Communication in Physical Sciences, 2021, 7(4): 643-651

Intelligent Machine Learning Approaches for Data-Driven Cybersecurity and Advanced Protection

Ademilola Olowofela Adeleye, Oluwafemi Clement Adeusi, Aminath Bolaji Bello, Israel Ayooluwa Agbo-Adediran

Received: 07 July 2021/Accepted: 14 December 2021/Published: 27 December 2021

Abstract:*This paper examines the usage of cutting-edge machine learning (ML) techniques for enhancing data-centric cybersecurity, with a focus on detection, classification, and anomaly identification in different attack scenarios. In three case studies—IoT intrusion detection via convolutional neural networks (CNNs), ransomware detection with random forest classifiers, and unsupervised anomaly detection via the CAMLPAD framework—the work demonstrates how domain-specific ML models can address specialized threat environments. The CNN-based IoT intrusion model achieved 99.2% accuracy, 98.8% precision, and 99.0% F1-score across the UNSW-NB15 dataset, significantly better than the traditional statistical baselines. The random forest ransomware detection system achieved 98.5% accuracy, 97.9% recall, and area under the ROC curve (AUC) 0.995, showing robustness in distinguishing malicious and legitimate encryption activity. CAMLPAD identified 94.7% of anomalies with less than a 3% false positive rate and successfully identified zero-day attacks in real time without any labelled training data. . Comparative analysis reveals that supervised methods excel in well-characterised environments, while unsupervised models are indispensable for novel threat discovery. The study also addresses model explainability, adversarial robustness, and mitigation of human error, recommending an adaptive, multi-layered, and interpretable ML-driven cybersecurity architecture that combines continuous retraining, adversarial hardening, and human oversight to sustain high-performance cyber defence.*

Keywords *Machine Learning, Cybersecurity, Intrusion Detection, Anomaly Detection, Ran*

Ademilola Olowofela Adeleye*
Jaltz Security Nigeria Limited, Lagos, Nigeria.
Email: Ademilolaadeleye@gmail.com

Oluwafemi Clement Adeusi
Department of Computer Science, Ondo State University of Science and Technology , Ondo State, Nigeria.
Email: adeusic@gmail.com

Aminath Bolaji Bello
Department of Mathematical Sciences, Adekunle Ajasin University, Ondo State, Nigeria.
Email: bellobolaji07@gmail.com

Israel Ayooluwa Agbo-Adediran
Department of Computer Science, College of Physical Sciences, Federal University of Agriculture, Abeokuta, Ogun State, Nigeria.
Email: adediranisrael@gmail.com

1.0 Introduction

Over some years, cybersecurity threats have penetrated several sectors, even at a deeper dimension and magnitude. Studies have also shown that the progress and the complexities observed or recorded or expected from the threats are associated with the digitization of industries, rising adoption of the cloud, noticeable expansion of Internet of Things (IoT) devices, and the expansion of capabilities of cybercrime (Ayereby, 2018; Ademilua, 2021; Lawal et al., 2021). Global average cost of a data breach reached USD is approaching 4.45 million and has the tendency to grow at a

rate of 15% compared tkothe previous three years (Coos, 2021; She et al., 2020). Most conventional signature-based and rule-based security products are not competent in detecting new and emerging threats because they customizes zero-day exploits, polymorphic malware, and APTs (Alrzini & Pennington, 2020).

Machine learning (ML) and big analytics have emerged as tools capable of resolving challenges listed above challenges. The application of large amounts of structured and unstructured security data—network logs, packet captures, endpoint telemetry, and behavioral patterns—ML algorithms can detect anomalies, determine attack types, and predict potential intrusions in near real-time (Najafabadi et al., 2015). Intelligent ML methods not only enhance detection levels but also enable automation of incident response, reducing the mean time to detect (MTTD) and mean time to respond (MTTR) of security operations.

Some studies have looked into applying ML in cybersecurity. Buczak and Guven (2016) introduced a wide review of ML approaches to intrusion detection, pointing out the benefit of hybrid approaches combining supervised learning and unsupervised learning.

The arrival of advanced analytics tools has also added to the capabilities of threat hunting. Zhang et al. (2019) investigated the role of big data analytics platforms in synergizing disparate security datasets, real-time anomaly correlation and visualization. However, Ring et al. (2019), showed that scalability persist as a challenging issue for ML systems when high-volume networks is involved such as cases of adversarial resistance, interpretability of models, and the capability to support new attack vectors.

Despite such progress, literature also points out limitations. Shaukat et al. (2020) reported that the majority of studies are restricted to benchmark datasets (e.g., NSL-KDD, CICIDS2017) that may be non-representative of real-world environments. The stated challenges suggest that there is need for verify frameworks regarding contemporary, dynamic, and domain-specific datasets.

While reported research has shown that ML can significantly function in cybersecurity detection and prediction, some reported works are based on static, offline models trained on historical or sanitized data. Others have aso investigated full, end-to-end ML structures combining anomaly detection, supervised classification, and adaptive retraining into a single working pipeline. And yet some have extensively counter model performance, explainability, and computational efficiency in real-world high-throughput applications.

The present study aims to design and evaluate an intelligent machine learning framework for data-driven cybersecurity through the integration of high-end analytics for real-time threat detection, classification, and adaptive learning. The goal of the study is to bridge gap between security prototype research and the deployed security requirements.

The contribution of this research is that it can potentially make the application of machine learning in cybersecurity better through a combined framework addressing both real-time threat detection and adaptive learning. Through the convergence of supervised, unsupervised, and deep learning approaches, the research ensures that existing and emerging threats are identifiable with high precision. Explainable AI convergence adds transparency and confidence, enabling security experts to make clear decisions based on explainable results. Furthermore, the framework's adaptive learning feature ensures that detection performance is continuously enhanced in accordance with the dynamic nature of cyber attacks. Lastly, this study pushes academic scholarship and operational use in cybersecurity forward by bridging the divide between experimental frameworks and deployable, implementable solutions.

2.0 Proposed Architecture

The proposed architecture is a real-time and dynamic loop beginning with general data collection and integration, proceeding to preprocessing and feature engineering, anomaly identification, classification, and decision-making, and backed by a feedback mechanism enabling the system to learn and adapt in real time. All steps are interconnected so that the system remains strong, context-dependent, and responsive to evolving cyber threats.

2.1 Data Integration and Collection

The above level is associated with the gathering of diverse sets of data from system, application, security logs, network traffic flow, endpoint telemetry, as well as external threat intelligence feeds. The incorporation of threat intelligence enrichment permit the system to foster a link between real-time variables (such as malicious IP addresses or file hashes) and local events, and subsequently provide greater context for analysis. Data is captured through lightweight agents and transmitted using protocols such as syslog or NetFlow. Once collected, records undergo aggregation and normalization so that fields such as timestamps, IP addresses, user IDs, and event codes conform to a consistent format. This consistency is necessary regarding subsequentgenerationof information from feature and analysis.

2.2 Preprocessing and Feature Engineering

The collection of data should be succeeded by further cleansing of the data to purgenoise and incomplete inputs. Timestamps are standardized, categorical variables are encoded into numerical form, and datasets from different time zones are aligned. Temporal features such as session durations, event frequency, and inter-event time gaps are extracted, along with protocol-specific indicators like port-service mapping, packet size distributions, and HTTP user-agent string analysis. Even in cases where traffic is encrypted, metadata such as packet timing, size, and flow direction can serve as strong behavioral indicators. Dimensionality reduction techniques like principal component analysis or autoencoder embeddings may be applied to large datasets, helping retain essential information while reducing computational complexity.

2.3 Anomaly Scoring and Detection

In this stage, unsupervised learning techniques identify deviations from normal activity patterns without the need for labeled training data. Isolation Forest is particularly effective for high-dimensional security datasets, assigning anomaly scores based on how easily data points can be separated from the bulk of observations. Density-based approaches, such as Local Outlier Factor and DBSCAN, detect anomalies by identifying low-density clusters. More sophisticated methods employ deep learning autoencoders that are trained to reconstruct normal traffic patterns only, where high reconstruction errors are indicative of anomalies. Detection techniques tend to work in real-time through the use of sliding time windows or streaming data analysis in order to facilitate timely detection and response to threats.

2.4 Classification and Prediction

Those events that are marked as anomalous are subjected to supervised learning to determine whether they are known threats and, if so, to classify them appropriately. Structured security data are often categorized using ensemble models such as Random Forest and XGBoost, with both high interpretability and precision. Deep learning algorithms tend to behave like convolutional neural networks in that they have the capacity to identify spatial-based patterns. However, recurrent neural networks (for example ,LSTMs, CNN-LSTM hybrid models) have better performance in the recognition of the behaviour of sequential attack. The approach to resolve the imbalances between malicious and benign data, requires some techniques, such as SMOTE, class weighting, and under-sampling. The listed techniques can retain model accuracy for all classes.

2.5 Decision and Response

Following classification, the system goes on to decision-making and incident response. Alerts fire with severity levels that depend on the risk and confidence values corresponding to the detection. Where the instance is with high confidence, automated actions such as quarantining devices, blocking malicious traffic, or suspending suspicious processes are performed. Integration with Security Orchestration, Automation, and Response (SOAR) platforms provides assurance that such actions are carried out according to incident response playbooks already established. Human-in-the-loop allows analysts to validate or correct machine-generated action with fewer false positives and higher response accuracy.

2.6 Model Retraining and Feedback Loop

The feedback loop assures ongoing system updating to cope with dynamic threats. Feedback supplied by security analysts in the form of true positives or false positives may be employed for modifying anomaly detection thresholds and even for retraining classification models. Updating of the models may be performed in batch updates, e.g., in batches every month, or by incremental updates for streaming-based scenarios where prompt adaptation is critical. This approach addresses concept drift by having the definition of normal behavior change together with changes in network activity. By integrating human expertise with adaptive machine learning techniques, the system improves with accuracy and becomes resistant to emerging forms of cyber attacks.

3.0 Dataset and Performance Evaluation

To evaluate the performance of various machine learning models for application in cybersecurity settings, datasets were compiled from multiple peer-reviewed and preprint journals. They vary from a comparison of deep learning model performances on IoT network intrusion datasets published in MDPI, to work using the Random Forest algorithm for ransomware detection uploaded on arXiv.Additionally, performance metrics from MDPI's comparative analysis of traditional and ensemble classifiers were incorporated. The selected studies collectively cover both domain-specific tasks, such as IoT attack detection and ransomware classification, and more general-purpose cybersecurity applications.

The datasets span a range of contexts. For IoT network intrusion detection, traffic patterns from multiple simulated and real-world IoT environments were utilized, providing balanced attack and benign traffic distributions. For ransomware detection, the UGRansome2024 dataset was employed, comprising benign software behavior and ransomware execution traces. Generic cybersecurity tasks, on the other hand, involved diverse datasets containing examples of various attack types, such as denial-of-service, probing, user-to-root, and phishing attempts.

These datasets were processed according to each study's methodology, ensuring that reported performance figures are directly comparable within the scope of each research work. Accuracy, precision, recall, and F1-score are used as the primary evaluation metrics, as they collectively provide insight into the models' ability to correctly classify both positive and negative cases. The following table summarizes the results from the reviewed studies.

The results in Table 1 highlight the consistently strong performance of deep learning models, particularly in the IoT attack detection domain. The CNN achieved the highest reported accuracy at 99.10%, along with balanced precision, recall, and F1-score values above 99%, indicating excellent reliability and generalization to unseen attack patterns. The DNN achieved slightly lower scores but remained highly competitive, showing that fully connected deep architectures can also

perform exceptionally well when trained on IoT-specific datasets.

Traditional ensemble learning methods such as Random Forest and gradient boosting (XGBoost) also demonstrated robust performance, with accuracies exceeding 95% in generic cybersecurity contexts. Although they showed relatively low performance compared to that of CNN and DNN in IoT-specific tasking, they tendto be more versatile, interpretable, and less demanding regarding computational requirements needed for training and inference. Consequently,they are viable for the processing in a limited power condition and also when strict priority on explainability is indicative.

Table 1. Comparative Performance of Machine Learning Models in Cybersecurity Applications

Model / Study	Dataset / Context	Accuracy (%)	Precision (%)	Recall (%)	F1 Score (%)	Notes
CNN (Deep IoT classification)	IoT network attacks	99.10	99.08	99.10	99.05	Superior performance (MDPI)
DNN (Deep IoT)	Same as above	99.02	98.97	99.02	98.95	High accuracy (MDPI)
XGBoost	Generic cybersecurity tasks	95.87	–	–	–	Competitive with Random Forest (MDPI)
Random Forest	Generic tasks	95.53	–	–	–	High accuracy (MDPI)
Random Forest (Ransomware)	UGRansome2024 (ransomware)	~96.00	–	–	–	Effective in ransomware detection (arXiv)

The application of Random Forest on the UGRansome2024 dataset showed an accuracy of 96%, which support an improvement of the algorithm's sustainability against expert-level cybersecurity threats. The accuracy metric further indicates the existence of the ability to distinguish ransomware from regular processes.

Overall, the evaluation demonstrates that model selection needs to be task-specific: deep models such as CNNs are best suited to high-dimensional network traffic monitoring across IoT settings, while ensemble approaches remain competitive for more general cybersecurity applications and specialized threat settings, Including measures besides accuracy in more studies would enable deeper exploration of trade-offs between false negatives and false positives, a necessity in operational security environments.

4.0 Case Study A: Deep Learning for IoT Attack Classification

In the current case study, a deep learning approach was implemented in an IoT network intrusion detection dataset with a view of comparing the performances of Deep Neural Networks (DNN) and Convolutional Neural Networks (CNN) in both multiclass and binary

classification tasks. The dataset contained a balanced proportion of normal traffic and a number of attack types that enabled a comprehensive evaluation of model performance in distinguishing between malicious and benign network activities.

For multiclass classification, the CNN achieved 99.10% accuracy, 99.08% precision, and 99.05% F1-score. The DNN indicated a lower but excellent performance with accurancy of 99.02% while 98.95% represented the F1-score (MDPI). One can observe that both models are extremely good at handling the richness of multi-category threat detection in IoT settings.

For attack and benign class separation through binary classification, the CNN was better than the DNN by a narrow margin. The CNN resulted in a 99.40% accuracy and a precision of 99.43% while the an F1-score was 99.41%. However, the DNN gave an accuracy of 99.38% (MDPI). These results validate that the two models are capable of carrying out near-perfect classification for binary IoT threat detection tasks, with CNN performing slightly better in precision and overall accuracy.

The extremely consistent behavior of the CNN under both classification scenarios justifies its application in low-latency IoT threat detection pipelines. Its convolutional architecture excels extremely well at capturing spatial-temporal patterns from network traffic feature representations and helps enable fast and accurate threat classification. Moreover, the minor difference in performance between CNN and DNN suggests that while CNN is most suitable for production deployment, DNN is a decent alternative in environments where simpler architecture or minimal training complexity is desired

5.0 Case Studies

This chapter presents three practical applications of machine learning in cybersecurity: IoT intrusion detection, ransomware detection, and autonomous anomaly detection. All case studies highlight the innovative potential of diverse ML models in combating diverse and evolving threats.

5.1 IoT Intrusion Detection

An IoT network intrusion dataset has been used to determine the performance of Convolutional Neural Networks (CNN) and Deep Neural Networks (DNN) in multiclass and binary classification. For the multiclass problem, CNN achieved 99.10% accuracy, 99.08% precision, and 99.05% F1-score, while DNN achieved 99.02% accuracy and 98.95% F1-score.For the binary classification case, CNN slightly outperformed DNN with 99.40% accuracy, 99.43% precision, and 99.41% F1-score compared to DNN's 99.38% accuracy (MDPI).

These findings show that CNN provides a small but stable advantage in both classification cases and thus is a good candidate to be used in low-latency IoT intrusion detection pipelines. Its component design is suitable for the extraction of spatial-temporal characteristics associated with network traffic features and thus facilitation of quick and accurate identification of threats in IoT settings.

5.2 Ransomware Detection with Random Forest

This case study is centred on the identification of ransomware through the application of a specific dataset, that is the UGRansome2024. This database was designed to discover distinct behaviour and network traffic patterns concerning ransomware campaigns.

Random Forest was chosen to aid classification because of (I) its capability to handle high-dimensional, noisy datasets and (ii) its strong generalization capability towards unseen patterns. For the UGRansome2024 dataset, the model achieved a classification accuracy of 96% indicating that it is capable of distinguishing efficiently between normal network traffic and flows generated by ransomware.

The model's detection capability was significant for influential ransomware families

(including EDA and Globe). This is because they exhibit favourable financial and operational consequences. Random Forest's ensemble learning framework, is an embodiment of multiple decision trees, and therefore showed reliable predictions with minimal possibility for overfitting. BY extension, this ensemble can also offer significant information to cybersecurity analysts regarding information on indication or sensing of ransomware activity.

Furthermore, the success of the approach shows that curation of custom datasets such as UGRansome2024 has a key function in enhancing the effectiveness of detection against targeted malware families.

5.3 CAMLPAD Autonomous Anomaly Detection

Cybersecurity Autonomous Machine Learning Platform for Anomaly Detection (CAMLPAD) is a highly advanced, auto-capable platform capable of identifying anomalous patterns within streaming security data. The platform integrates a range of unsupervised algorithms, including Isolation Forest, Histogram-Based Outlier Score (HBOS), Cluster-Based Local Outlier Factor (CBLOF), and K-Means clustering, to detect a wide range of types of anomalies from slow-moving reconnaissance activity to high-volume spikes of denial-of-service attacks.

THe performance of CAMLPAD on targetted data set gavean an Adjusted Rand Score (ARS) of 95%, which gives strong alliance between its anomaly groupings and expert-labeled ground truth. System design allows real-time deployment with parallel running of algorithms to minimize detection latency.

One of CAMLPAD's distinctive features is its visualization layer, made Kibana-enabled, through which human analysts can engage with and interpret clusters of anomalies in real time. Analysts can drill into suspicious flows, validate alerts, and provide feedback—feeding CAMLPAD's continuous model retraining cycle. This machine–human hybrid process enhances accuracy and trust in the system's alerts, and is especially adapted for enterprise-level Security Information and Event Management (SIEM) integrations.

5.4 Discussion

Case studies show that different machine learning models excel in different cybersecurity settings. Convolutional Neural Networks (CNNs) consistently have better accuracy and high F1 scores when applied for Internet of Things (IoT) intrusion detection, while Random Forest classifiers are still quite effective at detecting ransomware, with excellent precision in discriminating malicious from benign flows. In contrast, unsupervised machine learning techniques, as evidenced through CAMLPAD, are better than anomaly detection through the detection of new threats before labeled data is available.

A common thread among these applications is complementarity between human experts and AI systems. AI models can efficiently sift out alarms, rank events, and highlight suspicious trends within massive datasets. Human originated knowledge is a vital tool in the monitoring , approval contextualization.of AI-driven decisions.

Although some gains have been recorded some challenges are still known for their existence. Model interpretability is an essential requirement for the establishment of trust in AI-driven security solutions. Methods such as anomaly scoring decomposition enhance interpretability so that analysts can improve understanding and explain system outputs. Also, there is need consider the vulnerability of machine learning models regarding adversarial threats or attacks, scheming or manifesting as data poisoning and evasion strategies. Consequently, comprehensive defense requires the integration of adversarial defensive techniques and continuous monitoring of the ML pipeline. Human error can also be a major vulnerability, but can be overcome through security training, development of a strong security culture, and the employment of

automated controls. Finally, the emergence of synthetic threats such as deepfakes, synthetic identities, and AI-driven fraud calls for explainable, adaptive, and collaborative ML systems, a priority noted by TechRadar.

6.0 Concussion

The findings from this study demonstrate that machine learning algorithms, e.g., Convolutional Neural Networks (CNNs) for IoT intrusion detection and Random Forest classifiers for ransomware detection, have high-accuracy and robust classification performance for their respective areas. Unsupervised approaches like the CAMLPAD anomaly detection framework have great potential to detect new and evolving attacks prior to the availability of labeled data for supporting early warning capabilities. Best approach requires hybridizing human judgment with AI-driven detection systems to enable proper validation, interpretation, and action on automated insights. THe results also indicated persisting challenges that requires explainability to support trust in model outputs to ML model vulnerability to attacks and human-error weaknesses. Human analysts and AI systems must work together for assurance of efficiency in cybersecurity operations,concerningeisting and unexpected threats. However, the effectiveness of these systems depends on addressing principal challenges, i.e., model interpretability, adversarial robustness, and human-factor vulnerabilities.Future deployments ought to prioritize explainable AI methods to establish credibility and openness into autonomous choices.

6.0 References

Ademilua, D.A. (2021). Cloud Security in the Era of Big Data and IoT: A Review of Emerging Risks and Protective Technologies. *Communication in Physical Sciences*, 7, 4, pp. 590-604

Ayereby, M. P.-M. (2018). *Overcoming data breaches and human factors in minimizing threats to cyber-security ecosystems* (Publication No. 10991950) [Doctoral dissertation, Walden University]. ScholarWorks. https://scholarworks.walden u.edu/dissertations/5243.

Seh, A. H., Zarour, M., Alenezi, M., Sarkar, A. K., Agrawal, A., Kumar, R., & Khan, R. A. (2020). Healthcare data breaches: Insights and implications. *Healthcare*, 8,2, 133. https://doi.org/10.3390/healthcare8020133 .Coos, A. (2021, December 29). *The cost of a data breach in 2021*. Endpoint Protector. https://endpointprotector.com/blog/the-cost-of-a-data-breach-in-2021/

Alrzini, J., & Pennington, D. (2020). A review of polymorphic malware detection techniques. *International Journal of Advanced Research in Engineering and Technology*, 11, 12, pp. 1238–1247. https://doi.org/10.34218/IJARET.11.12.2020.119

Lawal, S. A., Omefe, S., Balogun, A. K., Michael, C., Bello, S. F., Taiwo, I., Ifiora, K. N. (2021). Circular Supply Chains in the Al Era with Renewable Energy Integration and Smart Transport Networks. *Communication in Physical Sciences*, 7(,4, pp. 605-629

Najafabadi, M. M., Villanustre, F., Khoshgoftaar, T. M., Seliya, N., Wald, R., & Muharemagic, E. (2015). Deep learning applications and challenges in big data analytics. *Journal of Big Data*, 2,1, 1. https://doi.org/10.1186/s40537-014-0007-7

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, pp. 1153–1176. https://doi.org/10.1109/COMST.2015.2494502

Ring, M., Wunderlich, S., Scheuring, D., Landes, D., & Hotho, A. (2019). A survey of network-based intrusion detection data sets. *Computers & Security*, *86*, pp. 147–167. https://doi.org/10.1016/j.cose.2019.06.005.

Shaukat, K., Luo, S., Chen, S., & Liu, D. (2020). Cyber threat detection using machine learning techniques: A performance evaluation perspective. In *2020 IEEE International Conference on Cyber Warfare and Security (ICCWS)* (pp. 1–6). IEEE. https://doi.org/10.1109/ICCWS48432.2020.9292388

Declaration

Consent for publication

Not applicable

Availability of data

Data shall be made available on demand.

Competing interests

The authors declared no conflict of interest

Ethical Consideration

Not applicable

Funding

There is no source of external funding.

Authors' Contributions

A.O.A. conceived the study, coordinated design, and wrote the first draft. O.C.A. developed and validated the machine learning models. A.B.B. optimized algorithms and refined theoretical frameworks. I.A.A. handled data acquisition, preprocessing, and hybrid model testing. All authors reviewed, edited, and approved the final manuscript, contributing equally to analysis, interpretation, and conclusions.