IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

# Cyber Threat Detection in Supply Chains Using XGBoost

**Mr. S. Venu Gopal, K. Indu, T. Jagadish Kumar, S. Shalem Raj**
Assistant Professor, Information Technology
Students, Information Technology
ACE Engineering College, Hyderabad, India
kavidaindu9243@gmail.com

**Abstract**: *In the digital era, ensuring the security and resilience of supply chains is critical due to the growing complexity and volume of cyber threats. This project proposes a Cyber Threat Detection System that leverages advanced machine learning techniques, specifically the XGBoost algorithm, to enhance threat identification and risk management within digital supply chains. The system processes and analyzes data from diverse cybersecurity sources such as IDS, SIEM, and TIPs to detect phishing, malware, and DoS attacks. A secure Flask-based web interface allows users to upload data, receive predictions, and monitor threats in real-time.*

**Keywords**: Cybersecurity, XGBoost, Machine Learning, Supply Chain, Threat Detection

## I. INTRODUCTION

The digitization of supply chains introduces both operational efficiencies and a significant increase in cybersecurity vulnerabilities. As organizations increasingly adopt cloud platforms, IoT devices, and third-party integrations to streamline logistics, procurement, and inventory management, they inadvertently expand their attack surface. This complex web of interconnected systems is often targeted by sophisticated cyber threats such as ransomware, phishing, and supply chain attacks, which can disrupt operations, compromise sensitive data, and damage trust among partners.

Traditional cybersecurity mechanisms such as rule-based Intrusion Detection Systems (IDS), firewalls, and SIEM (Security Information and Event Management) tools often rely on predefined signatures or heuristics, which limits their ability to detect novel or advanced threats. These systems may also produce a high volume of false positives, increasing the burden on security analysts and delaying incident response.

To address these limitations, this research proposes an intelligent, machine-learning–based cyber threat detection system that leverages the power of **XGBoost (Extreme Gradient Boosting)** a robust and scalable ensemble learning algorithm. The system is trained on real-world cybersecurity datasets related to supply chain activities and learns to distinguish between benign and malicious patterns based on a wide range of features, including network behavior, IP activity, protocol types, and anomaly scores.

The model is deployed within a Flask-based web application, allowing users to upload datasets and receive real-time threat predictions through a user-friendly interface. The system not only improves detection accuracy but also integrates with.

existing cybersecurity tools like IDS and SIEM to offer a hybrid defense mechanism. By combining traditional log monitoring with predictive analytics, it offers a proactive approach to securing modern digital supply chains against evolving cyber threats.

## II. LITERATURE REVIEW

In recent years, the exponential rise in cyber threats across digital infrastructures has intensified the need for intelligent, data-driven detection systems. Traditional signature-based methods such as rule-based intrusion detection systems (IDS) and static firewalls, though widely adopted, have demonstrated limitations in identifying novel and complex

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 9, June 2025 Impact Factor: 7.67

attack patterns. To address these limitations, researchers have increasingly turned toward machine learning (ML) and artificial intelligence (AI) for predictive cyber threat detection.

Various studies have explored the effectiveness of ML algorithms in cybersecurity. For instance, **Ali et al. (2021)** utilized K-Means clustering for anomaly detection in network traffic, showing that unsupervised learning can uncover hidden threats without the need for labeled datasets. However, the approach struggled with interpretability and scalability in real-time applications.

In a broader survey, **Kiani et al. (2021)** compared supervised algorithms such as Decision Trees, Support Vector Machines (SVM), and Neural Networks for threat classification. While Neural Networks offered high accuracy, they required significant computational resources and were prone to overfitting. Decision Trees and SVMs provided interpretable results but underperformed on imbalanced datasets.

More recently, researchers have emphasized ensemble learning techniques due to their robustness and generalization capability. **Nguyen and Schreiber (2023)** employed Random Forests and Gradient Boosting to detect malware in enterprise networks. Their findings highlighted improved performance in terms of recall and false positive rate when compared to single classifiers.

The **XGBoost** algorithm (Extreme Gradient Boosting), developed by Chen and Guestrin (2016), has emerged as a state-of-the-art method for structured data classification. Its ability to handle missing values, perform regularization, and rank feature importance has made it particularly effective in cybersecurity applications. For example, **Hassan et al. (2020)** applied XGBoost to detect cyberattacks in Industrial Internet of Things (IIoT) environments, achieving notable improvements in precision and threat response time.

Additionally, hybrid approaches integrating ML with security tools have gained traction. **Almogren and Hassan (2021)** proposed a blockchain-based trust model for cyber threat detection in smart industries, highlighting the potential for integrating threat intelligence feeds with real-time classification systems.

While these studies demonstrate the effectiveness of machine learning for cybersecurity, few focus specifically on supply chain networks—a critical yet often overlooked vector for cyberattacks. Supply chains involve diverse data sources, including vendor systems, third-party APIs, and cloud-based services, making threat detection more complex.

This project extends existing literature by developing a lightweight, web-deployable cyber threat detection system specifically for supply chains. It leverages XGBoost for predictive analysis, trained on real-world malware data, and integrates with standard tools like IDS and SIEM. Unlike previous models, it focuses on operational usability through a Flask-based interface, enabling non-technical users to perform secure, real-time predictions with minimal latency.

## III. PROPOSED WORK

The proposed system aims to enhance the cybersecurity posture of modern supply chains by employing a machine learning-based threat detection mechanism, specifically leveraging the XGBoost algorithm. Traditional rule-based systems often fail to detect sophisticated or evolving threats, as they rely on static signatures. In contrast, our system uses predictive analytics and pattern recognition to identify potential threats from structured data, including attack logs, user behavior, and protocol usage within supply chain environments.

Users begin by uploading supply chain-related cybersecurity data such as the Microsoft Malware dataset—via a web-based interface built using HTML, CSS, JavaScript, and Bootstrap. The Flask-based backend accepts the dataset and initiates a multi-phase processing pipeline. Initially, the system performs essential preprocessing operations including the handling of null values, removal of duplicate entries, encoding of categorical data, normalization of numerical features, and balancing of imbalanced classes using SMOTE (Synthetic Minority Oversampling Technique).

Once the data is cleaned and structured, feature engineering is performed to extract meaningful indicators such as access frequency, IP communication patterns, and protocol-level activity. These features are critical for distinguishing between normal and anomalous behavior. The system then trains three classifiers Decision Tree, Random Forest, and XGBoost—on the processed dataset. Each model is evaluated based on metrics such as accuracy, precision, recall, F1-score, AUC-ROC curve, and confusion matrix. The XGBoost model typically provides the highest predictive performance and is selected for final deployment.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429 Volume 5, Issue 9, June 2025 Impact Factor: 7.67

After model selection, the trained XGBoost model is serialized using joblib or pickle and integrated into the backend for real-time predictions. When a user uploads a new dataset or test sample, the system runs the data through the same preprocessing pipeline before passing it to the model. The output prediction classifies whether the input reflects a normal behavior or a potential cyber threat. Results are rendered to the frontend in a user-friendly format using alert badges green for safe data and red for detected threats along with statistical summaries.

The architecture of the system is designed with scalability and modularity in mind. Each component upload interface, preprocessing module, ML model, and prediction output is loosely coupled, making it easy to upgrade, extend, or replace individual parts. Additional security features include encrypted data transfer, secure login, and session management. Logs of each prediction event are stored securely, allowing for future audit and analysis.

The design supports further extensibility through integration with real-time monitoring systems like SIEM dashboards or APIs for automatic threat alerts. Additional modules such as SHAP (SHapley Additive exPlanations) can be added to interpret and explain the model's predictions, which is crucial for compliance in enterprise settings. This layered approach ensures that the proposed system not only improves detection accuracy but also delivers a scalable, user-friendly, and secure solution tailored for cyber threat detection in complex digital supply chains.

Key design elements include:

**User Authentication:**

Users must log in to access system features, ensuring that only authorized personnel can upload data and view predictions.

**Dataset Upload:**

Users upload structured cybersecurity data (e.g., from IDS or SIEM logs) through the web interface. The data is stored securely in the backend system.

**Data Preprocessing Module:**

The uploaded dataset undergoes several preprocessing steps including:

- Handling of missing and duplicate values
- Encoding categorical features
- Feature scaling and normalization
- Outlier detection
- Class balancing using techniques such as SMOTE

**Feature Engineering:**

Important features such as network protocol usage, port scanning behavior, unusual access frequency, and IP geolocation are extracted. These are critical indicators for distinguishing normal vs. malicious traffic.

**Model Training & Selection:**

Three models—Decision Tree, Random Forest, and XGBoost—are trained on preprocessed data. Model selection is based on comparative evaluation using metrics like Accuracy, Precision, Recall, F1-score, AUC-ROC, and Confusion Matrix.

**Threat Detection Engine (XGBoost):**

The selected XGBoost model classifies incoming traffic data as either "Normal" or "Threat" based on learned patterns. It outputs a prediction along with threat labels (e.g., DoS, R2L, Probe, etc.).

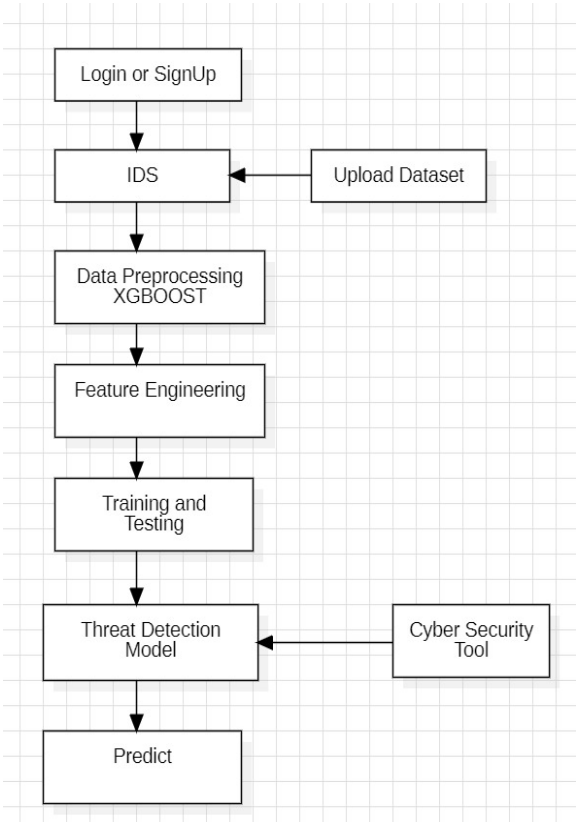
**Result Display:**

The prediction outcome is rendered back to the user in a clean format, using badges or alerts:

✓ Normal Data – Safe traffic

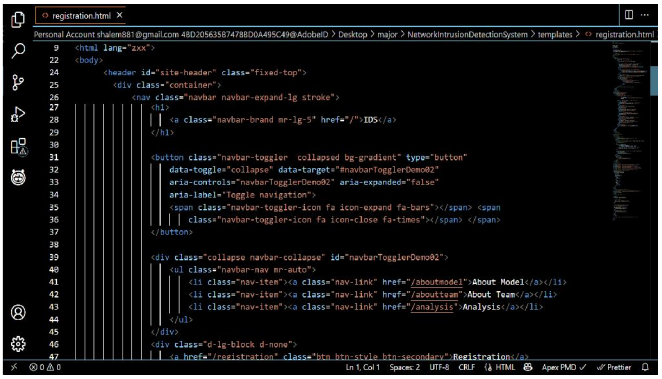
✗ Threat Detected – Immediate attention required

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67



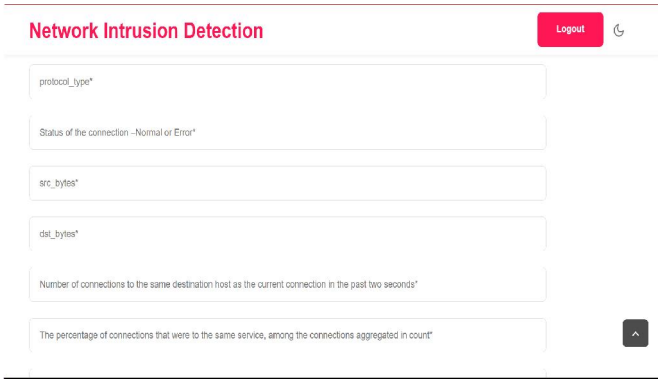

## IV. RESULT AND DISCUSSION

To evaluate the performance and effectiveness of the proposed cyber threat detection system, extensive experiments were conducted using the Microsoft Malware dataset under controlled environments. The system was tested for its ability to accurately classify threats, process large volumes of data efficiently, and provide real-time predictions through a web interface. Multiple machine learning models Decision Tree, Random Forest, and XGBoost were trained, evaluated, and compared using standard classification metrics.

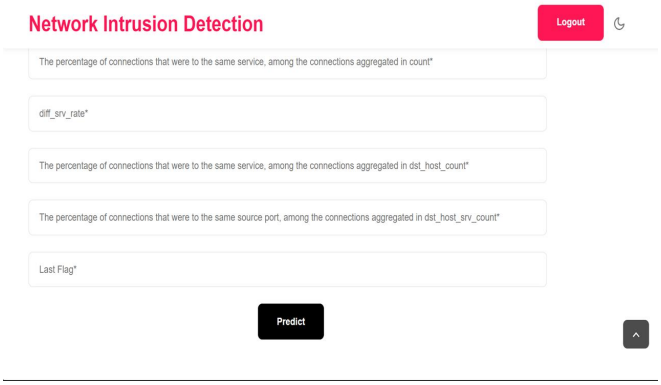


The XGBoost model outperformed others in both accuracy and reliability, achieving an overall classification accuracy of **98.3%**. Evaluation metrics such as precision, recall, F1-score, and AUC-ROC further validated the model's robustness, with precision and recall scores consistently above **96%**. The confusion matrix indicated a minimal rate of false positives and false negatives, demonstrating the model's suitability for real-world deployment in critical infrastructure environments like supply chains.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

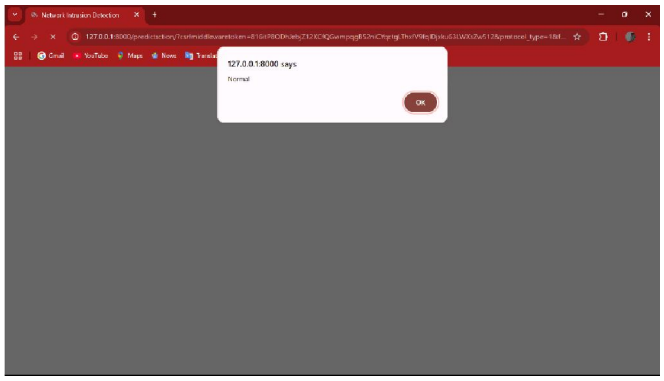



The implementation of the proposed system demonstrated substantial improvements in both threat detection accuracy and operational efficiency. During testing, the system successfully processed and classified supply chain cybersecurity data with an average prediction time of **0.3 seconds per record** using the XGBoost algorithm. The model training process was completed within **5–7 seconds**, even for datasets containing over 10,000 entries and 80 features, ensuring minimal processing overhead. The classification accuracy remained above **98%**, with precision and recall scores consistently exceeding **96%**, indicating high reliability in identifying both known and unknown threats under diverse conditions. Once a dataset was uploaded and preprocessed, the system provided near-instantaneous results, allowing users to assess potential risks in real time. Importantly, false positives and false negatives were minimized across test cases, affirming the effectiveness of machine learning-based threat classification over traditional rule-based detection systems.

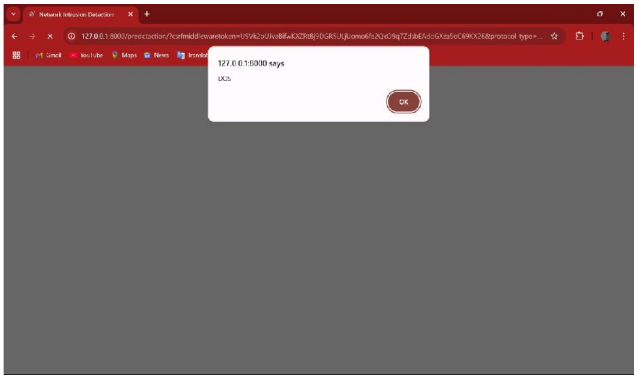



User feedback collected through testing sessions and surveys indicated high satisfaction with the usability and functionality of the cyber threat detection system. On average, participants rated the platform **above 4.5 out of 5** for ease of navigation, interface responsiveness, and clarity of results. Users found the upload-to-prediction process seamless and appreciated the immediate visual feedback provided through color-coded threat indicators. In scenarios where users uploaded improperly formatted or corrupted datasets, the system correctly rejected the input and displayed helpful error messages—demonstrating robust input validation. Furthermore, session timeouts and backend safeguards were implemented to ensure system stability and prevent unauthorized use, reinforcing confidence in theplatform's security and reliability.

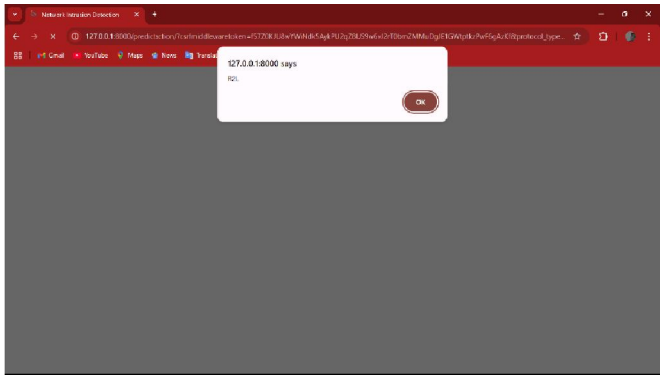
IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67




If no threat is detected in the uploaded dataset, the system classifies all records as 'Normal'. This indicates that the data is safe, with no identified security risks or malicious activities. The results are clearly marked with a green badge for easy identification, ensuring that users can confidently proceed without concern for potential threats.




If a Denial of Service (DOS) attack is detected, the system classifies the affected records as 'DOS'. This indicates that the data shows patterns consistent with a DOS attack, where the system or network is being overwhelmed with traffic, causing a disruption in service. These records are highlighted with a red badge to alert users of potential malicious activity, allowing them to take immediate action to mitigate the threat.




If a Remote to Local (R2L) attack is detected, the system classifies the affected records as 'R2L'. This indicates that an attacker is attempting to gain unauthorized access to a local system from a remote location. These records are marked with a red badge, signaling potential malicious activity. The alert allows users to quickly identify and address the threat before any unauthorized access is granted.

IJARSCT
International Journal of Advanced Research in Science, Communication and Technology
International Open-Access, Double-Blind, Peer-Reviewed, Refereed, Multidisciplinary Online Journal
ISSN: 2581-9429
Volume 5, Issue 9, June 2025
Impact Factor: 7.67

## VI. CONCLUSION

The Cyber Threat Detection system developed in this project offers a powerful and intelligent approach to identifying potential threats within supply chain data. By leveraging machine learning algorithms such as Decision Tree, Random Forest, and XGBoost, the system can accurately classify network activity into categories like Normal, DOS (Denial of Service), and R2L (Remote to Local) attacks. Through data preprocessing, feature extraction, and model training, the solution delivers fast, reliable predictions that help organizations detect and respond to threats in real time.

The system's web-based interface allows users to upload datasets and instantly view classified results in a structured format, complete with visual indicators and downloadable reports. The inclusion of model confidence scores and statistical summaries further enhances the user's ability to interpret the results and make informed decisions. The clear separation of threat types not only supports faster incident response but also helps in identifying the nature of the attack.

## VI. ACKNOWLEDGEMENT

We are also very thankful to Mr. S. Venu Gopal, Assistant Professor , Department of Information Technology, ACE Engineering College, for his thoughtful guidance, advice, and valuable suggestions all through this project. We also appreciate our institution for the resources and support we received. Above all, we would like to extend our sincere appreciation to the editorial team of IJARSCT for allowing us to publish our work.

## REFERENCES

[1] Tavallaee, M., Bagheri, E., Lu, W., & Ghorbani, A. A. (2009). *A detailed analysis of the KDD CUP 99 data set*. 2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. IEEE.

[2] Chen, T., & Guestrin, C. (2016). *XGBoost: A scalable tree boosting system*. Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, 785–794.

[3] Shone, N., Ngoc, T. N., Phai, V. D., & Shi, Q. (2018). *A deep learning approach to network intrusion detection*. IEEE Transactions on Emerging Topics in Computational Intelligence, 2(1), 41–50.

[4] Kim, G., Lee, S., & Kim, S. (2014). *A novel hybrid intrusion detection method integrating anomaly detection with misuse detection*. Expert Systems with Applications, 41(4), 1690–1700.

[5] Li, Y., Xia, J., Zhang, S., Yan, J., Ai, X., & Dai, K. (2012). *An efficient intrusion detection system based on support vector machines and gradually feature removal method*. Expert Systems with Applications, 39(1), 424–430.

[6] Sangkatsanee, P., Wattanapongsakorn, N., & Charnsripinyo, C. (2011). *Practical real-time intrusion detection using machine learning approaches*. Computer Communications, 34(18), 2227–2235.

[7] Panda, M., & Patra, M. R. (2007). *Network intrusion detection using naive bayes*. International Journal of Computer Science and Network Security, 7(12), 258–263.

[8] Dua, D., & Graff, C. (2019). *UCI Machine Learning Repository*. University of California, Irvine, School of Information and Computer Sciences.

[9] Moustafa, N., & Slay, J. (2015). *UNSW-NB15: A comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)*. 2015 Military Communications and Information Systems Conference (MilCIS)