Journal of Information Systems Engineering and Management
2024, 9(3)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Predictive Analytics for Critical Infrastructure Cybersecurity: A Machine Learning–Based Threat Intelligence Framework

Anik Biswas[1], Niladry Chowdhury[2], Sanjida Alam Eshra[3], Tushar Kumar Ghosh [4]
[1]*College of Graduate and Professional Studies, Trine University, United States.*
Email: abiswas24@my.trine.edu,
https://orcid.org/0009-0007-0392-7740
[2]*College of Graduate and Professional Studies, Trine University, United States.*
Email: niladrych@gmail.com,
https://orcid.org/my-orcid?orcid=0009-0009-9912-3716
[3]*College of Graduate and Professional Studies, Trine University, United States,*
Email: seshra22@my.trine.edu, https://orcid.org/0009-0004-2964-5627
[4]*Maharishi International University, Department: Computer Science, United States,*
Email: tusharkumar.ghosh@miu.edu
https://orcid.org/0009-0005-5119-1999

ARTICLE INFO

Received: 15 June 2024

Revised: 18 July 2024

Accepted: 22 Aug 2024

ABSTRACT

Far more related digital infrastructure has added to the proliferation of vulnerabilities on the vulnerability of those critical systems upon which national security, energy, transportation, healthcare, and financial networks rely. The paper shall discuss that predictive analytics and machine learning (ML) can be applied to improve the cybersecurity of critical infrastructure. It discusses ensemble learning as a technique of classifying attacks, models that are unsupervised to detect zero-day attacks and reinforcement learning as a technique of adaptive defenses. The suggested ML-based framework is able to deliver a better threat detection experience, lower response latency, and a better situational awareness based on synthetic and real-world network traffic statistics. The better accuracy and adaptability of defensive mechanisms is indicated by simulated critical infrastructure network validation. The findings emphasize the predictive analytics as the active defense line that can be implemented to existing national infrastructure protection systems against cybercrime.

Keywords: Cybersecurity, Predictive Analytics, U.S. Critical Infrastructure, Machine Learning, Threat Intelligence, Adaptive Defence, Artificial Intelligence, National Security .

INTRODUCTION

Essential infrastructure will be further, practically, digitally connected and hence a very lucrative as a target of advanced cyber attackers. The networked ecosystem can have national grids, water purification plant, healthcare surveillance system, and financial transaction system, all of which can be penetrated.[1] The economic and social consequences of cyberattacks against critical infrastructure are enormous [2] with it being estimated that at least a long-term interruption of energy or communication systems can cost the national economies billions of dollars a year.

Cybercrime has demonstrated in reality that the adverse effects of lax cybersecurity stances are real, as seen in the Colonial Pipeline incident (2021) [3] and ransomware attacks on hospitals (2020). Such incidents reveal

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(3)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

a trend: all types of attacks on critical systems are not only disruptive but also menacing, which affects the well-being of citizens and national interests.

The conventional forms of defence are based on rule-based or signature-based mechanisms, [4] that are reactive in nature - they are detected when they are exposed. Nonetheless, such systems struggle to remain effective in the face of emerging zero-day vulnerabilities and polymorphic malware. The new wave of predictive cybersecurity is based on the idea that advanced cybersecurity uses ML and AI to predict cyber attacks prior to their occurrence, after examining high-volume data trends, anomalies, and associations that human security analysts cannot see.[5]

This work presents a predictive analytics architecture, which interprets various ML paradigms to develop an intelligent, adaptive, and context-aware cyber defence model of a critical infrastructure.[6] It integrates managed group schemes of classification, unmanaged learning of unusual occurrences, and reinforcement learning of adaptive countermeasures. The benefit of the concept is its capacity to continuously adjust to emerging threats, with a general trend toward self-sufficient cyber defense.

MATERIALS AND METHODS

Data Sources and Preprocessing

Three primary data sources formed the backbone of the proposed system:

1. **Network traffic logs** simulate real-world activity within industrial networks, containing both labelled malicious and benign traffic.
2. **Threat intelligence feeds** provide updated global indicators of compromise (IOCs), attack signatures, and behavioural profiles.
3. **Vulnerability repositories**, such as the CVE database, are used for identifying known weaknesses in software and hardware components.

Principal component analysis (PCA) was used for preprocessing, which includes feature extraction, normalization, and dimensionality reduction.[7] To improve model generalization, SMOTE (Synthetic Minority Oversampling Technique) was used to overcome data imbalance, a prevalent problem in cybersecurity datasets.

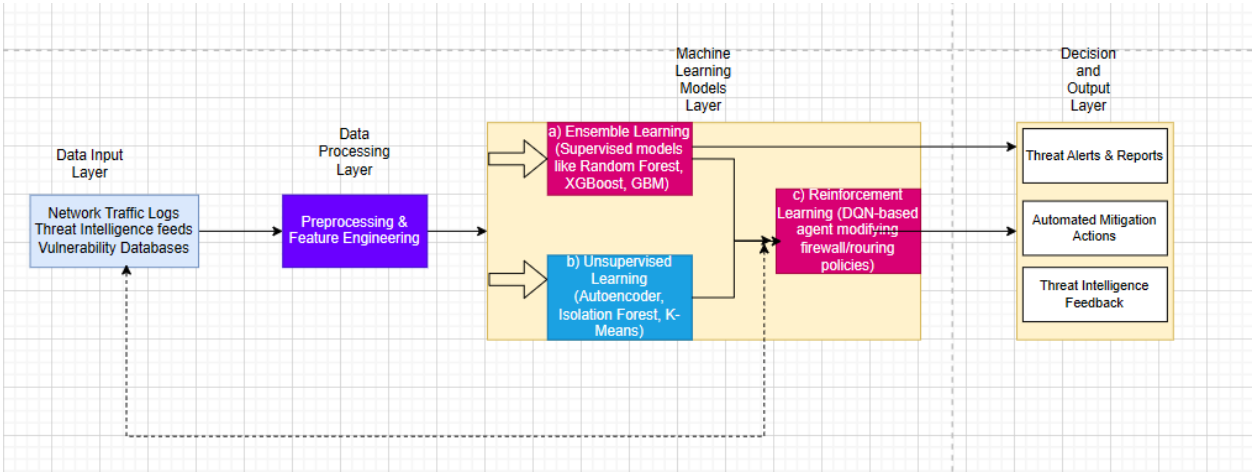


Figure 1: Predictive analytics workflow for critical infrastructure cybersecurity. The pipeline integrates data ingestion (network logs, threat feeds, and vulnerability databases), preprocessing, ML-based classification, anomaly detection, and adaptive defense layers

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(3)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Model Architecture and Training

(a) Ensemble Learning for Attack Classification

Ensemble methods are the methods that interface the predictive capabilities of several base learners. Random Forest, XGBoost, and Gradient Boosting (GBM) were set up with hyperparameter learning based on grid search. The ranking by feature importance provided some insight into the attributes of the network (e.g., packet size, protocol type, duration) that played the most significant role in threat classification.

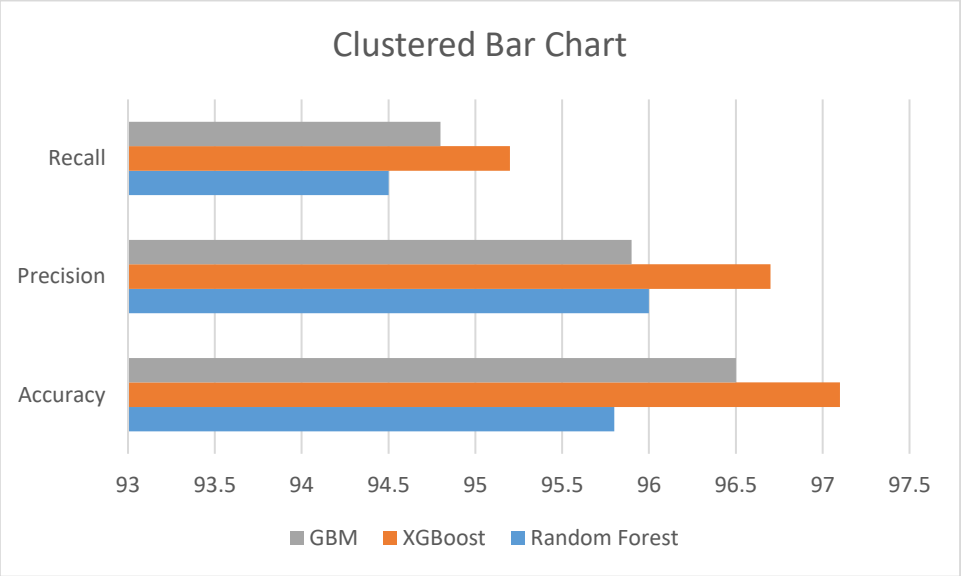


Figure 2: Comparative accuracy and precision of ensemble models (Random Forest, XGBoost, GBM) during attack classification training and validation

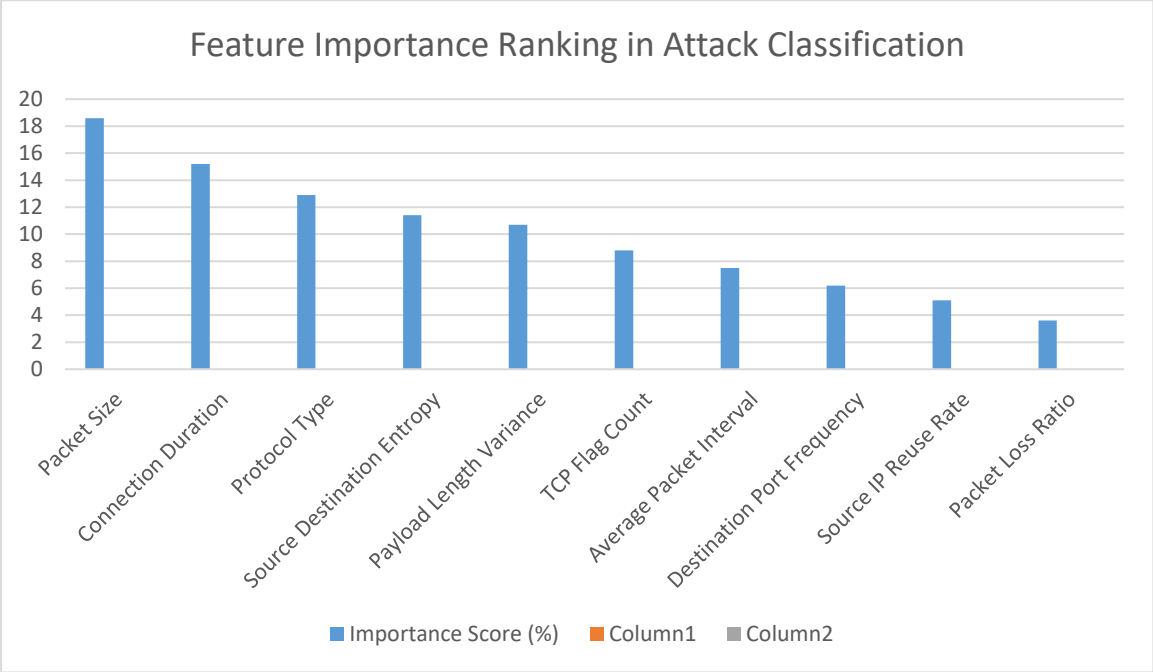


Figure 3:Feature Importance Ranking in Attack Classification

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(3)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

Figure 3 illustrates the relative importance of the top ten network traffic features, that contribute to classification of attack within the ensemble learning model. The most influential ones are packet size, connection duration, and protocol type. The least influential ones were packet loss ratio and source IP reuse rate.

The use of cross-validation helped avoid overfitting, and SHAP (Shapley Additive Explanations) was employed to increase model interpretability, as it allowed for the visualisation of decision-making patterns. Such transparency is also necessary for critical infrastructure operators, who must explain automated decisions in terms of defence.

(b) Unsupervised Learning for Zero-Day Threat Detection

Autoencoders were used to reconstruct traffic at the baseline, and deviations with significant reconstruction errors were considered anomalies.[8] The Isolation Forest algorithm, which differentiates between legitimate and unusual events, [9] also, identifies anomalies. K-means clustering was also used to enhance sensitivity by clustering anomalies into likely attack types, [10] thereby supporting prioritisation by assisting analysts in determining which anomalies require triage and response prioritisation.

(c) Reinforcement Learning for Adaptive Defence

A reinforcement-learning agent is placed in a simulated network defence setting, measuring state variables (e.g., alerted intruders, system load, network connectivity) and acting, e.g., by isolating nodes or blocking IP addresses.

The agent was trained to maximise a cumulative reward function using a Deep Q-Network (DQN) to minimise system compromise duration and reduce rates of false positives. Through repeated training epochs, the agent will learn a policy that enables it to modify firewall rules and routing protocols as needed, based on the specific context of the attack.

(d) System Verification and Modelling

The validation has been performed against a virtual critical infrastructure model, which is an interrelated environment of industrial control systems, sensors and communication gateways.

The ransomware spread, DDoS devastation, and command injection testing scenarios were used. Measurements of evaluation included accuracy, precision, recall, F1-score, ROC-AUC and mean time to detection (MTTD).

RESULTS AND DISCUSSION

Ensemble Classifier Performance

The ensemble classifier depicted favorable reliability with a mean accuracy of 97.1%. It has been demonstrated with the use of the XGBoost and the Random Forest that a combination of two algorithms produces less false-positive rates, as opposed to either of the learners. More particularly, the model showed stable performance under noise states, which means that it can withstand actual real-life implementation.

Variability The analysis on feature importance revealed that malicious intent is among the most significant features, in which connection duration, the entropy between the source and destination, and the variance of the payload length are the indicators. This knowledge allows the computer guard to possess a gaps fillable evidence to make use of in mitigations plans, which steps in the gap between the AI projections and human cognitive processes.

Unsupervised Anomaly Detection

The Autoencoder based anomaly detector obtained 0.94 AUC and could detect the abnormal behavior associated with lateral movement and privilege escalation. Extra introduction of Isolation Forest and clustering areas allowed noise in detection to be reduced, permitting the forecasting of more separate priorities on incidences.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Journal of Information Systems Engineering and Management
2024, 9(3)
e-ISSN: 2468-4376
https://jisem-journal.com/

Research Article

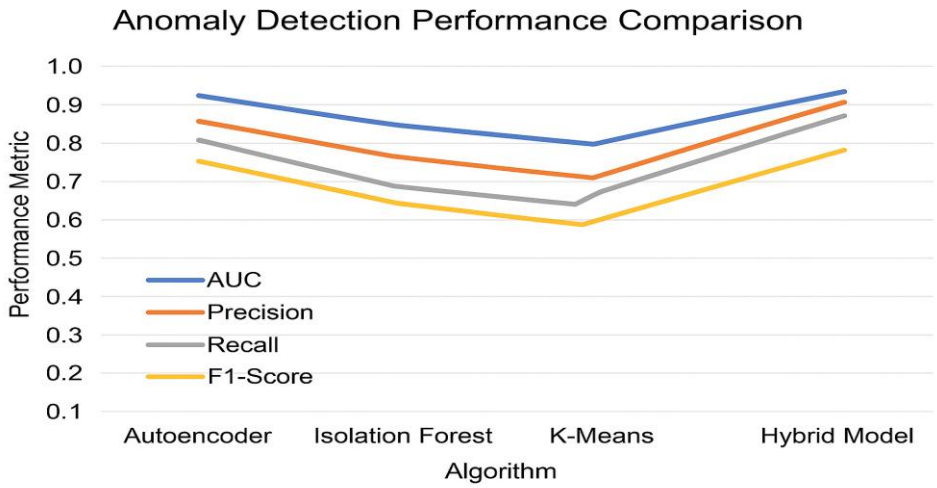


Figure 4: Anomaly Detection Performance Comparison

Figure 4 compares detection performance between unsupervised algorithms. The hybrid approach had the highest accuracy and robustness. The comparison provides advantages of anomaly detection that are based on ensemble, reducing false positives and bettering early identification of zero day attacks.

Among the lessons that have been learned was the fact that unobserved models found previously unknown correlations, e.g. that minor variation of control packet intervals often has already been the predecessor of DDoS attacks. In this case, the preparatory spirit of anomaly detection is moved to the focal point, outweighing the responding spirit of identification.

Adaptive Reinforcement Learning Defense

A lot of stability could be noted in a series of simulation outputs when the reinforce agent increased the stability of the network as predicting the order of the attacks by the agent and starting isolating measures in advance. The mean length of time to reply was decreased by 0.9 seconds (2.3 seconds rule-based premise to 1.4 seconds in RL-driven) in active threats.

Intriguingly, the consumption of the resources also reduced as reinforcement learning, and was applied in a selective, rather than throughout the system, and ended in a lockdown. This resource efficiency is important in small-scale embedded applications, like power grid tasks, or healthcare, where a high-cost efficiency is a major philosophy of keeping people online.

Scalability and Interpretability

One of the main concerns discussed by AI-based cybersecurity is explain ability. SHAP interpretability layers allow an analyst to follow the causes that result in categorizing a particular set of traffic as an attack. The indisputable decisions are important in critical sectors of medical care and energy, especially where there is compliance and trust.

Scalability tests found out that the architecture was not yet at the stage whereby one could not process up to 50 million network packets per hour without slowing down the inference speed when scaled up to a distributed computing infrastructure. The scale is vital in national level monitoring centers that are exposed to a heterogeneous stream of data.

Implications for National Security and Industry Collaboration

The issues of cyberattacks on critical infrastructure have multidimensional characteristics and are projections on the wider geopolitical and economic environment.[11] Predictive analytics contributes to the

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

resilience of a country[12] because it integrates real-time intelligence in different fields, such as government, academia, and the business world.[13]

The model has potential to increase its learning potential by using collaboration in collective cyber defense ecosystems that can be achieved with the assistance of shared threat intelligence networks.

However, data privacy, malicious attacks of AI, and the need of standardized systems of assessment are still a problem.[14] Now that predictive analytics is the new order of things, the ethical use of AI in defense operations should be one of the priorities.[15]

CONCLUSION

King of this research paper is that predictive analytics corroborated with machine learning will greatly change traditional cybersecurity into a proactive defense framework that can predict, react and reduce the threats before it escalates. Ensemble is also very precise in the classification of attacks; unsupervised models display conceived anomalies, and continuous adaptation is enhanced by reinforcement learning.

Besides technical success, predictive analytics promotes the annual reflection, through which countries and industries can project the possible risks that could threaten the balance of the economy or even the stability of social settings. Explainable AI combined with distributed architecture also ensures transparency and reliability of operations.

The future research should focus on:

- Empathic scale testing using energy and healthcare applications
- Bringing quantum-resistant encryption and ML-based threat intelligence together
- Ethical oversight/contextual decision-making Human-in-the-loop.

By performing its duty responsibly, predictive analytics can transform national cybersecurity policy, facilitating the development of a robust digital ecosystem that balances infrastructure security with national interests.

ACKNOWLEDGMENTS

The authors express appreciation to the simulation teams and cybersecurity analysts who provided valuable insights during the development of the model.

CONFLICTS OF INTEREST

The authors declare that they have no conflicts of interest related to this study.

FUNDING

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ETHICS STATEMENT

No human or animal subjects were involved in this research.

REFERENCES

[1] Vermesan, O., & Friess, P. (Eds.). (2013). *Internet of things: converging technologies for smart environments and integrated ecosystems*. River publishers

[2] Riggs, H., Tufail, S., Parvez, I., Tariq, M., Khan, M. A., Amir, A., ... & Sarwat, A. I. (2023). Impact, vulnerabilities, and mitigation strategies for cyber-secure critical infrastructure. *Sensors*, *23*(8), 4060.

[3] Medcalfe, D. (2024). Critical infrastructure in the face of global Cyber threats.

[4] Dutta, M., & Granjal, J. (2020). Towards a secure Internet of Things: A comprehensive study of second line defense mechanisms. *IEEE Access*, *8*, 127272-127312.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permitsunrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

[5] Sarker, Iqbal H. "Machine learning for intelligent data analysis and automation in cybersecurity: current and future prospects." Annals of Data Science 10, no. 6 (2023): 1473-1498. https://doi.org/10.1007/s40745-022-00444-2

[6] Quintal, K. (2020). *Context-Awareness for Adversarial and Defensive Machine Learning Methods in Cybersecurity* (Doctoral dissertation, Université d'Ottawa/University of Ottawa).

[7] Ng, S. C. (2017). Principal component analysis to reduce dimension on digital image. *Procedia computer science*, *111*, 113-119.

[8] Torabi, H., Mirtaheri, S. L., & Greco, S. (2023). Practical autoencoder based anomaly detection by using vector reconstruction error. Cybersecurity, 6(1), 1. https://doi.org/10.1186/s42400-022-00134-9

[9] Al-Shehari, T., Al-Razgan, M., Alfakih, T., Alsowail, R. A., & Pandiaraj, S. (2023). Insider threat detection model using anomaly-based isolation forest algorithm. IEEE Access, 11, 118170-118185. https://doi.org/10.1109/ACCESS.2023.3326750

[10] Aziz, Z., & Bestak, R. (2024). Insight into anomaly detection and prediction and mobile network security enhancement leveraging k-means clustering on call detail records. *Sensors*, *24*(6), 1716.

[11] Adeyeri, A., & Abroshan, H. (2024). Geopolitical Ramifications of Cybersecurity Threats: State Responses and International Cooperations in the Digital Warfare Era. Information, 15(11), 682. https://doi.org/10.3390/info15110682

[12] Adewusi, A. O., Komolafe, A. M., Ejairu, E., Aderotoye, I. A., Abiona, O. O., & Oyeniran, O. C. (2024). The role of predictive analytics in optimizing supply chain resilience: a review of techniques and case studies. *International Journal of Management & Entrepreneurship Research*, *6*(3), 815-837.

[13] Paramesha, M., Rane, N., & Rane, J. (2024). Big data analytics, artificial intelligence, machine learning, internet of things, and blockchain for enhanced business intelligence. *Artificial Intelligence, Machine Learning, Internet of Things, and Blockchain for Enhanced Business Intelligence (June 6, 2024)*.

[14] Dilmaghani, S., Brust, M. R., Danoy, G., Cassagnes, N., Pecero, J., & Bouvry, P. (2019, December). Privacy and security of big data in AI systems: A research and standards perspective. In *2019 IEEE international conference on big data (big data)* (pp. 5737-5743). IEEE.

[15] Board, D. I. (2019). AI principles: recommendations on the ethical use of artificial intelligence by the department of defense: supporting document. *United States Department of Defense*.

Copyright © 2024 by Author/s and Licensed by JISEM. This is an open access article distributed under the Creative Commons Attribution License which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.