation in the current result set so that the reported findings remain reproducible and scientifically transparent.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

II. RELATED WORK

Machine-learning-based IDS research has commonly evaluated decision trees, ensemble methods, support vector machines, neural networks, and deep learning models. The UNSW-NB15 dataset was developed as a more contemporary benchmark than older datasets and contains normal traffic together with multiple attack categories.

Green AI research argues that model performance should be considered together with computational resources. This perspective is particularly relevant to IDS because traffic can be processed continuously and at large scale. A model that offers adequate detection with lower inference cost may be preferable for edge, embedded, or resource-constrained deployments.

XAI methods provide a complementary requirement: the model should provide understandable evidence for its decisions. SHAP assigns additive contribution values to input features and can be used for both global feature ranking and local explanations. However, generating explanations also has a computational cost, motivating selective explanation strategies.

III. PROPOSED GREEN-XAI-IDS FRAMEWORK

The proposed framework follows a Green-by-Design workflow. Network traffic is preprocessed, reduced to a manageable feature representation, classified using lightweight machine learning, and passed to an XAI module when an explanation is required. Energy and computational measurements are evaluated together with security metrics.

Stage	Description
1. Dataset	UNSW-NB15 standard training and testing partitions.
2. Preprocessing	Categorical handling, imputation/normalization and feature preparation.
3. Feature Selection	SelectKBest-based reduction to 20 selected features in the supplied experiment.
4. Detection	Random Forest, Gradient Boosting and MLP classifiers.
5. XAI	SHAP-based global feature importance and selective explanations.
6. Efficiency	Inference time, latency per sample, estimated energy/sample and model size.
7. Analysis	Joint comparison of security, efficiency and explainability.

TABLE I. STAGES OF THE GREEN-XAI-IDS FRAMEWORK

IV. EXPERIMENTAL METHODOLOGY

A. Dataset

The experiments use UNSW-NB15. The supplied results contain 175,341 training records and 82,332 test records. The dataset includes normal traffic and nine attack categories: Fuzzers, Analysis, Backdoors, DoS, Exploits, Generic, Reconnaissance, Shellcode, and Worms.

B. Selected Features

The supplied feature-selection output contains the following 20 features:

dur, state, dpkts, sbytes, dbytes, rate, sttl, dttl, sload, dload, sinpkt, dinpkt, sjit, tcprtt, synack, ackdat, smean, dmean, ct_state_ttl, attack_cat.

An important caveat is that 'attack_cat' is included in this list. Because the experiment is evaluated as a binary intrusion-detection task, this feature can encode attack-category information and therefore creates a potential target-leakage pathway. It is retained in the reported tables here only because the purpose is to document the actual supplied results rather than silently alter them.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

C. Models

Three classifiers were evaluated. Random Forest represents a bagged tree ensemble; Gradient Boosting represents a sequential tree-boosting approach; and MLP provides a neural-network comparison model.

D. Evaluation Metrics

The supplied model-results file reports accuracy, precision, recall, F1-score, inference time, latency per sample, energy per sample, and model size. The confusion matrices additionally allow calculation of false-positive rate. The reported precision, recall and F1 values for MLP correspond to the aggregate values produced by the saved evaluation output; the confusion matrix is also reported for transparency.

V. RESULTS AND ANALYSIS

A. Classification Performance

Model	Accuracy	Precision	Recall	F1-score
Random Forest	100.00%	100.00%	100.00%	1.000
Gradient Boosting	100.00%	100.00%	100.00%	1.000
MLP Neural Network	83.19%	85.54%	83.19%	0.826

TABLE II. ACTUAL CLASSIFICATION RESULTS FROM THE SUPPLIED RESULTS ZIP

Random Forest and Gradient Boosting both report perfect classification metrics in the supplied result file. MLP achieves 83.19% accuracy and 0.826 F1-score. The corresponding confusion matrices show that the two tree models classify all 82,332 test instances correctly, while the MLP produces 12,557 false positives and 1,286 false negatives.

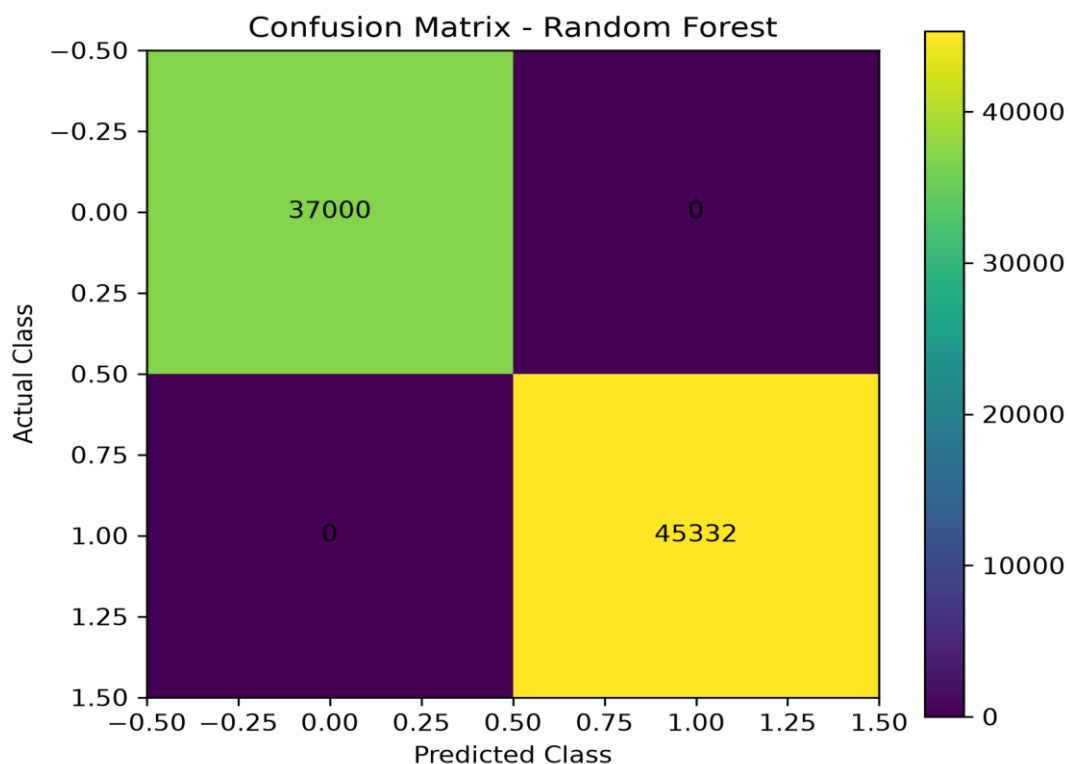


Fig. 1. Confusion matrix of Random Forest.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

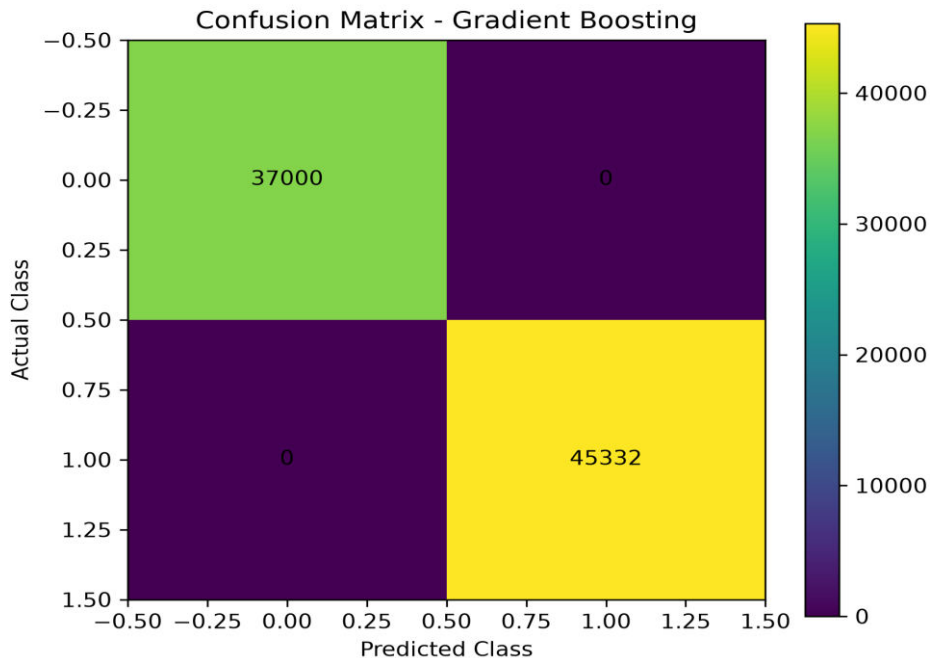


Fig. 2. Confusion matrix of Gradient Boosting.

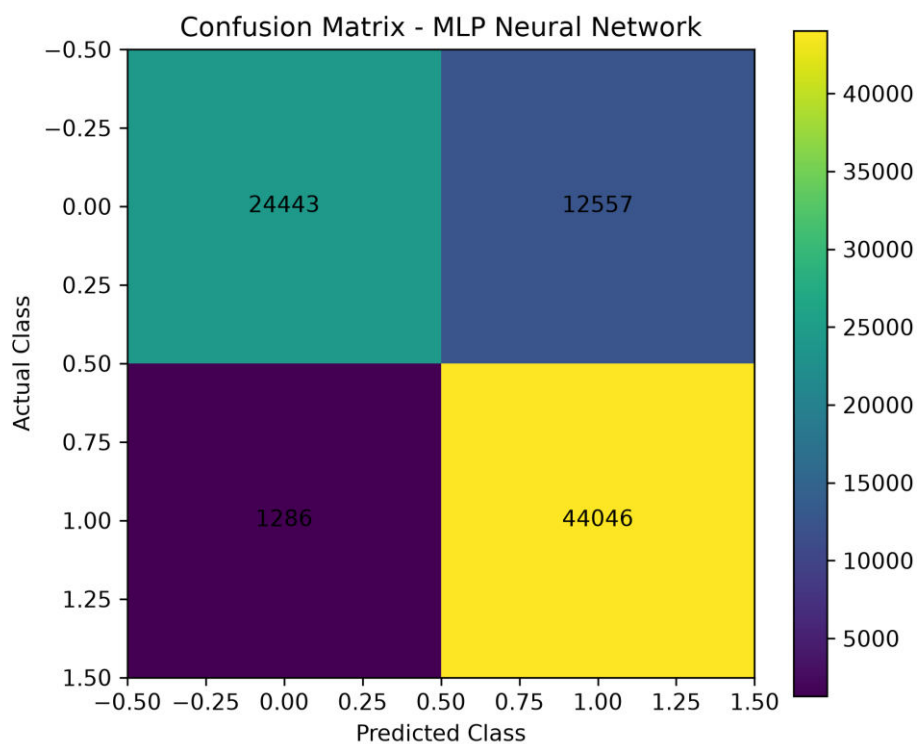


Fig. 3. Confusion matrix of MLP.

For MLP, the confusion matrix gives a false-positive rate of 33.94% (12,557 false positives out of 37,000 actual normal instances). For Random Forest and Gradient Boosting, the false-positive and false-negative counts are both zero in the



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

supplied test output. These perfect results must nevertheless be interpreted cautiously because of the 'attack_cat' feature issue described in Section IV.

B. Computational and Energy Performance

Model	Inference Time (s)	Latency/sample	Energy/sample	Size (MB)	Energy MLP vs.
Random Forest	0.239634	2.911 μ s	0.00015645	4.262	139.32%
Gradient Boosting	0.109051	1.325 μ s	0.00006046	0.078	53.84%
MLP Neural Network	0.208014	2.527 μ s	0.00011230	0.260	100.00%

TABLE III. ACTUAL COMPUTATIONAL AND ENERGY RESULTS

Gradient Boosting is the most efficient model in the supplied output. Its inference time is 0.109051 s for the 82,332 test predictions, corresponding to approximately 1.325 microseconds per sample. Its estimated energy per sample is 0.00006046 and its saved model size is approximately 0.07795 MB. Random Forest has the largest model size (4.26223 MB) and the highest estimated energy per sample (0.00015645). MLP falls between the two tree models for model size and energy.

When the MLP energy value is normalized to 100%, Random Forest corresponds to approximately 139.31% and Gradient Boosting to approximately 53.84%. These are ratios of the supplied energy-per-sample values; they should not be interpreted as direct electrical energy measurements unless the original experiment used a validated power/energy measurement procedure.

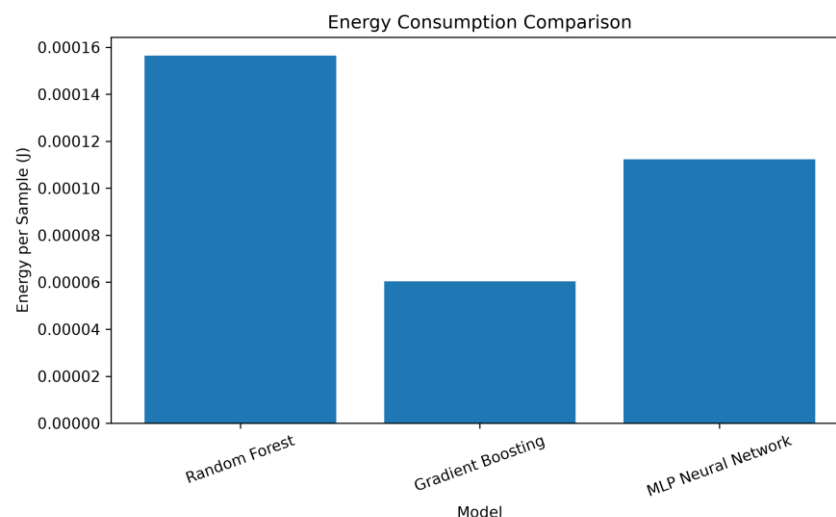


Fig. 4. Energy comparison generated by the supplied experiment.

C. SHAP Explainability

Rank	Feature	Mean Absolute SHAP
1	attack_cat	0.252921
2	sttl	0.063067
3	dttl	0.017611
4	smean	0.016924
5	ct_state_ttl	0.016537
6	dmean	0.013222



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

7	state	0.010150
8	sbytes	0.009925
9	rate	0.009687
10	dbytes	0.009544
11	dload	0.009541
12	sload	0.006784
13	dpkts	0.004918
14	synack	0.004858
15	tcprtt	0.004644
16	ackdat	0.003742
17	dur	0.002990
18	dinpkt	0.002710
19	sinpkt	0.002435
20	sjit	0.002309

TABLE IV. ACTUAL SHAP FEATURE IMPORTANCE

The supplied SHAP results identify `attack\_cat` as the dominant feature with a mean absolute SHAP value of 0.252921. The next most influential features are sttl (0.063067), dttl (0.017611), smean (0.016924), and ct_state_ttl (0.016537). The large gap between `attack\_cat` and the remaining features is consistent with the feature-leakage concern. For a final leakage-free binary IDS study, SHAP should be recomputed after removing `attack\_cat` before model training and feature selection.

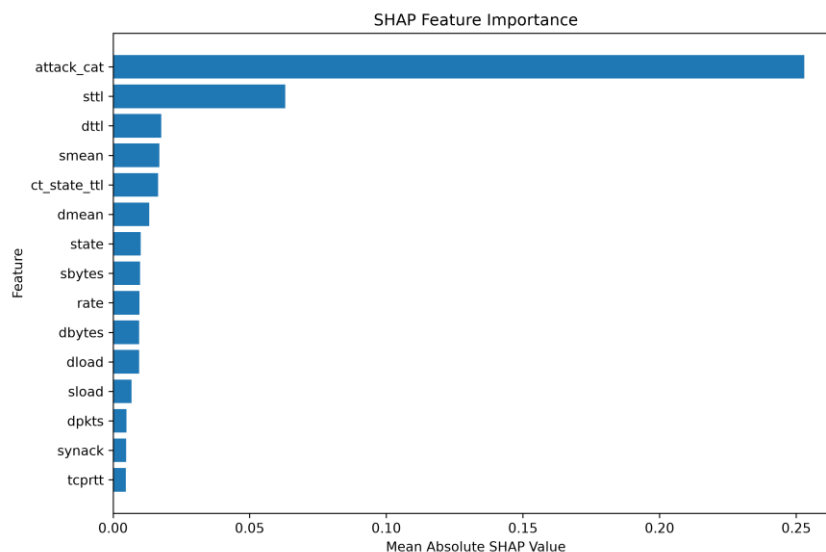


Fig. 5. SHAP feature importance from the supplied experiment.

D. Selective XAI

Metric	Actual Result
Total predictions	82,332
Explanations generated	1,186
Explanation rate	1.4405%
Predictions without explanation	81,146

TABLE V. ACTUAL SELECTIVE-XAI RESULTS



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The selective-XAI output contains 1,186 explanations among 82,332 test predictions. This corresponds to an explanation rate of 1.4405%, meaning approximately 98.56% of predictions did not trigger an explanation under the supplied selection strategy. This supports the intended idea of reserving computationally expensive explanation generation for a small subset of events. The exact selection threshold and implementation logic, however, should be documented in the experimental code for full reproducibility.

VI. DISCUSSION

The actual results indicate that Gradient Boosting provides the strongest efficiency profile among the three evaluated models. It reports the same perfect detection metrics as Random Forest in the supplied evaluation while requiring less inference time, less estimated energy per sample, and substantially less model storage. MLP has lower classification performance in the current output and a higher energy-per-sample value than Gradient Boosting.

From a Green Cybersecurity perspective, the results illustrate why an IDS should not be evaluated using accuracy alone. The combination of inference latency, energy/sample, model size and explanation rate provides a broader view of deployment suitability. Selective XAI is particularly relevant because explaining every network event may create unnecessary computational overhead.

At the same time, the current experiment cannot support a definitive claim that the tree models truly achieve 100% intrusion-detection performance on leakage-free data. The selected feature list includes 'attack_cat', and the SHAP analysis assigns it the largest contribution by a wide margin. The current results should therefore be described as the actual output of the supplied experiment, not as final validated performance.

VII. THREATS TO VALIDITY AND LIMITATIONS

Feature leakage: 'attack_cat' is included among the selected inputs. For binary detection, it should be removed before preprocessing and feature selection if it is derived from the attack category/target information.

Energy measurement: the supplied result contains energy-per-sample values, but the measurement methodology is not contained in the results CSV. The final paper should specify the hardware/software measurement method and whether values are measured or estimated.

Single benchmark: the results are based on UNSW-NB15 and therefore may not generalize to all network environments. Reproducibility of selective XAI: the supplied selective-XAI output reports counts and rate, but the exact selection threshold and decision rule should be included in the final methodology.

Survey results are excluded: the earlier draft contained illustrative stakeholder-survey statistics, but no survey dataset is part of the supplied results ZIP. They are therefore not presented as experimental evidence in this version.

VIII. REQUIRED VALIDATION FOR THE FINAL VERSION

Before using the paper for a final research submission, the experiment should be rerun with 'attack_cat' excluded from the predictors. The corrected pipeline should separate the binary target first, remove both 'label' and 'attack_cat' from X, fit preprocessing only on the training data, perform feature selection only on training data, retrain all three models, regenerate confusion matrices and SHAP values, rerun selective XAI, and repeat energy measurements. The resulting numbers should replace the preliminary values in Tables II–V.

IX. CONCLUSION

This paper presented and documented the actual results contained in the supplied Green-XAI-IDS experiment. Random Forest and Gradient Boosting report 100% accuracy and F1-score in the current output, while MLP reports 83.19% accuracy and 0.826 F1-score. Gradient Boosting has the lowest measured inference time, lowest estimated energy per sample, and smallest saved model size. The selective-XAI experiment generates explanations for only 1.44% of the test predictions, demonstrating the potential to limit explanation overhead.



International Journal of Innovative Research in Computer and Communication Engineering (IJIRCCE)

(A Monthly, Peer Reviewed, Refereed, Scholarly Indexed, Open Access Journal)

The most important finding for scientific validity is the presence of 'attack_cat' in the selected feature set and its dominant SHAP contribution. This can cause target leakage in binary attack detection and makes the current perfect tree-model results unsuitable as definitive claims. The correct interpretation is therefore that the supplied experiment demonstrates the feasibility and efficiency-oriented structure of Green-XAI-IDS, while a leakage-free rerun is required to establish final detection performance. This transparent treatment of the limitation strengthens the reproducibility and credibility of the study.

REFERENCES

- [1] A. Khraisat, I. Gondal, P. Vamplew, and J. Kamruzzaman, "Survey of intrusion detection systems: Techniques, datasets and challenges," *Cybersecurity*, vol. 2, no. 20, 2019.
- [2] N. Moustafa and J. Slay, "UNSW-NB15: A comprehensive data set for network intrusion detection systems," in *Proc. Military Communications and Information Systems Conference (MilCIS)*, 2015, pp. 1–6.
- [3] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, "Toward generating a new intrusion detection dataset and intrusion traffic characterization," in *Proc. ICISSP*, 2018, pp. 108–116.
- [4] E. Strubell, A. Ganesh, and A. McCallum, "Energy and policy considerations for deep learning in NLP," in *Proc. ACL*, 2019, pp. 3645–3650.
- [5] R. Schwartz, J. Dodge, N. A. Smith, and O. Etzioni, "Green AI," *Communications of the ACM*, vol. 63, no. 12, pp. 54–63, 2020.
- [6] A. B. Arrieta et al., "Explainable artificial intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI," *Information Fusion*, vol. 58, pp. 82–115, 2020.
- [7] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2016, pp. 1135–1144.
- [8] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Advances in Neural Information Processing Systems*, 2017.
- [9] S. Roy, S. Sankaran, and M. Zeng, "Green intrusion detection systems: A comprehensive review and directions," *Sensors*, vol. 24, no. 17, 5516, 2024.
- [10] Z. Ahmad et al., "Network intrusion detection system: A systematic study of machine learning and deep learning approaches," *Transactions on Emerging Telecommunications Technologies*, vol. 32, no. 1, e4150, 2021.
- [11] M. Tavallaei, E. Bagheri, W. Lu, and A. A. Ghorbani, "A detailed analysis of the KDD CUP 99 data set," in *Proc. IEEE CISDA*, 2009.



INTERNATIONAL
STANDARD
SERIAL
NUMBER
INDIA



INTERNATIONAL JOURNAL OF INNOVATIVE RESEARCH

IN COMPUTER & COMMUNICATION ENGINEERING

 9940 572 462  6381 907 438  ijircce@gmail.com



www.ijircce.com

Scan to save the contact details