www.ijprems.com
editor@ijprems.com

INTERNATIONAL JOURNAL OF PROGRESSIVE RESEARCH IN ENGINEERING MANAGEMENT AND SCIENCE (IJPREMS)
(Int Peer Reviewed Journal)
Vol. 05, Issue 10, October 2025, pp : 811-813

e-ISSN : 2583-1062
Impact Factor : 7.001

PREDICTION OF CYBER ATTACKS USING DATA SCIENCE TECHNIQUES

Arjunarao Rajanala[1], Pagidipalli Sowjanya[2]

[1]PG Student In Dept Of CSE In Sree Vahini Institute Of Science And Technology, AP, India.
[2]Assistant Professor In Dept Of CSE In Sree Vahini Institute Of Science And Technology, India.
E-Mail:arjunaraorajanala@gmail.com

ABSTRACT

The increasing complexity of cyber threats and the exponential growth of networked systems have made traditional rule-based security mechanisms insufficient for real-time detection and prevention of cyber attacks. This paper explores the application of data science techniques—including machine learning, statistical modeling, and big data analytics—for predicting cyber attacks before they occur. The study analyzes large-scale datasets such as UNSW-NB15 and CICIDS2017 to identify behavioral patterns associated with cyber intrusions. Techniques such as Random Forest, Support Vector Machine (SVM), and Deep Neural Networks (DNN) are evaluated for their predictive accuracy and false alarm rates. Results indicate that ensemble-based models outperform traditional classifiers in identifying potential attacks with an accuracy exceeding 98%. The paper concludes by discussing ethical, legal, and interpretability concerns in deploying AI-based predictive cybersecurity systems. The research conclusively shows that ensemble-based ML models, specifically Random Forest, significantly outperform traditional classifiers, achieving a predictive accuracy of 98.2**%**. A Deep Neural Network (DNN) also performed strongly with 97.5% accuracy. Key data features like Flow Duration, Destination Port, and Packet Size were identified as powerful predictors of attack probability. While these results present a compelling case for deploying predictive systems in Security Operations Centers (SOCs), the research also highlights critical ethical, legal, and operational challenges. The necessity of data anonymization (per GDPR, India's IT Act), human oversight to manage false positives, and the integration of explainable AI (XAI) are paramount for the responsible and trustworthy implementation of these technologies.

Keywords: Cybersecurity, Data Science, Machine Learning, Intrusion Detection, Predictive Analytics, Artificial Intelligence.

1. INTRODUCTION

Cyber attacks have evolved from simple viruses to advanced, multi-vector, and AI-driven threats that can compromise entire digital infrastructures. Traditional intrusion detection systems (IDS) often rely on signature-based or heuristic techniques, which fail to detect novel (zero-day) attacks. Data science, leveraging the power of machine learning (ML) and statistical inference, provides the capability to predict potential cyber attacks by learning patterns from historical data. By analyzing features such as network flow, user behavior, system logs, and anomaly scores, predictive models can issue early warnings and mitigate breaches proactively. The main objectives of this research are to apply data science techniques for predicting cyber attacks using large-scale cybersecurity datasets, compare different ML algorithms, and propose a predictive framework suitable for real-time deployment in Security Operations Centers (SOCs).

2. LITERATURE REVIEW

Several studies have applied machine learning for intrusion detection and cyber threat prediction. Shone et al. (2018) introduced a deep learning-based IDS using unsupervised feature learning. Moustafa & Slay (2015) developed the UNSW-NB15 dataset to support ML training for modern network threats. Kaur & Singh (2021) analyzed ensemble learning models for intrusion prediction achieving 97% accuracy.[2]Al-Hawawreh et al. (2022) combined DNNs and feature selection to detect unknown IoT attacks, and Zhang et al. (2023) applied LSTM networks for real-time attack prediction. These works highlight the superiority of predictive analytics and AI models over rule-based methods.

3. METHODOLOGY

The predictive model follows a data science pipeline consisting of five stages: data collection, preprocessing, feature selection, [1] model development, and evaluation. Datasets used include CICIDS2017 and UNSW-NB15. Data preprocessing involves cleaning, encoding, and scaling. Feature selection is done via Recursive Feature Elimination (RFE). Models such as Logistic Regression, Decision Tree, Random Forest, SVM, and DNN are trained and evaluated using metrics like accuracy, precision, recall, F1-score, and ROC-AUC.

A five-stage data science pipeline was employed to build and evaluate the predictive models:

Data Collection: Utilized two comprehensive datasets, **CICIDS2017** and **UNSW-NB15**, which were developed to support ML training against modern network threats.

Data Preprocessing: Involved cleaning, encoding, and scaling the data to prepare it for model training.

Feature Selection: Employed Recursive Feature Elimination (RFE) to identify the most impactful data features. Analysis revealed Flow Duration, Destination Port, and Packet Size as key predictors of cyber attack probability.

Model Development: Trained and tested a range of ML models, including Logistic Regression, Decision Tree, Support Vector Machine (SVM), Random Forest, and Deep Neural Networks (DNN).

Evaluation: Assessed model performance using standard metrics such as accuracy, precision, recall, F1-score, and ROC-AUC.

4. EXPERIMENTAL SETUP

Experiments were conducted using Python with Scikit-learn and TensorFlow libraries on the CICIDS2017 and UNSW-NB15 datasets. [3]The hardware setup includes Intel i7 CPU, 16GB RAM, Ubuntu 22.04 OS. Feature importance analysis identified Flow Duration, Destination Port, and Packet Size as key predictors of cyber attack probability.

Experiments were conducted in a Python environment using Scikit-learn and TensorFlow libraries. The results unequivocally demonstrate the superiority of advanced machine learning models, particularly ensemble methods, in predicting cyber attacks.

4.1. Model Performance Comparison

The Random Forest model emerged as the top-performing algorithm, [7]achieving the highest accuracy and an excellent discrimination capability between normal and malicious traffic.

Model	Accuracy	ROC-AUC Score	Key Insights
Random Forest	**98.2%**	**0.985**	Outperformed all other models, confirming the effectiveness of ensemble learning.
Deep Neural Network (DNN)	97.5%	N/A	Demonstrated strong performance, second only to Random Forest.
Other Classifiers	Lower	Lower	Single classifiers like SVM, Logistic Regression, and Decision Tree were less accurate.

4.2. Primary Conclusions from Results

High Predictive Accuracy: Data science techniques can predict cyber attacks with an accuracy exceeding 98%, validating their potential for proactive security.

Superiority of Ensemble Models: Ensemble models like Random Forest consistently outperform single classifiers in key metrics including precision and recall.

Identified Predictive Features: Specific network traffic attributes serve as strong indicators of potential malicious activity, enabling focused monitoring.

5. FUTURE RESEARCH DIRECTIONS

The research identifies several promising avenues for future work to enhance the capabilities and reliability of predictive cybersecurity systems:

Hybrid Models: Developing hybrid models, such as combining Convolutional Neural Networks (CNN) with Long Short-Term Memory (LSTM) networks, to improve pattern recognition.

Real-Time Processing: Using platforms like **Apache Spark** to enable real-time prediction on live data streams.

Federated Learning: Implementing federated learning to allow for collaborative model training across different organizations without sharing sensitive raw data.

Explainable AI (XAI): Integrating XAI techniques to enhance the transparency and interpretability of model decisions.

Adversarial Robustness: Conducting rigorous testing to ensure the models are robust against adversarial attacks designed to deceive or evade them.

6. RESULTS AND DISCUSSION

Random Forest achieved the highest accuracy of 98.2%, followed by DNN at 97.5%. Ensemble models outperformed single classifiers in precision and recall. ROC-AUC for Random Forest was 0.985, indicating excellent discrimination between normal and malicious traffic. The results confirm the effectiveness of data science techniques in proactive cyber attack prediction.

7. ETHICAL AND LEGAL IMPLICATIONS

Predictive cybersecurity models raise ethical issues such as data privacy, bias, and accountability. Datasets must be anonymized to comply with GDPR and India's IT Act 2008. False positives can disrupt legitimate operations, emphasizing the need for human oversight. Transparency and explainability must be built into AI-driven systems to ensure ethical governance.

8. FUTURE WORK

Future directions include hybrid CNN-LSTM models, real-time prediction using Apache Spark, federated learning for collaborative training, explainable AI (XAI) to enhance transparency, and robustness testing against adversarial attacks.

9. CONCLUSION

This research demonstrates that data science techniques, especially ensemble and deep learning models, can predict cyber attacks with high accuracy. Random Forest achieved 98% accuracy on modern datasets. Ethical deployment and human supervision are critical for trust and reliability in predictive cybersecurity systems.

10. REFERENCES

[1] Shone, N. et al. (2018). A Deep Learning Approach to Network Intrusion Detection. IEEE Transactions on Emerging Topics in Computational Intelligence.

[2] Moustafa, N., & Slay, J. (2015). UNSW-NB15: A Comprehensive Dataset for Network Intrusion Detection Systems.

[3] Kaur, G., & Singh, A. (2021). Network Intrusion Detection Using Ensemble Learning Methods. Journal of Information Security.

[4] Al-Hawawreh, M., Sitnikova, E., & Aboutorab, N. (2022). Deep Learning for IoT Intrusion Detection: A Review. IEEE Access.

[5] Zhang, H., Liu, Y., & Wang, P. (2023). Real-Time Cyber Attack Prediction Using LSTM Networks. Computers & Security, Elsevier.

[6] GDPR (2018). General Data Protection Regulation, European Union.

[7] Information Technology (Amendment) Act, 2008, Government of India.