Vol-3 Issue-1 2025
Scientific Research Journal of Science, Engineering and Technology
ISSN: 2584-0584, Peer Reviewed Journal

CYBER THREATS PREDICTON USING EXPERIENCE SHARING MODEL AND ENSEMBLE LEARNING ALGORITHM

Abubakar Bello[1], Abdulrashid Sabo[2]

[1]Africa Centre of Excellence on Technology Enhanced Learning (ACETEL), National Open University of Nigeria

[2]Department of Mathematics and Computer Science, Sule Lamido University, Kafin Hausa.

Corresponding Author: abusadeek46@gmail.com

ABSTRACT
The increasing complexity of cyber threats, particularly in critical industries such as oil and gas, necessitates proactive predictive models for early detection and response. Traditional frameworks such as the Common Vulnerability Scoring System (CVSS) are reactive, often addressing vulnerabilities post-incident, thereby exposing organizations to operational and financial risks. This study proposes a novel hybrid framework combining an experience-sharing model with ensemble machine learning algorithms, including bagging and boosting techniques. Using structured datasets such as VERIS and CAPEC, machine learning classifiers—logistic regression, k-Nearest Neighbors, and regression trees—were employed and validated using k-fold cross-validation. The results revealed a 94% prediction accuracy and a 0.96 AUC-ROC score with bagging ensembles, outperforming conventional models by 12%. A case study focused on Nigeria's oil and gas infrastructure validated the model's sector-specific applicability. This study contributes to cybersecurity analytics by demonstrating (1) the efficacy of ensemble learning, (2) a validated experience-sharing paradigm, and (3) the development of dynamic cyber-risk metrics suited for modern threats. The proposed framework offers cost-effective and scalable solutions for proactive threat mitigation.

Keywords: *Cybersecurity, Ensemble Learning, Threat Prediction, Machine Learning, Risk Assessment, Oil and Gas Sector*

---

1. Introduction

1.1 Background of the Study

The rapid integration of internet technologies into both public and private systems has heightened their vulnerability to cyberattacks. These threats range from malware and spyware to ransomware, phishing, and denial-of-service attacks. According to the Ponemon Institute (2020), the average cost of a data breach in 2019 was $3.92 million, reflecting the escalating economic consequences of cyber incidents.

Organizations are increasingly adopting artificial intelligence (AI) and machine learning (ML) to build predictive systems for identifying anomalies and potential threats in real time. The Capgemini Research Institute (2020) noted that 64% of firms observed improved threat response efficiency and a 12% decrease in latency due to AI-based predictive analytics.

1.2 Statement of the Problem

Conventional cybersecurity tools such as CVSS assess threats reactively, post-deployment. This leaves critical sectors—such as oil and gas—vulnerable during the window between deployment and response. Moreover, endpoint visibility challenges, data incompleteness, and resource limitations impair proactive defenses (Predicting Infection of Organization Endpoints, 2020). As Jaganathan et al. (2015) emphasized, organizations need dynamic frameworks that factor in real-time environmental and endpoint variables.

2. LITERATURE REVIEW

Cybersecurity threats continue to evolve in both complexity and impact, driving the need for advanced prediction models. The literature offers a diverse range of frameworks and approaches aimed at improving cyber threat prediction, particularly using machine learning and artificial intelligence.

Pathade and Bhosale (2021) emphasized the growing importance of using machine learning algorithms—such as regression analysis, k-nearest neighbors (kNN), and decision trees—to forecast potential cyber threats. Their study also highlighted the role of ensemble methods like bagging and boosting to improve the accuracy and robustness of threat detection models. The use of structured datasets, such as those provided by Microsoft and other open repositories, was key to developing these models.

The work by Jaganathan et al. (2015) focused on integrating environmental and vulnerability metrics into a proactive prediction model. Using multiple regression, they quantified the impact of cyber threats on critical systems and demonstrated the utility of incorporating CVSS scores, network traffic, and system vulnerability data into predictive analytics.

Another study, Predicting Infection of Organization Endpoints by Cybersecurity Threats Using Ensemble Machine Learning Techniques (n.d.), emphasized endpoint-level threat prediction by analyzing device-specific data (e.g., operating system, firewall status, RAM, disk capacity) and using ensemble algorithms to forecast infection probabilities. Their findings validated the strength of boosting algorithms—particularly stochastic gradient boosting—in identifying threats in high-dimensional, incomplete datasets.

In addition, Mehta et al. (2015) and Dalal & Rele (n.d.) demonstrated that integrating host-based intrusion detection systems (e.g., OSSEC) with machine learning models can significantly enhance detection capabilities. These systems, when coupled with honeypots and sandbox environments, allow real-time data collection and contextual analysis of attack behavior.

Collectively, these works emphasize the value of ensemble learning techniques, structured

incident schemas (e.g., VERIS and CAPEC), and dynamic feature selection in building robust, predictive cybersecurity models.

3. METHODOLOGY

This study adopts a data-driven, experimental design to evaluate the effectiveness of ensemble learning algorithms in predicting cybersecurity threats. A quantitative approach is utilized to extract, preprocess, train, and validate prediction models using real-world datasets.

Two well-established cybersecurity schemas were used:
1. **VERIS (Vocabulary for Event Recording and Incident Sharing):** Provides standardized incident reporting including vectors, actors, and impacts.
2. **CAPEC (Common Attack Pattern Enumeration and Classification):** Offers structured knowledge about adversary tactics and behavioral patterns.

Data preprocessing involved normalization, encoding of categorical features, and handling missing values. Missing data imputation was performed using Multiple Imputation by Chained Equations (MICE), which introduces randomness and preserves statistical validity.

The following algorithms were implemented and compared:
a. Multiple Linear Regression (MLR)
b. Nearest Neighbors (kNN)
c. CART (Classification and Regression Trees)
d. Logistic Regression
e. K-Means Clustering
f. Random Forest (Bagging)
g. Gradient Boosting Machines (GBM)

Models were evaluated using 10-fold cross-validation with metrics such as Accuracy, Precision, Recall, F1-Score, and AUC-ROC.
Tools used include Scikit-learn, Pandas, Matplotlib, Seaborn, and Jupyter/Google Colab.

4. RESULTS AND DISCUSSION

The comparative analysis showed that ensemble learning algorithms substantially outperformed individual classifiers.
a. **Random Forest (Bagging)** achieved 94% accuracy and 0.96 AUC, validating its robustness.
b. **Gradient Boosting Machines (GBM)** delivered high F1-score (0.89) and 0.95 AUC, ideal for high-impact threat detection.
c. **Logistic Regression and kNN** underperformed on complex data.

The model was adapted to simulate Nigeria's oil and gas infrastructure, providing localized insights for early warning and strategic resource allocation. This aligns with Jaganathan et al.'s (2015) recommendation for integrating environmental context into threat models.

The findings highlight the value of ensemble learning and structured incident data for proactive cybersecurity. Future enhancements may include real-time retraining, integration with SIEM, and robustness testing against adversarial manipulation.

5. CONCLUSION

This study demonstrates the applicability of ensemble machine learning models for predicting cybersecurity threats, with a focus on critical infrastructure such as the Nigerian oil and gas sector. Using structured datasets and cross-validated ensemble models, the research achieved high accuracy and reliability. Notably, Random Forest and Gradient Boosting models performed best across key evaluation metrics.

Key contributions include the development of a domain-specific cyber threat prediction model, integration of experience-sharing frameworks, and validation of ensemble methods for cyber-risk quantification. These outcomes are particularly relevant for sectors requiring preemptive resource allocation and security incident mitigation.

Future research should explore deep learning models, zero-day threat detection, and real-time deployment integration with SIEM platforms. Localized datasets and cross-organizational collaboration can further enhance the model's utility and adaptability.

Author Contributions:

A.B.: Conceptualization, Methodology (ensemble learning model), Writing – Original Draft. A.B.: Software (Python implementation), Data Curation (VERIS/CAPEC datasets), Formal Analysis. A.S.: Validation (k-fold cross-validation), Writing – Review & Editing. A.B.: Supervision, Project Administration

Funding:

This research received no external funding

Conflict of Interest Statement:

The authors declare no conflict of interest

Data Sharing Statement:

The datasets analyzed in this study—VERIS (Vocabulary for Event Recording and Incident Sharing) and CAPEC (Common Attack Pattern Enumeration and Classification)—are publicly available at their respective sources: VERIS Community Database and CAPEC MITRE Repository. The derived datasets and code used for ensemble learning analysis are available from the corresponding author upon reasonable request.

Software And Tools Use:

This study was implemented using Python 3.8 with key libraries including Scikit-learn (v1.0) for ensemble learning algorithms (bagging/boosting), Pandas (v1.3) for data processing, and

Matplotlib (v3.4) for visualization. The analysis was conducted in Jupyter Notebook and Google Colab environments. Anaconda (v2021.05) was used for package management

Acknowledgements:
We thank the Petroleum Technology Development Fund (PTDF), Nigeria, for their institutional support. We also acknowledge the VERIS and CAPEC communities for providing open-access datasets critical to this research. Special gratitude to ACETEL at National Open University of Nigeria for their technical guidance and to the anonymous reviewers for their constructive feedback.

REFERENCES

1. Axelrad, E. T., Sticha, P. J., Brdiczka, O., & Shen, J. (2013). A Bayesian network model for predicting insider threats. In 2013 IEEE Security and Privacy Workshops (pp. 82–89). IEEE.
2. Capgemini Research Institute. (2020). Reinventing cybersecurity with artificial intelligence: The new frontier in digital security. https://www.capgemini.com/research/reinventing-cybersecurity-with-ai/
3. Dalal, D., & Rele, M. (n.d.). Cyber attack prediction using machine learning and sandbox environment. [Conference paper].
4. Dalton, A., Bonnie, D., Leon, L., & Kristy, H. (2017). Improving cyber-attack predictions through information foraging. In 2017 IEEE International Conference on Big Data (BigData) (pp. 3326–3331). IEEE.
5. Freund, Y., & Schapire, R. E. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. Journal of Computer and System Sciences, 55(1), 119–139.
6. Hosmer, D. W., Lemeshow, S., & Sturdivant, R. X. (2013). Applied logistic regression (3rd ed.). Wiley.
7. Jaganathan, V., Cherurveettil, P., & Sivashanmugam, P. M. (2015). Using a prediction model to manage cyber security threats. The Scientific World Journal, 2015, Article ID 703713.
8. James, G., Witten, D., Hastie, T., & Tibshirani, R. (2013). An introduction to statistical learning: With applications in R (6th ed.). Springer.
9. Khan, M. A., & Hameed, M. (2010). Cyber security quantification model. Bahria University Journal of Information and Communication Technology, 3(1), 23–27.
10. MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability (Vol. 1, pp. 281–297). University of California Press.
11. Mehta, V., Bahadur, P., Kapoor, M., Singh, P., & Rajpoot, S. (2015). Threat prediction using honeypot and machine learning. In 1st International Conference on Futuristic Trends in Computational Analysis and Knowledge Management (ABLAZE-2015) (pp. 615–620). IEEE.
12. MITRE Corporation. (2022). Common Attack Pattern Enumeration and Classification (CAPEC). https://capec.mitre.org
13. Moore, A. P., Cappelli, D. M., & Trzeciak, R. F. (2013). A system dynamics model for investigating early detection of insider threat risk (CMU/SEI-2013-TR-004). Software Engineering Institute, Carnegie Mellon University.
14. Pathade, C., & Bhosale, T. (2021). Cyber threats prediction using machine learning. International Research Journal of Engineering and Technology (IRJET), 8(12), 1250–1255.
15. Ponemon Institute. (2020). 2019 Cost of a data breach report. IBM Security. https://www.ibm.com/security/data-breach

16. Predicting infection of organization endpoints by cybersecurity threats using ensemble machine learning techniques. (n.d.). [Unpublished manuscript].
17. Sheyner, O., Haines, J., Jha, S., Lippmann, R., & Wing, J. M. (2002). Automated generation and analysis of attack graphs. In Proceedings 2002 IEEE Symposium on Security and Privacy (pp. 273–284). IEEE.
18. Tahia, A., Soujanya, T. S., & Vasavi, D. S. (2012). Study on techniques for providing enhanced security during online exams. International Journal of Engineering Inventions, 1(1), 32–37.
19. Tittel, E. (2013). Preventing and avoiding network security threats and vulnerabilities. Tom's IT Pro.
20. van Buuren, S., & Groothuis-Oudshoorn, K. (2011). mice: Multivariate imputation by chained equations in R. Journal of Statistical Software, 45(3), 1–67.
21. VERIS Community. (2022). Vocabulary for Event Recording and Incident Sharing (VERIS). http://veriscommunity.net
22. Wu, J., Yin, L., & Guo, Y. (2012). Cyber attacks prediction model based on Bayesian network. In 2012 IEEE 18th International Conference on Parallel and Distributed Systems (pp. 730–731). IEEE.
23. Zhou, Z.-H. (2012). Ensemble methods: Foundations and algorithms. CRC Press.