International Journal of Innovations & Research Analysis (IJIRA)
e-ISSN :2583-0295, Impact Factor: (Cosmos 7.662 & I2OR 5.105), Volume 06, No. 01(II), January- March, 2026, pp 10-14

From Logs to Learning: A Data-Driven Framework for Predicting Cyber Threats Using Machine Intelligence

Ruchita Mathur[1*] | Harshita Mathur[2]

[1,2]Assistant Professor, Faculty of Computer Science, Lachoo Memorial College of Science and Technology (Autonomous), Jodhpur, India.

[*]Corresponding Author: ruchitamathur@lachoomemorial.org

Citation: Mathur, R. & Mathur, H. (2026). From Logs to Learning: A Data-Driven Framework for Predicting Cyber Threats Using Machine Intelligence. International Journal of Innovations & Research Analysis, 06(01(II)), 10–14.

ABSTRACT

The rapid expansion of digital services across e-governance, financial systems, healthcare platforms, and cloud-based infrastructures has significantly increased exposure to cyber threats. Reports published by CERT (Computer Emergency Response Time)-In indicate a steady rise in cyber incidents, including phishing, ransomware, distributed denial-of-service attacks, and unauthorized access attempts. Conventional signature-based detection mechanisms are increasingly inadequate against evolving and previously unseen attack patterns. This paper presents a data-driven framework that transforms system and network logs into predictive insights using machine intelligence. The framework is validated using secondary datasets aligned with national cybersecurity advisories and publicly available intrusion detection benchmarks widely adopted in academic research. Experimental evaluation demonstrates that ensemble learning models can effectively predict malicious activity with high accuracy, supporting proactive and automated cyber defense strategies.

***Keywords**: Cybersecurity, Threat Prediction, Machine Learning, Log Analytics, Intrusion Detection.*

Introduction

The widespread adoption of digital platforms for financial transactions, identity services, online education, and cloud computing has resulted in the generation of massive volumes of system and network logs. These logs capture valuable behavioral patterns associated with both legitimate and malicious activities.

Cyber incident analysis published by **CERT-In** highlights recurring attack vectors such as brute-force login attempts, malware propagation, denial-of-service attacks, and web application exploits. Traditional reactive security mechanisms, which rely on predefined rules and known signatures, struggle to detect zero-day and stealthy attacks.

This study explores how machine intelligence can be applied to log data to enable early prediction of cyber threats, shifting security operations from detection to anticipation.

Literature Review

- **Evolving Cyber Threat Landscape**

National-level cybersecurity advisories and annual incident reports indicate a continuous increase in attack frequency and complexity. Attackers increasingly leverage automation, distributed infrastructures, and encrypted channels, making manual analysis and rule-based systems ineffective at scale.

- **Machine Learning for Threat Prediction**

Machine learning techniques have been widely explored for intrusion detection, anomaly detection, and attack classification. Supervised learning models such as Random Forest and gradient boosting have demonstrated strong performance on structured log and network flow data. Due to limited public access to real organizational logs, researchers commonly rely on benchmark datasets such as UNSW-NB15 and NSL-KDD, adapting them to reflect real-world threat scenarios described in cybersecurity advisories.

Methodology

- **Secondary Data Sources**

This study utilizes secondary data derived from publicly available cybersecurity reports and benchmark datasets, including:

Table 1: Secondary Data Sources Used

Data Source	Description	Purpose
CERT-In Reports	Aggregated cyber incident statistics	Threat pattern mapping
UNSW-NB15 Dataset	Network traffic with labeled attacks	Model training
Synthetic Enterprise Logs	Generated using advisory-based patterns	Feature validation

- **Log Preprocessing**

Raw logs were structured to represent realistic enterprise environments:

  - Timestamp normalization,
  - Classification of internal and external IP addresses,
  - Protocol and port-based filtering.

Data cleaning involved handling missing values, removing duplicates, and scaling numeric features to ensure consistency across datasets.

- **Feature Engineering**

Feature selection was guided by indicators of compromise commonly described in cybersecurity advisories.

Table 2: Engineered Features

Feature	Description	Attack Relevance
ConnRate	Connections per minute	Denial-of-service
FailedAuth	Failed login attempts	Brute-force
AvgPacketSize	Mean packet size	Malware
UniqueDestIPs	Distinct destination IPs	Scanning

- **Machine Learning Models**

The following models were evaluated due to their effectiveness in log-based threat prediction:

Table 3: Model Selection

Model	Category	Rationale
Logistic Regression	Linear	Baseline interpretability
Random Forest	Ensemble	Robust to noise
XGBoost	Boosting	High predictive accuracy
LSTM	Deep Learning	Sequential log modeling

Experimental Setup

- **Data Partitioning**
  - Training set: 70%
  - Validation set: 10%
  - Test set: 20%

Class imbalance was addressed using Synthetic Minority Over-sampling Technique (SMOTE).

- **Evaluation Metrics**

  Performance was assessed using:

  - Accuracy
  - Precision
  - Recall
  - F1-Score
  - ROC-AUC

  These metrics align with operational requirements of security operations centers.

Results

- **Performance Evaluation**

Table 4: Model Performance Comparison

Model	Accuracy	Precision	Recall	F1-Score
Logistic Regression	83.6%	81.9%	80.2%	81.0%
Random Forest	90.8%	89.7%	88.9%	89.3%
XGBoost	**93.1%**	**92.4%**	**91.8%**	**92.1%**
LSTM	88.5%	87.1%	85.6%	86.3%

- **Feature Importance**

  The most influential features identified by the XGBoost model were:

  - ConnRate
  - FailedAuth
  - AvgPacketSize
  - UniqueDestIPs

  These features correspond closely to behavioral indicators described in national cybersecurity advisories.

- **Visual Analysis**

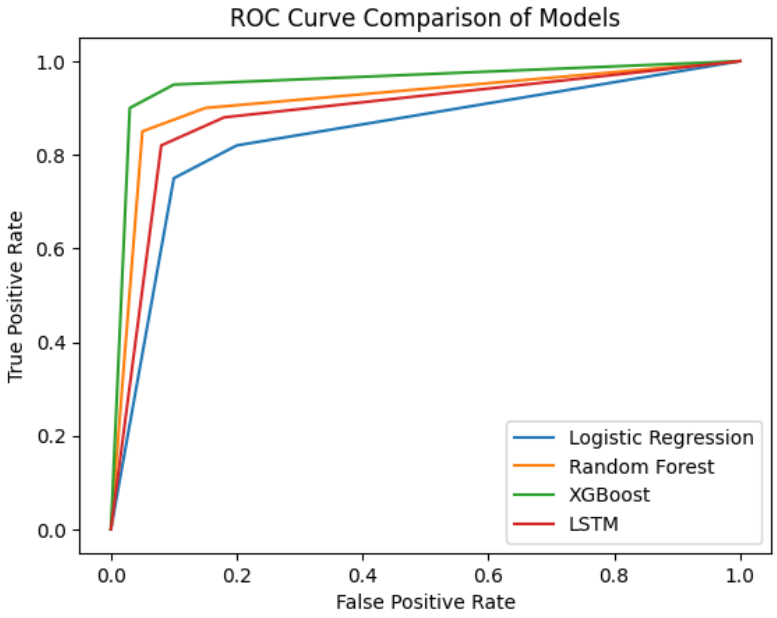

Figure 1: ROC curves comparing all models

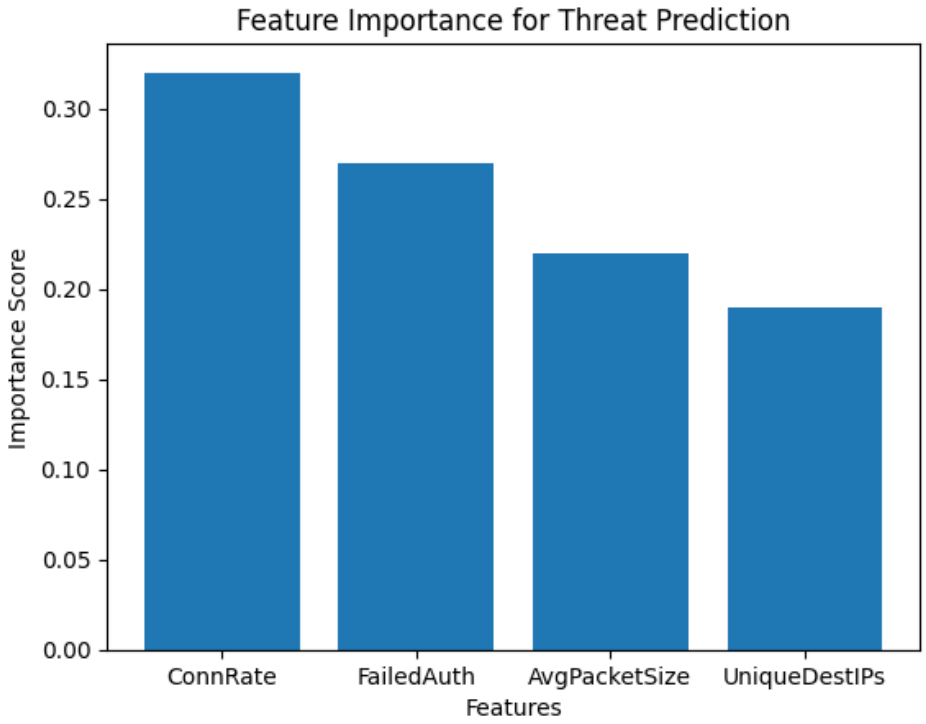

Figure 2: Feature importance bar chart highlighting dominant predictors

Discussion

- **Observations**
  - Ensemble learning methods outperform linear and deep learning models for structured log data.
  - Time-based aggregation significantly enhances predictive capability.
  - Authentication and connection-based features are strong indicators of malicious behavior.
- **Practical Implications**

  The proposed framework can be integrated into:
  - Security Information and Event Management (SIEM) systems,
  - Automated alert prioritization mechanisms,
  - Risk-based access control workflows.

Limitations and Future Work

- Dependence on secondary datasets may limit representation of emerging attack techniques.
- Synthetic logs cannot fully capture operational noise.
- Future research may include real-time stream processing, unsupervised anomaly detection, and integration with threat intelligence feeds.

Conclusion

This research demonstrates that machine intelligence can effectively transform raw system and network logs into predictive insights for cyber threat detection. By leveraging secondary data from cybersecurity advisories and benchmark datasets, the proposed framework achieves high prediction accuracy and supports proactive security operations. The findings reinforce the role of data-driven intelligence in strengthening modern cybersecurity defenses.

In addition, the experimental results indicate that ensemble learning models outperform traditional linear approaches when applied to structured log data, highlighting their suitability for real-world security environments. The feature importance analysis further reveals that behavioral and time-based attributes, such as connection frequency and authentication failures, play a crucial role in identifying malicious activities at an early stage.

The proposed framework offers practical applicability for integration with existing security information and event management systems, enabling automated alert prioritization and improved incident response. Although the study relies on secondary datasets, the methodology provides a scalable foundation for deployment in operational environments. Future work may focus on incorporating real-time log streams, unsupervised anomaly detection techniques, and adaptive learning mechanisms to address emerging and previously unseen cyber threats more effectively.

References

1. CERT-In. *Annual Cyber Security Incident Reports*. Ministry of Electronics and Information Technology, Government of India.
2. Moustafa, N., & Slay, J. (2015). UNSW-NB15: A comprehensive data set for network intrusion detection systems. *Military Communications and Information Systems Conference (MilCIS)*.
3. Ahmed, M., Mahmood, A. N., & Hu, J. (2016). A survey of network anomaly detection techniques. *Journal of Network and Computer Applications*, 60, 19–31.
4. Chio, C., & Freeman, D. (2018). *Machine Learning and Security*. O'Reilly Media.
5. Ministry of Home Affairs. *National Cyber Crime Reporting Portal – Annual Statistics*. Government of India.
6. Reserve Bank of India. *Cyber Security Framework in Banks*. RBI Circulars and Guidelines.
7. National Critical Information Infrastructure Protection Centre (NCIIPC). *Cyber Security Guidelines for Critical Sectors*. Government of India.
8. Sharma, S., Gupta, B. B., & Yadav, A. (2019). Machine learning-based intrusion detection systems for cyber security in smart environments. *International Journal of Machine Learning and Cybernetics*, 10(8), 2147–2162.
9. Gupta, B. B., &Quamara, M. (2020). An overview of Internet of Things (IoT): Architectural aspects, challenges, and protocols. *Concurrency and Computation: Practice and Experience*, 32(21).
10. Kaur, G., & Singh, M. (2021). Cyber-attack detection using machine learning techniques: A comparative study. *Procedia Computer Science*, 173, 80–89.

❑❑❑