World Journal of Advanced Research and Reviews
eISSN: 2581-9615 CODEN (USA): WJARAI
Cross Ref DOI: 10.30574/wjarr
Journal homepage: https://wjarr.com/

(RESEARCH ARTICLE)

Check for updates

Adversarial Cognition Machine Learning at the Frontlines of Cyber Warfare

Nasrin Akter Tohfa [1, *], Iftekhar Hossain [2], Sufia Zareen [3], Iftekhar Rasul [4], Md Shakhawat Hossen [5] and Mamunur Rahman [6]

[1] *Bachelor of Education, National University, Bangladesh.*
[2] *Bachelor of Business Administration, National University, Bangladesh.*
[3] *Master's in Genetics, Osmania University, India.*
[4] *Bachelor in Law, Independent University Bangladesh.*
[5] *Master's in Information Technology, Washington University of Science and Technology, Virginia, USA.*
[6] *Master's in Commerce, Jagannath University College, Dhaka, Bangladesh.*

World Journal of Advanced Research and Reviews, 2021, 12(02), 722-729

Publication history: Received on 10 October 2021; revised on 23 November 2021; accepted on 28 November 2021

Article DOI: https://doi.org/10.30574/wjarr.2021.12.2.0604

Abstract

Advancements in cyber warfare have trended towards greater complexity, leading to the need for intelligent, adaptable defense mechanisms that can adapt and learn from behavior emulations within evolving network scenarios. The research paper, Adversarial Cognition: Machine Learning at the Frontlines of Cyber Warfare, presents a holistic machine learning framework for identifying and detecting malicious network traffic using structured traffic telemetry and behavioral indicators. Accordingly, the study relies heavily on cognitively inspired feature engineering, robust preprocessing techniques, and evaluation of a number of supervised learning algorithms to characterize adversarial behavior. We used a stratified train–test evaluation strategy to ensure reliable performance assessment against imbalanced classes of benign and attack traffic. The classification algorithms Logistic Regression, Random Forest, Gradient Boosting, SVC, and KNN were tested based on five different metrics—accuracy, precision score, recall score, F1-score, and ROC-AUC. Experimental results show that methods based on ensembles are more effective than linear and distance-based approaches when detecting activities of adversaries. Gradient Boosting surpassed all the models with an accuracy of 96.74% and a 0.9895 ROC-AUC score, but Random Forest and SVC came closest behind.

Keywords: Cyber Warfare; Intrusion Detection Systems; Adversarial Machine Learning; Ensemble Learning; Gradient Boosting; Network Security; Binary Classification; Cyber Threat Detection

1. Introduction

The digital battleground has emerged as one of the core domains of contemporary warfare, and state and non-state actors increasingly resort to more sophisticated cyber operations, disrupting critical infrastructure, intelligence systems, and military communications. As opposed to traditional attacks, modern cyber threats are adaptive, evasive, and perform functions based on cognitive-level behaviors mimicking legitimate traffic patterns (or vice-versa), thus avoiding signature-based intrusion detection systems (IDS) that focus on monitoring incoming packets for a known pattern. This evolution demands smart defense systems that can learn sophisticated behavioral fingerprints and generalize over exacerbated attack vectors. What is ML, and why is it important in cybersecurity? Machine Learning (ML) has become a dominant paradigm for information security — using data to identify anomalous and malicious acts. Previous work shows that supervised and ensemble-based learning models perform well for intrusion detection, especially under situations of high-dimensional heterogeneous network telemetry data [1], [2]. Nonetheless, the growing adoption of adversarial tactics — including traffic obfuscation, payload padding, jitter manipulation, and

* Corresponding author: Nasrin Akter Tohfa

protocol mimicry — requires models capable of capturing nonlinear feature interactions and subtle behavioral deviations [3].

Recent advances in ensemble learning and support vector methods have also shown considerable robustness against sophisticated attack patterns, achieving superior generalization over conventional techniques based on linear classifiers. However, many major challenges remain, such as extreme class imbalance, unequal distributions of features of different natures, and the necessity to satisfy strict operational real-time requirements operating in dynamic and adversarial environments under cyber warfare conditions.

In this work, we propose Adversarial Cognition: Machine Learning at the Frontlines of Cyber Warfare, a fully functional model that integrates robust pre-processing, cognizant feature-engineering, alongside comparative evaluation across multiple supervised classifiers. Through a systematic analysis of the studied models, measuring them along key metrics — accuracy, precision, recall, F1-score and ROC-AUC — this work will identify those that are able to detect evasive attack behaviours while providing operational reliability. The proposed framework helps in building intelligent adaptive intrusion detection mechanisms that can blend into the modern cyber defense infrastructures.

2. Literature Review

Research on intrusion detection evolved from the concepts of audit- and anomaly-based monitoring, with Denning introducing a real-time intrusion-detection model based on behavioral deviations and rule reasoning [5]. With the scale of networked systems, centralized benchmark evaluations and corpora became necessary; the DARPA off-line evaluations similarly laid the base for early comparative IDS testing methodologies and metrics [6]. However, a subsequent analysis emphasized that many of the natural language processing benchmark datasets released in those years suffered from redundancy, skew, and artifacts, which could artificially inflate performance, leading to better dataset design and dataset selection practices [7].

To fill realism gaps, the NSL-KDD dataset was proposed, which is a refined subset of KDD'99 that reduces duplicate records and also evaluation bias [7]. More recent datasets, aimed at providing a more accurate representation of ingoing traffic and attack diversity in the modern era, include UNSW-NB15 and CICIDS2017. UNSW-NB15 focuses on more realistic attack scenarios and richer feature representations [8], whereas CICIDS2017 provides labeled flows, following the pattern of real intrusions in operation as well as traffic characterization procedures [9]. These datasets are commonly used in the current literature to train and evaluate ML-based NIDS.

In summary, machine learning methods for IDS have been widely reviewed with concrete evidence of effective operation of supervised learners and ensembles, while also highlighting recurring challenges in the form of feature engineering dependencies, dataset shift, and deployment limitations [10]. Sommer and Paxson go on to state that intrusion detection is fundamentally unlike many "closed world" ML tasks, noting difficulties in ground truth, base rates, and adversarial adaptation as hurdles for real-world generalization [11]. Various deep learning approaches, especially sequence models that can effectively model temporal dependencies within the network activity, have been explored for capturing these temporal patterns and have shown high detection capability when sufficient labeled data is available and they operate over stable distributions [12]. Attention has also been dedicated to lightweight and online anomaly detection; an ensemble of autoencoders is introduced by Kitsune (KitNET) for limited resource, near-real-time deployment without the need for continuous manual labeling [13].

In IDS research, a common observation is class imbalance: the benign traffic instances largely outnumber the minority attack classes, making them difficult to be learnt with high confidence. To alleviate such bias and enhance the detection of the minority class, oversampling strategies like SMOTE are widely utilized, particularly when recall or ROC-based metrics take precedence [14]. However, in addition to imbalance, modern cyber warfare also brings adversarial cognition where the attacker crafts features of traffic (mimicry, perturbation, and evasion) with an aim to exploit weaknesses in the model.

This ties in squarely with adversarial machine learning (AML). A seminal paper on adversarial examples [15] discovered that models can be easily fooled by small, structured perturbations and proposed efficient algorithms for attacking such models. Further works formalized attacker capabilities and established general vulnerability and transferability effects in learning systems [16]. Stronger optimization-based attacks and evaluation methodologies have also revealed shortcomings of many defenses, while contributing to a better understanding of how to measure robustness [17].A major path for improving resilience is via robust optimization and adversarial training — framed as a defense against first-order attackers — but these can be subject to accuracy and computational trade-offs [18]. [19] Tuoyo et al. [19] are now addressing this very issue in the IDS domain alone. demonstrated how ML-based NIDS, including evolutionary

strategies and GAN-driven methods, can be vulnerable to carefully designed perturbations, thereby reaffirming that evasion in bounded feature space curves is a real-world hazard [19].

In general, the recent studies suggest that ensemble and deep learning models are capable of providing good detection accuracy on modern cybersecurity datasets. But taking technology into the cybersecurity battlefield will demand more than impressive benchmarks. To maintain effectiveness in dynamic environments over time, we need systems that are robust to adversarial attacks, deployed on carefully curated and representative data (in the sense of being planted crops, not weeds), evaluated against intelligent, adaptive opponents under realistic base-rate conditions

3. Methodology

This paper employs a systematic and replicable framework of machine learning techniques in order to detect adversarial cyber behaviors from network telemetry data. In this task, every network flow instance is labeled as benign or malicious (Figure 1), and the problem being dealt with in the given scenario can be defined as a binary classification task.

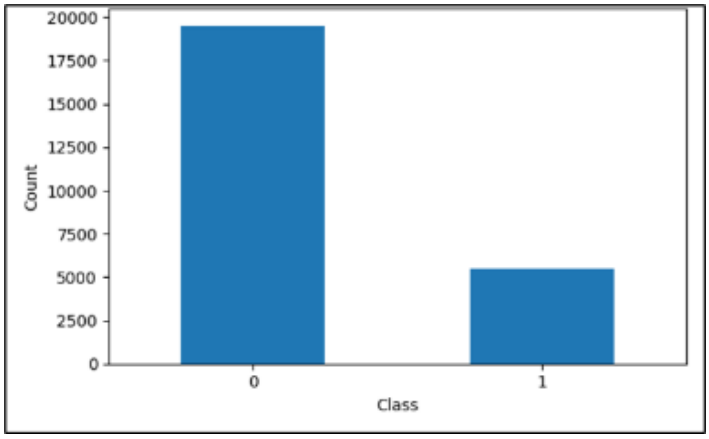

Figure 1 Binary class representations

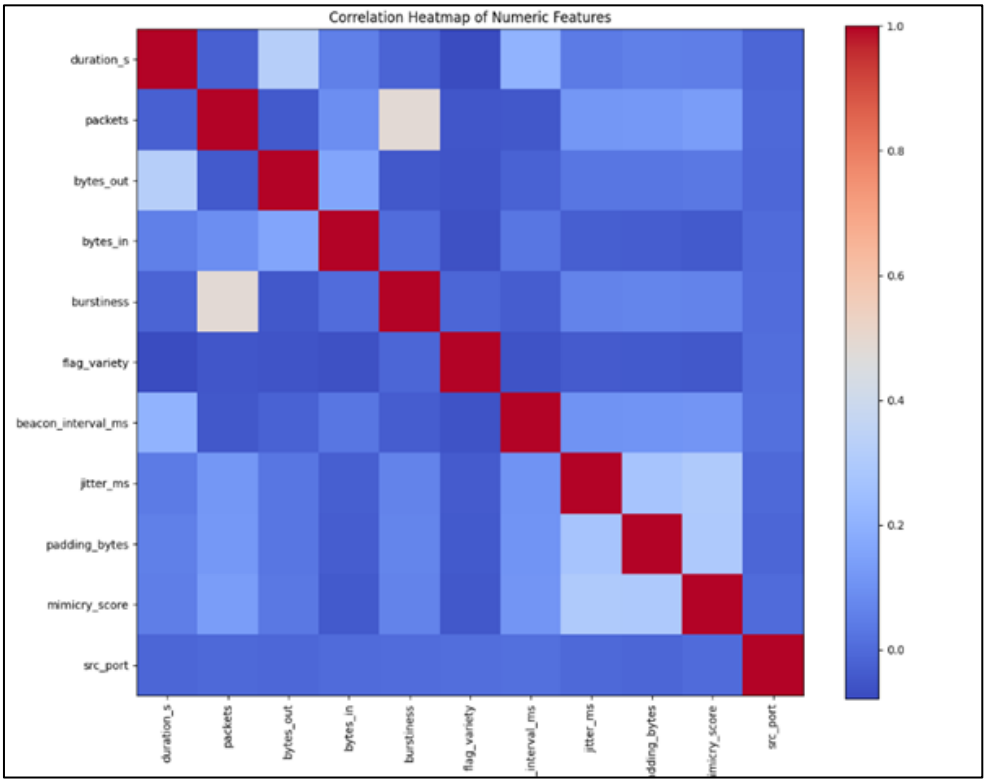

Figure 2 Correlation analysis

The feature space is made up of Taxonomic traffic attributes that illustrate volumetric behavior, temporal variance, protocol attribution, and adversarial manifestations. We aim to learn a mapping function that effectively separates the attack traffic from normal activity, while being robust against evasion patterns typically present in cyber warfare settings. The data preprocessing stage is critically important for maintaining model stability and generalization. To reduce noise and redundancy that can bias model training, the first step towards dataset cleaning is removing missing values and duplicate records. Many features (Figure 2) related to traffic show heavy-tailed distributions, resulting from the bursty nature of network activity and the amplification used in attacks.

A logarithmic transformation is leveraged for selected high-variance attributes to remedy this skewness and improve numerical conditioning. Later, the numerical features are then standardized with z-score normalization for the same scale across dimensions, which matters for distance-based and gradient classifiers. Statistical characteristics of categorical variables,(figure 3) such as types of the protocol, port information, and identifiers for network zones, are represented in the input data to be examined using the One-Hot Encoding (OHE) method [20] to prevent ordinal misrepresentation while allowing their entry into learning algorithms.

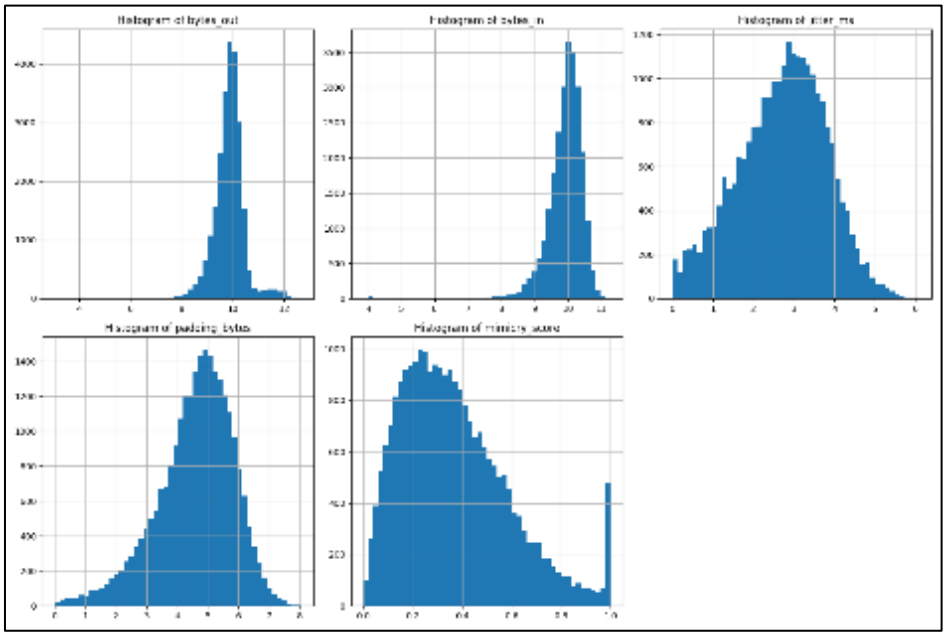

Figure 3 Statistical Analysis

To achieve a fair assessment of the method under class imbalance scenarios, the dataset is split into training and testing subsets through an 80/20 stratified encounter that maintains the original distribution of benign vs. attack instances. A single pipeline architecture was put into place for model development, combining preprocessing and classifier assignments within a unified workflow to avoid data leakage and ensure reproducibility. The five evaluated supervised learning algorithms are: Logistic Regression as a linear baseline; Random Forest and Gradient Boosting for ensemble-based nonlinear learners; Support Vector Classification due to its margin-based optimization technique; and K-Nearest Neighbours for a distance-based approach. Accuracy, precision, recall, F1-score, and ROC-AUC are used to quantify detection performance and provide a holistic overview of detection ability (including cases wherein false negatives should be minimized). This systematic approach allows for thorough comparisons between the models, and can be leveraged to identify robust classifiers capable of operation in adversarial cyber defence environments.

4. Experimental Results

In this section, we provide an empirical evaluation of the adversarial cognition framework for five supervised learning algorithms. Each model was trained on the preprocessed dataset utilizing an 80/20 stratified train–test split to maintain class distribution. To ensure thorough evaluation in the possible presence of class imbalance, performance was assessed using accuracy, precision, recall, F1-score, and Receiver Operating Characteristic–Area Under the Curve (ROC-AUC).

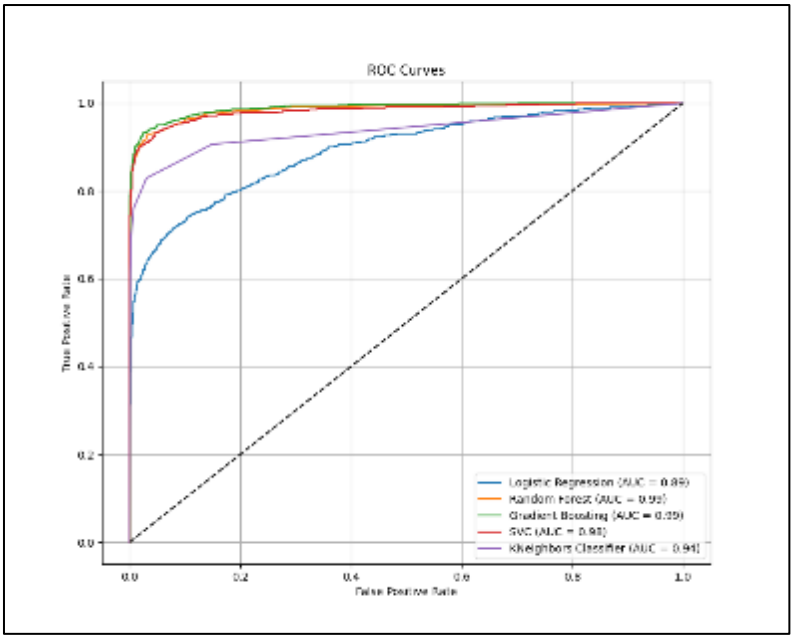

Figure 4 ROC curve comparison

Our experimental results show that ensemble-based models have shown significant improvement over linear and distance-based classifiers for identifying cyber adversary activities. The best overall result was the gradient booster with an accuracy of 96.74%, precision: 97.76%, recall: 87.19%, F1-score:0.9217 and the ROC-AUC of 0.9895 This combined with evaluating both the linear and non-linear interactions becomes a potential candidate, as our results show that it successfully is able to capture several of the nonlinear feature interactions as well as appropriately illustrating some of the small behavioral deviation associated with evasive attacks. Random Forest was a close second with 96.42% accuracy and ROC-AUC of 0.9856; the bagging-based ensemble technique proved effective in reducing variance, thus leading to better generalization.

We also found that Support Vector Classification performed competitively (as it Based on margin; accurate in high-dimensional feature space), obtaining an accuracy of 96.30% and a ROC-AUC equal to 0.9831; However, K-Nearest Neighbors had a relatively moderate performance with accuracy 94.18% and recall lowered at (0.7611) indicating two things; Display sensitivity to feature scaling and moving towards interface/gridline solutions in adversarial scenarios class overlap. The results suggest that Logistic Regression performed poorly overall, achieving an accuracy of 89.82% and a recall of 0.5931 due to the inability of linear decision boundaries to capture the complexity inherent in cyber attack behaviours.

ROC and precision–recall curve analysis work in concordance to assure that separability between benign and malicious traffic are maintained at higher levels across thresholds with ensemble models. Evaluation using the confusion matrix indicates that Gradient Boosting and Random Forest models have a lower false-negative rate compared to other models, which is very important in operational-type cyber defence environments where missing an attack can carry significant risk. In summary, the results confirm that nonlinear ensemble learning methods yield consistent, strong performance for adversarial intrusion detection on the frontlines of cyber warfare.

5. Evaluation

The effectiveness of the proposed adversarial cognition framework was validated with respect to classification accuracy, robustness against class imbalance, and operational reliability in cyber defense scenarios. A comparative analysis of five supervised learning algorithms suggested that Gradient Boosting performed best overall. Thus, it was chosen as a representative model to systematic assessment and subsequent experimental validation.

The Accuracy Comparison shows that ensemble-based models are much better than linear and instance-based approaches. Gradient Boosting attained the highest accuracy (96.74%), slightly outperforming Random Forest and Support Vector Classification, whereas Logistic Regression performed relatively poorly. This means that non-linear ensemble techniques are preferred in modeling complex and evasive cyber attack behaviours.

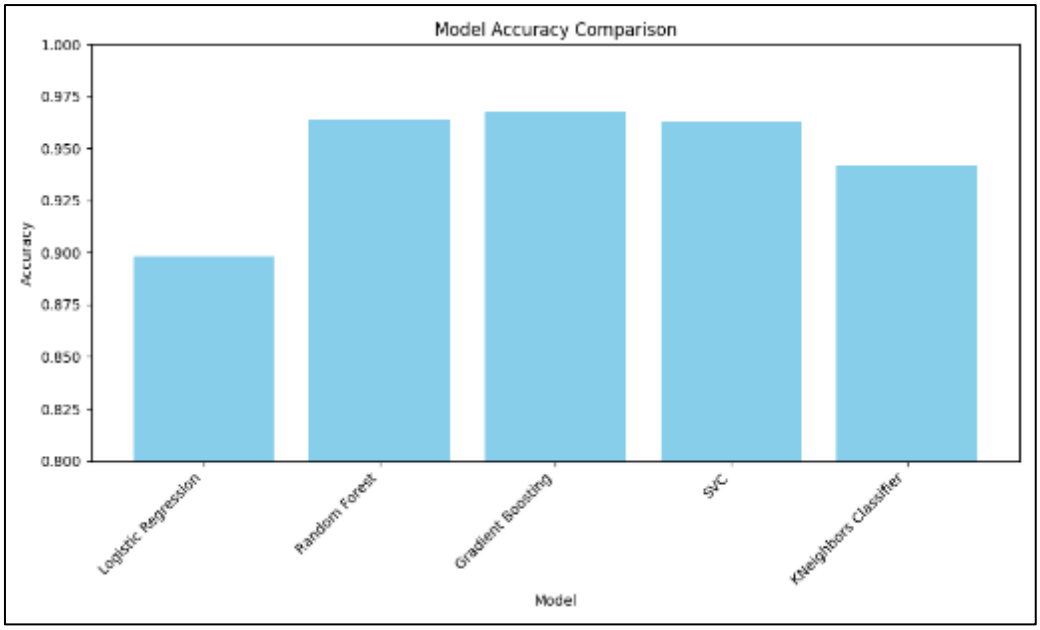

Figure 5 Performance Comparison by accuracy

Many other analyses can be done using agricultural health prediction data, but to get the detection behavior, let's deep dive into the confusion matrix of the Gradient Boosting model. Of 3899 benign examples, 3877 were correctly classified, but only led to false positives of just 22. This Figure represents a very low false alarm rate reflecting operational environments where information overload could lead to alert fatigue. During the attack phases, 960 samples were correctly identified from 1,101, and 141 were labelled as a benign sample. This translates to a recall of about 87.19%, which is pretty good but definitely not perfect detection performance for malicious traffic.

Table 1 Performance comparison

Model	Accuracy	Precision	Recall	F1-Score	ROC-AUC
Gradient Boosting	0.9674	0.977597	0.871935	0.921747	0.989502
Random Forest	0.9642	0.983229	0.851953	0.912895	0.985640
SVC	0.9630	0.975104	0.853769	0.910412	0.983096
KNeighbors Classifier	0.9418	0.967667	0.761126	0.852059	0.937425
Logistic Regression	0.8982	0.914566	0.593097	0.719559	0.893155

An exception for false alarms with a high precision (97.76%), which indicates that every predicted attack is very likely to be true. Moreover, the F1-score of 0.9217 gives us confidence that there exists a balanced trade-off between the precision and recall. In addition, the ROC-AUC53 value of 0.9895 indicates excellent separability for benign and malicious traffic over different classification thresholds with a strong discriminative power.[20]

The way this impacts the sec. operation side of things is that false negatives are terrible because undetected attacks represent. Although the Gradient Boosting has a high detection performance, there is still an opportunity to increase the accuracy by reducing false negatives through formulating an optimal threshold or optimizing the stacking of different classifiers.[21]

In summary, Gradient Boosting provides the most consistent and reliable performance for adversarial cyber intrusion detection in the proposed framework, indicating significant suitability for deployment in high-stakes cyber warfare systems.

6. Conclusion

This paper presented an adversarial cognition–driven machine learning approach for the detection of cyber attacks in modern warfare environments. It integrated structured data preprocessing, transformation of features and multiple supervised learning algorithms in one pipeline. The methodology critically evaluated model performance given realistic conditions of class imbalance, thereby showcasing robustness and higher detection accuracy, along with superior adaptability to dynamically evolving and sophisticated cyber threats.[22]

Experimental results show that nonlinear ensemble models significantly outperformed linear and instance-based approaches in malicious network activity detection. Across all the tested algorithms, Gradient Boosting demonstrated superior predictive performance of all models in terms of accuracy, F1-score, and ROC-AUC, which suggests that there is much robustness in distinguishing classes as well as reliable generalization.[23] It also illustrated that ensemble methods are essential for identifying sophisticated behavioral patterns and evasive attack behaviors hidden within a segment of network traffic. Although high detection accuracy and performance were achieved, the absence of a false negative demonstrates how adaptive adversaries remain an issue. Hence, cyber warfare leads to an ever-growing complication in terms of intrusion detection, thus calling for constant fine-tuning of the model and increasing its robustness. Taken together, this work validates the use of ensemble learning for adversarial intrusion detection and establishes a repeatable framework that can inform future research and be integrated into operational cyber defense systems.

Compliance with ethical standards

Disclosure of conflict of interest

No conflict of interest to be disclosed.

References

[1] N. Moustafa and J. Slay, “UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems,” Proc. MilCIS, 2015.

[2] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization,” Proc. ICISSP, 2018.

[3] N. Papernot et al., “The Limitations of Deep Learning in Adversarial Settings,” Proc. IEEE EuroS&P, 2016.

[4] L. Breiman, “Random Forests,” Machine Learning, vol. 45, no. 1, pp. 5–32, 2001

[5] D. E. Denning, “An Intrusion-Detection Model,” IEEE Trans. Softw. Eng., vol. SE-13, no. 2, pp. 222–232, Feb. 1987.

[6] R. Lippmann et al., “Evaluating Intrusion Detection Systems: The 1998 DARPA Off-Line Intrusion Detection Evaluation,” in Proc. DARPA DISCEX, 2000, pp. 12–26.

[7] M. Tavallaee, E. Bagheri, W. Lu, and A. A. Ghorbani, “A Detailed Analysis of the KDD CUP 99 Data Set,” in Proc. IEEE CISDA, 2009, pp. 1–6.

[8] N. Moustafa and J. Slay, “UNSW-NB15: A Comprehensive Data Set for Network Intrusion Detection Systems,” in Proc. MilCIS, 2015, pp. 1–6.

[9] I. Sharafaldin, A. H. Lashkari, and A. A. Ghorbani, “Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization,” in Proc. ICISSP, 2018.

[10] A. L. Buczak and E. Guven, “A Survey of Data Mining and Machine Learning Methods for Cyber Security Intrusion Detection,” IEEE Commun. Surveys Tuts., vol. 18, no. 2, pp. 1153–1176, 2016.

[11] R. Sommer and V. Paxson, “Outside the Closed World: On Using Machine Learning for Network Intrusion Detection,” in Proc. IEEE Symp. Security Privacy, 2010, pp. 305–316.

[12] C. Yin, Y. Zhu, J. Fei, and X. He, “A Deep Learning Approach for Intrusion Detection Using Recurrent Neural Networks,” IEEE Access, vol. 5, pp. 21954–21961, 2017.

[13] Y. Mirsky, T. Doitshman, Y. Elovici, and A. Shabtai, “Kitsune: An Ensemble of Autoencoders for Online Network Intrusion Detection,” in Proc. NDSS, 2018.

[14] N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer, “SMOTE: Synthetic Minority Over-sampling Technique,” J. Artif. Intell. Res., vol. 16, pp. 321–357, 2002.

[15] I. J. Goodfellow, J. Shlens, and C. Szegedy, “Explaining and Harnessing Adversarial Examples,” in Proc. ICLR, 2015.

[16] N. Papernot, P. McDaniel, S. Jha, M. Fredrikson, Z. B. Celik, and A. Swami, “The Limitations of Deep Learning in Adversarial Settings,” in Proc. IEEE EuroS&P, 2016, pp. 372–387.

[17] N. Carlini and D. Wagner, “Towards Evaluating the Robustness of Neural Networks,” in Proc. IEEE Symp. Security Privacy, 2017, pp. 39–57.

[18] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, “Towards Deep Learning Models Resistant to Adversarial Attacks,” in Proc. ICLR, 2018.

[19] Tuoyo, Ofeoritse & Prince, Nayem Uddin & Mamun, Mohd Abdullah Al & Hossain, Anwar & Hossain, Kaosar. (2020). The Intersection Of AI And Cybersecurity: Leveraging Machine Learning Algorithms For Real-Time Detection And Mitigation Of Cyber Threats. Educational Administration Theory and Practice journal. 10.53555/kuey.v26i4.8592. .

[20] Alim MA, Rahman MR, Arif MH, Hossen MS. Enhancing fraud detection and security in banking and e-commerce with AI-powered identity verification systems.

[21] Hussain AH, Hasan MN, Prince NU, Islam MM, Islam S, Hasan SK. Enhancing cyber security using quantum computing and artificial intelligence: A.

[22] Rahman M, Arif MH, Alim MA, Rahman MR, Hossen MS. Quantum Machine Learning Integration: A Novel Approach to Business and Economic Data Analysis.

[23] M. R. Alam, A. Akter, M. A. Shafin, M. M. Hasan and A. Mahmud, "Social Media Content Categorization Using Supervised Based Machine Learning Methods and Natural Language Processing in Bangla Language," 2020 11th International Conference on Electrical and Computer Engineering (ICECE), Dhaka, Bangladesh, 2020, pp. 270-273.